Noname manuscript No.
(will be inserted by the editor)

A Study of Bag-of-Visual-Words Representations for Handwritten Keyword Spotting

David Aldavert · Marçal Rusiñol · Ricardo Toledo · Josep Lladós

Received: date / Accepted: date

Abstract The Bag-of-Visual-Words (BoVW) framework has gained popularity among the document image analysis community, specifically as a representation of handwritten words for recognition or spotting purposes. Although in the computer vision field the BoVW method has been greatly improved, most of the approaches in the document image analysis domain still rely on the basic implementation of the BoVW method disregarding such latest refinements. In this paper we present a review of those improvements and its application to the keyword spotting task. We thoroughly evaluate their impact against a baseline system in the well-known George Washington dataset and compare the obtained results against nine state-of-the-art keyword spotting methods. In addition, we also compare both the baseline and improved systems with the methods presented at the Handwritten Keyword Spotting Competition 2014.

Keywords Bag-of-Visual-Words; Keyword Spotting; Handwritten Documents; Performance Evaluation

1 Introduction

Keyword spotting can be defined as the pattern recognition task aimed at locating and retrieving a particular keyword within a document image collection without explicitly transcribing the whole corpus. Its use is particularly interesting when applied in scenarios where Optical Character Recognition (OCR) performs poorly or can not be used at all, such as in historical document collections, handwritten documents, etc. Being a mature research problem [30], many different keyword spotting approaches have been proposed thorough the years.

In the document image analysis literature, we can distinguish two different families of keyword spotting methods depending on the representation of the handwritten words [26]. On the one hand, *sequential* word representations [35] describe handwritten words as a time series by using a sliding window in the writing direction. On the other hand, *holistic* word representations [29] extract a single feature vector of fixed dimensionality that characterizes the word as a whole.

Sequential word representations exploit the sequential nature of handwritten words formed by the concatenation of individual characters. However, since the size of the word's descriptors will depend on the width of the word, two different words cannot be directly compared by means of a distance between points, but some sort of alignment technique has to be used instead. The seminal work by Kołcz et al. [19] achieved a breakthrough in the handwritten keyword spotting domain by proposing the use of the Dynamic Time Warping (DTW) method (often used in speech analysis) for nonlinear sequence alignment. The use of DTW together with profile features was popularized by the well-known works by Rath and Manmatha [37,38] and Rath et al. [39] and many flavors of DTW-based handwritten keyword spotting methods appeared since those publications. Adamek et al. proposed in [1] to use DTW to align convexity and concavity features extracted from contours. Khurshid et al. presented in [18] a method that first aligned features at character level by DTW and then the resulting character prototypes are aligned at word level. Papandreou et al. [33], proposed an adaptive zoning description that can be matched by DTW. Besides direct matching strategies, learning-based methods have also

D. Aldavert · M. Rusiñol · R. Toledo · J. Lladós
Computer Vision Center, Dept. Ciències de la Computació
Edifici O, Universitat Autònoma de Barcelona
08193 Bellaterra (Barcelona), Spain
E-mail: {aldavert,marcal,ricard,josep}@cvc.uab.es

been proposed over the years. Hidden Markov Models are the most widely used techniques to model the keywords' sequential features [41,43,42,12,44], although other machine learning approaches such as Neural Networks [13] have also been used in the keyword spotting domain.

Holistic word representations have also received some attention thorough the years. Their main advantage is that by representing handwritten words by feature vectors of fixed size, the alignment step (which usually is very time consuming) is bypassed, and thus, two handwritten words can be compared using standard distances, or any statistical pattern recognition technique. We can find many different holistic word descriptions used in the literature for keyword spotting tasks. For example, simplified versions of the shape context descriptor, have been used in example-based keyword spotting architectures by Lladós and Sánchez [27] or by Fernández et al. [11]. Zoning-based characteristics have also been widely used to represent word images holistically, e.g. [20,17]. A combination of Histogram of Oriented Gradients (HOG) and Local Binary Patterns descriptors has been proposed by Kovalchuk et al. in [21] in a segmentation-free keyword spotting scenario. A set of biologically inspired features formed by a cascade of Gabor descriptors was proposed by van der Zant and Schomaker in [57]. The combination of gradient, structural and concavity features was proposed by Srihari and Ball in [54]. All of these word representations present their strengths and weaknesses and is hard to argue that a set of features is steadily better than another. Although in the latest years a trend towards using gradient-based features can be appreciated [40].

1.1 Keyword Spotting as an Object Recognition Task

Since the publication of the SIFT method [28], the computer vision task of recognizing and finding objects in cluttered scenes has been driven by methods extracting local descriptors that are further matched between the query model and the scene images. Many authors from the document analysis field, understanding keyword spotting as being a particular case of the object recognition task, started to apply such keypoint matching techniques to the problem of keyword spotting [48, 23,58,56]. Such matching techniques have been either used to directly estimate similarities between word images, or by searching the query model image within full pages in segmentation-free scenarios. However, the keypoint matching framework presents the same disadvantage than the sequential methods since an alignment between the keypoint sets has to be computed.

In order to avoid exhaustively matching all the keypoints among them, the classic bag-of-words paradigm from the information retrieval field was reformulated as the Bag-of-Visual-Words (BoVW) [53,8]. Such paradigm yield an holistic and fixed-length image representation while keeping the discriminative power of local descriptors such as SIFT.

Soon enough, researchers from the document image analysis domain adapted such BoVW representations to the keyword spotting problem [5,49,47,51,10, 44,50,46], obtaining very competitive results. However, we have the feeling that although the computer vision community kept proposing improvements on the BoVW framework in the last years, in the document analysis field, such improvements are still scarcely used. As an exception, it is worth to cite the works from Shekhar and Jawahar [52], or our last contribution [2], where more complex BoVW setups are used for the keyword spotting task.

1.2 Contributions and Outline of the Paper

In this paper we will review some of the latest improvements over the BoVW framework, namely sparse coding, spatial pyramids, and power normalization and its application to the keyword spotting task. We will thoroughly evaluate the impact of such improvements as well as the different parameters of the BoVW method by comparing their performances against a baseline system. We will finally compare the obtained results against nine state of the art segmentation-based keyword spotting methods by using the well-known George Washington dataset. In addition, we also compare both the baseline and improved systems with the methods presented at the Handwritten Keyword Spotting Competition 2014.

The paper is structured as follows, in Section 2, the different parts of the BoVW pipeline used to characterize the word images are presented. Then, the effects that each BoVW enhancement have in the performance of a keyword spotting system are evaluated in Section 3 and the results obtained by the system are compared with the state of the art in Section 4. Finally, we review the most important conclusions of the paper in Section 5.

2 Bag-of-Visual-Words Representations

In order to spot keywords in document images, we start by a layout analysis step devoted to segment the document images into individual words. The interested reader is referred to [25,31]. Once the words are segmented,

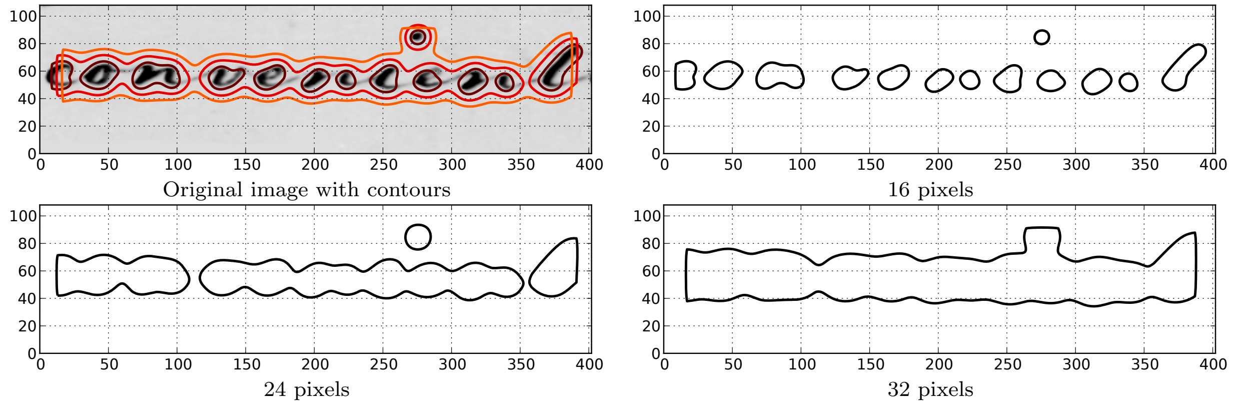


Fig. 1: Norm of the descriptors extracted from regions of 16, 24 and 32 pixels width sampled at each pixel of the image. The bold contours encircle the regions where descriptors which have a large enough norm and are considered reliable.

a visual signature is computed for each of them. The keyword spotting will be then performed by calculating the similarity between the description of the query word and all the descriptors of the words in the corpus. These visual signatures are created using a Bag-of-Visual-Words (BoVW) framework which has obtained good performances in keyword spotting tasks [47,51].

The BoVW framework has many variants in the literature, but all of them can be roughly divided into four basic steps: *sampling*, *description*, *encoding* and *pooling*. In order to increase the retrieval performance of the spotting system, we need to carefully select the methods used at each step. In this paper, we will mainly focus on the BoVW improvements that bring better word representations for recognition or spotting tasks.

2.1 Sampling

The first step is to select the regions of the image which contain meaningful information to describe the word snippets. Although covariant or salient region detectors can be used, it has been proven that the performance of BoVW representations is correlated with the number of sampled regions. For instance, Nowak et al. demonstrate in [32] that the larger the number of regions, the better the results. They show that the combination of several region detectors usually improves the performance of the BoVW framework, but this performance gain is related to the number of regions rather than the kind of sampled regions. Therefore, for our baseline implementation we decided to densely sample regions at different scales over the image instead of using a keypoint detector.

Regions are densely sampled using a fixed step and at different scales. The different scales are selected so that words are going to be modeled at different levels of detail: small regions will model portions of characters while large regions will model the relationships between characters.

2.2 Description

Once regions have been sampled, we need to characterize them with a local descriptor. Although descriptors specifically tailored for document analysis can be used, gradient based descriptors have recently shown better performances in keyword spotting tasks [3,47,2].

We are going to use the Histogram of Oriented Gradients (HOG) descriptor [9] to characterize the regions. This descriptor is derived from the SIFT descriptor [28], but it is more suited for dense sampling scenarios when rotation invariance is not needed. In our case, it is safe to assume that the orientation of the word images has been corrected by the word segmentation algorithm or intermediate slant correction steps. The HOG algorithm takes advantage of the information redundancy between overlapping regions, so that descriptors can be calculated at a much lower computational cost [59,14].

Although the dense sampling strategy will generate a large amount of HOG descriptors, only reliable descriptors are eventually accepted. Since HOG descriptors are based on gradient information, descriptors are more reliable when gradient vectors have a large module. Therefore, the norm of the descriptor can be used as a reliability indicator. For instance, Fig. 1 shows the norm of the HOG descriptors calculated at each pixel of the image. It can be appreciated that descriptors calculated near character locations have a high norm while descriptors sampled over other image regions have a low norm. Therefore, the BoVW signature can focus on

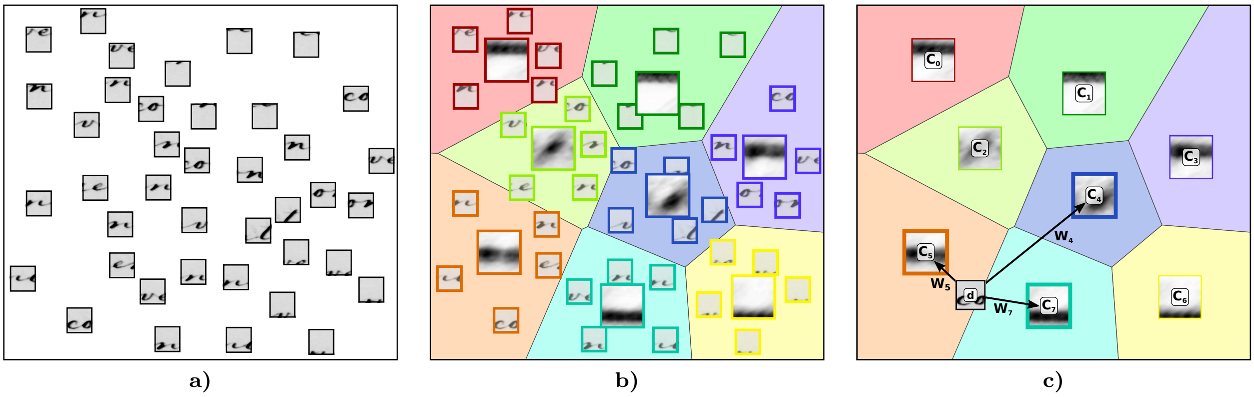


Fig. 2: Codebook creation and descriptor encoding example: **a)** Descriptors are randomly sampled from the indexed images, **b)** the k-means algorithm is used to build the codebook and **c)** descriptors are encoded using sparse coding with the cluster centroids.

the visual information from characters by filtering the descriptors depending on the value of their norm. The bold contours in the Fig. 1 encircle the zones where the descriptors have a norm higher than the threshold used in the paper. Descriptors which have a value lower than this threshold, i.e. descriptors outside the contours, are simply disregarded.

2.3 Encoding

After calculating the descriptors, we have to encode them into visual words. First, we need a codebook which quantizes the descriptor space into an arbitrary set of m salient regions. This codebook is created by randomly sampling descriptors from the indexed word snippets and using the k-Means algorithm to calculate m clusters. Then, a descriptor $\mathbf{d}_i$ is encoded by a vector $\mathbf{W}_i \in \mathbb{R}^m$ which weights the contribution of each codeword (i.e. cluster centroid). The most straightforward method to calculate $\mathbf{W}_i$ is to use hard-assignment [53], i.e. the weight vector has a single non-zero element corresponding to the nearest codeword to the descriptor.

This encoding approach has problems near the boundaries between codewords. Small changes in the descriptor may lead to a completely different visual words vector $\mathbf{W}_i$. This problem can be alleviated by using soft-assignment instead, i.e. encoding a descriptor using a weighted combination of codewords. Besides, combining the information of several codewords also reduces the information loss resulting of the descriptor quantization. Therefore, we decided to encode descriptors using the sparse coding technique proposed in [55], known as Locality-constrained Linear Coding (LLC). This method generates a compact BoVW signature that have a higher discriminative power than more complex representations [6].

Given a descriptor $\mathbf{d}_i$, the LLC method tries to find the linear combination of codewords which better approximates the original descriptor:

$$\mathbf{d}_i \approx \sum_{j=1}^{m} w_j \mathbf{C_j}, \tag{1}$$

where $\mathbf{C_j}$ is the j-th codeword and w_j its associated weight. Unlike other sparse coding algorithms, LLC emphases locality over sparsity and it only uses the t nearest codewords to encode a descriptor. This ensures that the resulting encoding is locally smooth, so that similar descriptors are likely to be encoded using the same codewords. Therefore, the LLC encoding is more robust compared to other sparse coding solutions. Another advantage is that the weights $(w_1, w_2, \ldots, w_m)$ can be derived analytically. Hence, the computational cost is drastically reduced compared to other sparse coding algorithms which require computationally demanding optimization procedures to find a solution. Then, a descriptor $\mathbf{d}_i$ is encoded by searching the t nearest codewords and using the LLC algorithm to calculate the weights vector $\mathbf{W}_i = (w_1, w_2, \ldots, w_m)$.

An example of the codebook creation and descriptor steps is summarized in Fig. 2. The randomly sampled descriptors of Fig. 2.a) are clustered into eight clusters in Fig. 2.b). In Fig. 2.c), we can see that the closest codewords to the descriptors $\mathbf{d}_i$ are $\mathbf{C}_4$, $\mathbf{C}_5$ and $\mathbf{C}_7$. Using hard-assignment, the descriptor will be encoded as $\mathbf{W}_i = (0, 0, 0, 0, 0, 1, 0, 0)$ as its nearest centroid is $\mathbf{C}_5$. On the other hand, the LLC algorithm will calculate the weights w_4, w_5 and w_7 so that $\mathbf{d}_i \approx w_4\mathbf{C}_4 + w_5\mathbf{C}_5 + w_7\mathbf{C}_7$ and the resulting encoding will be $\mathbf{W}_i = (0, 0, 0, 0, w_4, w_5, 0, w_7)$. Notice that the encoded descriptor is close to a boundary between codewords, so that a small variation of the descriptor can shift the closest codeword from $\mathbf{C}_5$ to $\mathbf{C}_7$. This would result in a completely different encoding when hard-assignment is used. In contrast, the LLC algorithm will generate a similar weight vector $\mathbf{W}_i$ since it still uses the same codewords and the weights w_4, w_5 and w_7 are slightly different.

2.4 Pooling

Once descriptors are encoded into visual words, the BoVW signature is obtained by simply accumulating the weight vectors $\mathbf{W}_i$:

$$\mathbf{s} = \sum_{i=1}^{N} \mathbf{W}_i, \tag{2}$$

where N is the number of valid descriptors extracted from the word image. In the following, we are going to see how to improve this representation.

2.4.1 Spatial information

In Eq. 2, visual words are accumulated without taking into account their spatial location, so the signature lacks any spatial information. However, spatial information is quite important in keyword spotting tasks since it helps to reduce the perceptual aliasing problem. Different instances of the same character are expected to be represented by similar visual words. Hence, the obtained BoVW signatures mostly depends on the characters that form the word, and it is possible that dissimilar words are represented by similar signatures when spatial information is not taken into account. For instance, anagrams will obtain a very similar visual signature in this scenario.

This problem can be addressed by using using the Spatial Pyramid Matching (SPM) technique proposed by Lazebnik et al. in [22] in order to add some spatial information into the unstructured BoVW model. This method roughly takes into account the visual word distribution over the image by creating a pyramid of spatial bins.

The spatial pyramid defines an initial set of horizontal P_x^0 and vertical P_y^0 partitions which create $P_x^0 \times P_y^0$ spatial bins. Then, these spatial bins are further divided into P_x horizontal and P_y vertical partitions at each level of the pyramid. Therefore, a spatial pyramid of L levels creates a collection of overlapping D_{sp} spatial bins, where

$$D_{sp} = P_x^0 P_y^0 \sum_{l=0}^{L-1} (P_x P_y)^l. \tag{3}$$

The final BoVW signature $\overline{\mathbf{W}}_i$ is created by independently accumulating the visual words for each spatial bin obtaining a $D_W = mD_{sp}$ dimensions descriptor. The amount of visual words assigned to each bin is lower at higher levels of the pyramid, due to the fact that the spatial bins are smaller. This is compensated by multiplying the contribution of each visual word to each spatial bin by the factor $s_l = P_x^0 P_y^0 (P_x P_y)^l$.

2.4.2 Normalization

Once we have obtained $\overline{\mathbf{W}}_i$, we can normalize the contribution of each visual word in order to obtain a better representation. First, we can reduce the importance of overrepresented visual words by using the method proposed by Perronnin et al. in [34] which applies the following normalization function to each bin of the signature:

$$g(x) = sign(x)|x|^{\alpha}, \tag{4}$$

where $0 < \alpha < 1$ is the power normalization factor. The power normalization improves the BoVW model since it removes the assumption that visual words come from an identically and independently distributed population [7]. Avoiding the i.i.d. assumption is important in keyword spotting as the frequency of visual words is highly correlated to the characters forming the word. For instance, the visual words modeling the character *e* will be overrepresented in words like *freeze* or *exceed* and hence their visual signature is going to be somehow similar. Therefore, by lessening the contribution of the overrepresented visual words, we are highlighting the other visual words and making both signatures more dissimilar.

Finally, the BoVW signature is ℓ_2-normalized to account that the amount of visual words accumulated in $\overline{\mathbf{W}}_i$ may change between two instances of the same word due to scale difference or image noise.

3 BoVW Parameter Evaluation

In order to evaluate the different parameters of the BoVW signature in a keyword spotting framework, we use a straightforward method to index and retrieve the word snippets from a database. The image signatures are indexed using an inverted file structure taking advantage that the BoVW representation is sparse, specially when SPM is used. The system is evaluated by calculating the mean Average Precision (mAP) score from the ranked list obtained by sorting in ascending order the Euclidean distances between the query and the indexed signatures.

3.1 Experimental Setup

The keyword spotting system is evaluated in the George Washington dataset described in [38]. This dataset consists of 20 handwritten pages with a total of 4860 words written by several Washington's secretaries. Although it was written by several authors, the writing style is pretty uniform and shows less variation than typical multi-writer collections. The database provides a set of word bounding-boxes with their transcription. These bounding-boxes are obtained using the segmentation algorithm proposed in [31] by Manmatha and Rothfeder.

The baseline BoVW configuration densely samples the HOG descriptors at every 5 pixels and at three different scales: 20, 30 and 45 pixel wide regions. The codebook has $m = 1024$ codewords and the histogram is created without using any improvement, i.e. descriptors are encoded using hard-assignment, no spatial information is added and the power normalization is not used (i.e. $\alpha = 1$). At each step of the experimental evaluation, we are going to assess the effects that a single improvement has on the spotting performance of the system. These evaluations are conducted by calculating the mAP score using two different setups:

- **Setup A**: Use as queries all words in the collection which appear at least twice.
- **Setup B**: Use as queries only words which have at least ten occurrences and with 3 or more characters.

The configuration *setup A* is defined to use all possible word snippets as queries while the configuration *setup B* cast queries which are more likely to be used in a real world scenario (e.g. avoiding short queries like "*a*" or "*to*").

In both setups, word snippets which have been discarded as queries are still used as distractors in the database. Therefore, the system has a 100% recall since it always returns a ranked list with all the 4859 elements, corresponding to all indexed images except the query.

3.2 LLC Encoding

First, we evaluate the effects of using a different amount of nearest neighbors t in the LLC encoding step. The mAP scores obtained while testing from 1 to 16 nearest neighbors are shown in Fig. 3. Note that using a single nearest neighbor corresponds to hard-assignment encoding, since only the closest codeword is used.

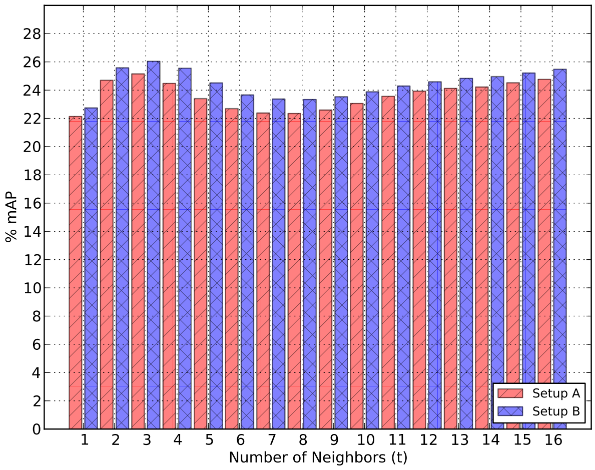


Fig. 3: mAP score obtained using different number of neighbors with LLC.

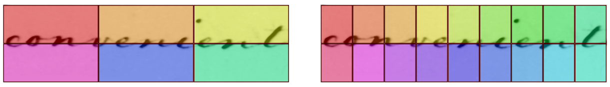
The results show that using LLC encoding slightly increases the performance of the word spotting system. The best results are obtained when three nearest neighbors are used to encode the descriptors: for *setup A* the mAP score improves from 22,13% to 25,15% while for *setup B* the score raises from 22,74% to a 26,04%. Although the selected number of neighbors may seem small, this result is coherent with the results shown in the original LLC paper [55] where using a small number of neighbors results in a better performance than when a large number of neighbors is employed. In the remaining experiments, we are going to use 3-nearest neighbors for the encoding step with LLC.

3.3 Spatial Pyramids

After evaluating the encoding, we are going to evaluate the importance of spatial information in the BoVW signature. In Table 1 we can see that the addition of spatial information greatly increases the performance of the system. In both setups, the mAP score increases two and a half times between the orderless representation and the best spatial pyramid configuration. From the obtained results, we can see that horizontal partitions are more important than vertical partitions. This is to be expected as adding more horizontal partitions helps to increase the representation of the word characters. For instance, in Fig. 4 we can see an example of the spatial bins defined by a two level spatial pyramid. In the first level, spatial bins roughly model syllables while

Table 1: mAP score obtained using different spatial configurations.

P_x^0	P_y^0	P_x	P_y	L	D_{sp}	D_W	**Setup A**	**Setup B**
1	1	1	1	1	1	1024	25,15%	26,04%
1	1	2	2	2	5	5120	40,96%	43,43%
1	1	2	2	3	21	21504	51,49%	54,03%
1	1	2	2	4	85	87040	57,65%	60,47%
1	1	3	2	2	7	7168	46,45%	48,79%
1	1	3	2	3	43	44032	58,09%	60,91%
1	1	3	2	4	259	265216	61,11%	64,26%
1	1	2	3	2	7	7168	42,45%	45,05%
1	1	2	3	3	43	44032	51,38%	53,91%
2	2	2	2	2	20	20480	55,46%	58,53%
3	2	2	2	2	30	30720	60,32%	63,56%
2	3	2	2	2	30	30720	55,71%	58,80%
2	2	3	3	2	40	40960	59,27%	62,43%
3	3	3	3	2	90	92160	62,01%	65,46%
3	1	2	2	2	15	15360	58,39%	61,43%
1	3	2	2	2	15	15360	43,37%	45,97%
3	1	2	1	2	9	9216	55,32%	58,50%
3	1	2	1	3	21	21504	58,98%	62,27%
3	1	3	2	2	21	21504	60,38%	63,66%
3	**2**	**3**	**1**	**2**	**24**	**24576**	**61,33%**	**64,75%**



First level **Second level**

Fig. 4: Distribution of the spatial bins in the two levels of the spatial pyramid.

in the second level bins are smaller and they model individual characters.

After evaluating the obtained results, we have selected a two level SPM with 3 × 2 spatial bins in the first level and 9 × 2 in the second (row in bold in Table 1) as the SPM configuration used in the following experiments. With this configuration the retrieval performance grows from 22, 15% to 61, 33% using *setup A* and from 26, 04% to 64, 75% in *setup B*. Although there is another configuration which obtains better results, the selected configuration offers a better compromise between performance and dimensionality growth. Additionally, we have re-checked the effects of LLC by disabling it and the performance is slightly reduced to 60, 62% and 64, 16% respectively.

3.4 Power normalization

Concerning power normalization, the retrieval performance obtained using different α power values can be found in Fig. 5. The results show that the use of power normalization also obtains an important boost of performance of the system. It attains the maximum performance of 68, 27% mAP at $\alpha = 0, 4$ for *setup A* and of 72, 20% mAP at $\alpha = 0, 3$ for *setup B*. Since the performance is pretty similar for $\alpha = 0, 3$ and $\alpha = 0, 4$, we are going to use a power normalization of $\alpha = 0, 35$ for both setups in the following experiments.

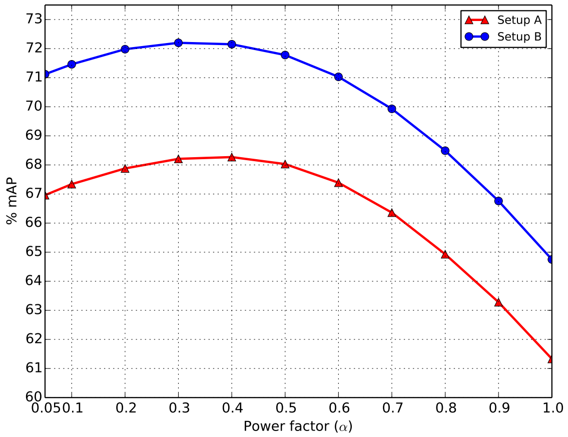


Fig. 5: Effect of the power norm to the performance of the word spotting system.

3.5 Codebook size

All the experiments until now have used a relatively small codebook of 1024 codewords. Since the performance usually increases as larger codebook are used, we compare the effects of different codebook sizes in Fig. 6.

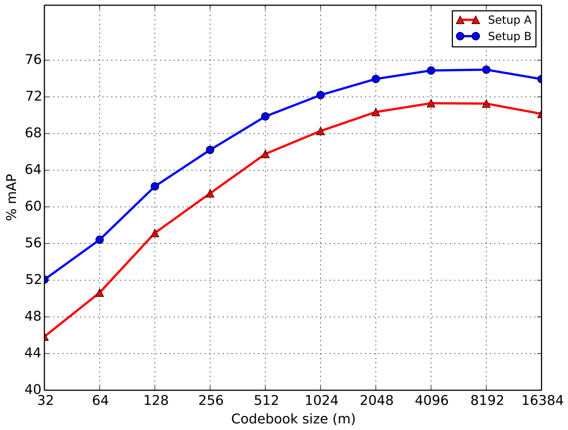


Fig. 6: Evolution of the mAP score while increasing the size of the codebook.

The performance of the system keeps improving until it saturates for the $m = 8192$ codebook. For larger codebooks, the performance degrades, because descriptor quantization errors start to be too frequent. Since the mAP score increase is marginal between codebooks of $m = 4096$ and $m = 8192$, we decided to use the 4096-codebook for the last experiment.

It is worth noting that the mAP score attained by the smallest codebook (with $m = 32$ codewords) in Fig. 6 doubles the score obtained by the baseline configuration: $45,85\%$ against $22,13\%$ for *setup A* and $52,07\%$ versus $22,74\%$ for *setup B*. Although the BoVW signature is more compact and it has 768 dimensions compared to the 1024 dimensions of the baseline configuration, the use of LLC, SPM and power normalization greatly increase the spotting capabilities of the system.

3.6 Descriptor sampling

Subsequently, we evaluate in Table 2 the effects of using different the descriptor sampling parameters. We have evaluated the use of larger regions to check which information is more important to characterize word images. The results show that it is more important that visual words model character fragments rather than the relationships among them. We have also evaluated the sampling density, observing that the performance increases as the descriptors are sampled more densely. Since the performance gap between the two configurations is quite important, it is safe to assume that works that used larger regions (e.g. our previous segmentation-free keyword spotting method [47]) will improve their performance by simply using smaller regions.

Table 2: mAP scores obtained when modifying the descriptor sampling parameters

Region size			**Step**	**Setup A**	**Setup B**
small	*medium*	*large*			
40	60	90	10	39,94%	43,71%
			8	43,37%	47,54%
			5	47,24%	51,61%
			4	47,75%	52,20%
			3	47,90%	52,35%
20	30	45	10	54,23%	58,25%
			8	62,94%	66,85%
			5	71,31%	74,88%
			4	72,35%	75,86%
			3	**72,98%**	**76,45%**

3.7 Summary of the Results

Finally, we present in Table 3 a summary of the results obtained by the different improvements over the baseline BoVW implementation. Besides the performance gains for each of the improvements, we also report the extra cost that each of the different steps might have. Both using sparse coding through LLC and tuning the descriptor sampling stage have a minimal cost in terms of computational complexity. In the encoding step the weights of the LLC have to be calculated instead of just using a hard-assignment strategy. When using denser and smaller HOG descriptors, the amount of descriptors to process per word image is increased, and thus the whole encoding and pooling steps are more complex to compute. When using an SPM configuration, the dimensionality of the word descriptors is exponentially increased, so one has to find a good trade-off between discriminative power and efficiency of the overall system in terms of speed and memory usages. The same goes for the codebook size, although we have seen that in that case, the system's performance degrades when starting to use too large dictionaries. Finally, the use of power normalization has no extra cost with regard to the baseline BoVW implementation. After the final experiment, the performance of the system has increased a 230% (from $22,13\%$ to $72,98\%$) in *setup A* and a 236% (from $22,74\%$ to $76,45\%$) in *setup B*.

4 Performance Comparison with the State of the Art

Now that we have shown that the performance of the BoVW model greatly varies depending on the methods used to create the signature, we can compare the baseline and enhanced BoVW implementations with the state of the art. In order to demonstrate that the enhanced BoVW implementation is competitive against most spotting methods, we are going to compare it against method which used the popular George Washington dataset and the H-KWS 2014 Competition benchmark [36] to assess their performance.

4.1 George Washington Dataset

The George Washington dataset has become a de-facto standard to evaluate handwritten recognition and keyword spotting methods. In order to conduct this comparison, we will only focus on segmentation-based methods to focus only on the performance of the word snippet descriptor. Segmentation-free and line-based methods follow a more general approach that is likely to obtain worse results due to processing a larger amount of information or due to errors introduced while locating words in the document image.

Table 3: Summary of the improvements over the baseline BoVW implementation with the gains in performance

	Setup A	**Setup B**	**Cost**
Baseline	22.13%	22.74%	
LLC	25.15% (↑ 13.65%)	26.04% (↑ 14.51%)	Computational complexity
SPM	61.33% (↑ 177.14%)	64.75% (↑ 184.74%)	Descriptor size
Power normalization	68.27% (↑ 208.50%)	72.20% (↑ 217.50%)	None
Codebook size	71.31% (↑ 222.23%)	74.97% (↑ 229.68%)	Descriptor size
Descriptor sampling	72.98% (↑ 229.78%)	76.45% (↑ 236.19%)	Computational complexity

Table 4: Comparison of the performance attained by the system using the baseline and final BoVW configurations against the results reported by each work. The methods in the first half are exemplar-based methods while second half methods are learning-based.

Reference	**Experimental Setup**	**Originally Reported**	**Baseline BoVW**	**Enhanced BoVW**	**Measure**
Example-based methods					
Rath and Manmatha [37]	10 good quality pages (2381 queries).	40.9%	28.1%	77.2%	mAP
Rothfeder et al. [45]	10 good quality pages (2381 queries).	36.2%	28.1%	77.2%	mAP
Kovalchuk et al. [21]	Same configuration as *setup B*	66.3%	22.7%	76.5%	mAP
Wang et al. [56]	Same configuration as *setup B*	17.5%	22.7%	76.5%	mAP
Howe [15]	4-folds: 3 train and 1 test folds. All non-stop words used as queries.	93.4% 78.9%	55.0% 19.0%	91.8% 79.0%	Mean Precision P@R=100%
Learning-based methods					
Howe et al. [16]	20-folds: 19 train and 1 test fold.	79.5%	38.5%	81.9%	mAP
Rodríguez-Serrano and Perronnin [42]	5-folds: 1 train, 1 validation and 3 test folds.	53.1%	23.6%	74.0%	mAP
Liang et al. [24]	5-folds: 4 train and 1 test folds. 38 words are selected as queries.	67.0%	39.9%	84.5%	mAP at rank 10
Almazán et al. [4]	5-folds: 1 train, 1 validation, 3 test folds. Words in the test set are used as queries.	85.7%	24.0%	74.3%	mAP

Although the George Washington dataset is widely used, there is not an standard experimental setup, and each work adapts it to the needs of their proposed algorithm. For instance, learning based algorithm usually use cross-validation to avoid evaluating the method on the same data used to fit their model. This reduces the amount of queries since query words must appear both in train and test folds. Also, the number of distractors is reduced as the number of putative results is trimmed. These changes make that a direct comparison between methods is not possible. Therefore, we have recalculated the results obtained by the proposed method employing the experimental setup used in each paper.

A brief summary of the experimental setup and the performance comparisons are shown in Table 4. We can see that all exemplar-based algorithms but the method proposed by Howe [15] do not use cross-validation. In [15], the author compares his method with the learning-based method proposed by Frinken et al. in [13], hence the use of cross-validation. Also, most works use mAP to asses their performance, only Liang et al. [24] and Howe [15] use other measures. In [24] the mAP is calculated only using the ten best results of each query. In [15], the author first calculates the mean of the precision and recall curves for all the queries and then reports the area under this curve and the precision at full recall. Finally, learning-based methods use the training set as queries, except the work by Almazán et al. [4]. In this work, the authors use the test set as a completely new database so that both query and indexed images have not been seen in the training phase of the algorithm.

In the comparison table, we can see that the obtained results using the baseline BoVW implementation are significantly worse than the compared works. Only in Wang et al. [56] the baseline implementation obtains a better result. On the other hand, the results attained by the system when using the enhanced BoVW implementation are significantly better than most of the compared works. The proposed BoVW signature is only outperformed by the method proposed by Almazán et al. [4] while Howe [15] have comparable results. It is worth to note, that the method from [4] use a Canonical Correlation Analysis step over a BoVW signature, aimed at finding correlations between visual words and word transcriptions. Obviously, the integration of machine learning techniques over BoVW representations is expected to produce better results than a simple distance among descriptors [2]. Concerning the method by Howe [15], we have to consider the computational complexity of the keyword spotting system. The vectorial nature of BoVW allows to apply standard indexation techniques for an efficient retrieval. In addition, [15] needs an alignment step to compute the similarity between the query and the document's words.

4.2 H-KWS 2014 Competition

The H-KWS 2014 [36] is a recently proposed benchmark dataset to compare the advances in keyword spotting. It analyzes both segmentation-based and segmentation-free algorithms using performance measures frequently found in the literature. This benchmark is composed by the Bentham and Modern datasets. The Bentham dataset is a collection of 50 images written by Jeremy Bentham himself as well as his secretarial staff. This collection is similar to the George Washington dataset in the sense that the calligraphic differences between different instances of the same word are minimal. The Modern dataset is a collection of 100 handwritten pages written by several writers. The writers were asked to copy a text written in English, German, French or Greek. Therefore, this dataset has a high calligraphic variety and it uses different scripts.

The comparison between the results obtained by the proposed basic and enhanced configurations and the methods which participated in the segmentation-based track of the H-KWS 2014 competition are shown in Table 5. The results of this table have been obtained using the evaluation tool provided with the benchmark[1]. As we have seen in the George Washington comparison, Kovalchuk et al. [21] and Howe [15] are exemplar-based while Almazán et al. [4] is a learning-based algorithm. This algorithm is trained using the annotations of George Washington dataset while creating the model for the Bentham dataset and using the IAM dataset for the Modern dataset.

In Table 5, we can see that the baseline configuration obtains rather bad results whereas the enhanced configuration is competitive when compared with the other methods. Specifically, looking at the mAP indicator, the enhanced configuration only obtains slightly better results than Howe [15] in the Bentham dataset while in the Modern dataset it is only surpassed by Almazán et al. [4].

The results obtained in both comparisons stress the fact that the use of simple improvements of the BoVW signatures can lead to a great boost in performance of keyword spotting systems and that it is possible to attain better results than more complex solutions.

5 Conclusions

In this paper we have studied the effects of different BoVW representations for a handwritten word spotting task. Although the use of BoVW has gained attention as a way to represent segmented handwritten words, most of the literature still uses a basic implementation of the BoVW framework, neglecting the latest improvements of such method.

We have reviewed in this paper the improvements that we believe are more suitable for word representation and seen that applying them can lead to a huge boost on the spotting performance of the system. Some of those improvements have in addition no extra or negligible cost in the whole representation, such as using sparse coding instead of hard-assignment or performing a power normalization to each bin of the final descriptor.

Overall, the most important increase in performance came from the use of spatial pyramids, specifically when selecting a configuration that split the handwritten words across the horizontal axis. We believe that such performance boost comes from the fact that this SPM configuration led the descriptor to encode sequential information of the word, i.e. which character comes before another, mimicking the information that is encoded in sequential word representations, but while preserving the advantage of holistic word representations.

Acknowledgements This work has been partially supported by the Spanish Ministry of Education and Science under projects TIN2011-25606 (SiMeVé), and TIN2012-37475-C02-02, by the People Programme (Marie Curie Actions) of the Seventh Framework Programme of the European Union (FP7/2007-2013) under REA grant agreement no. 600388, and by the Agency of Competitiveness for Companies of the Government of Catalonia, ACCIÓ.

References

1. Adamek, T., O'Connor, N., Smeaton, A.: Word matching using single closed contours for indexing handwritten historical documents. International Journal on Document Analysis and Recognition **9**(2–4), 153–165 (2007)
2. Aldavert, D., Rusiñol, M., Toledo, R., Lladós, J.: Integrating visual and textual cues for query-by-string word spotting. In: Proceedings of the International Conference on Document Analysis and Recognition, pp. 511–515 (2013)
3. Almazán, J., Gordo, A., Fornés, A., Valveny, E.: Efficient exemplar word spotting. In: Proceedings of the British Machine Vision Conference, pp. 67.1–67.11 (2012)
4. Almazán, J., Gordo, A., Fornés, A., Valveny, E.: Handwritten word spotting with corrected attributes. In: Proceedings of the International Conference on Computer Vision, pp. 1017–1024 (2013)
5. Ataer, E., Duygulu, P.: Matching ottoman words: an image retrieval approach to historical document indexing. In: Proceedings of the International Conference on Image and Video Retrieval, pp. 341–347 (2007)

[1] H-KWS 2014 competition homepage: `http://vc.ee.duth.gr/h-kws2014/`

Table 5: Comparison of the performance attained by the system using the baseline and enhanced BoVW configurations with the methods that participated in the H-KWS 2014 competition.

	Bentham Dataset				Modern Dataset			
Method	P@5	MAP	NDCG (Binary)	NDCG	P@5	MAP	NDCG (Binary)	NDCG
G1 (Kovalchuk et al. [21])	0.738	0.524	0.742	0.762	0.588	0.338	0.611	0.612
G2 (Almazán et al. [4])	0.724	0.513	0.744	0.764	0.706	0.523	0.757	0.757
G3 (Howe [15])	0.718	0.462	0.638	0.657	0.569	0.278	0.484	0.485
Baseline	0.491	0.292	0.565	0.578	0.231	0.091	0.349	0.350
Enhanced	0.629	0.465	0.707	0.723	0.619	0.389	0.680	0.681

6. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: An evaluation of recent feature encoding methods. In: Proceedings of the British Machine Vision Conference, pp. 76.1–76.12 (2011)
7. Cinbis, R., Verbeek, J., Schmid, C.: Image categorization using fisher kernels of non-iid image models. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, pp. 2184–2191 (2012)
8. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Proceedings of the ECCV International Workshop on Statistical Learning in Computer Vision, pp. 1–22 (2004)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005)
10. Dovgalecs, V., Burnett, A., Tranouez, P., Nicolas, S., Heutte, L.: Spot it! finding words and patterns in historical documents. In: Proceedings of the International Conference on Document Analysis and Recognition, pp. 1039–1043 (2013)
11. Fernández, D., Lladós, J., Fornés, A.: Handwritten word spotting in old manuscript images using a pseudo-structural descriptor organized in a hash structure. In: Pattern Recognition and Image Analysis, *Lecture Notes on Computer Science*, vol. 6669, pp. 628–635 (2011)
12. Fischer, A., Keller, A., Frinken, V., Bunke, H.: Lexicon-free handwritten word spotting using character HMMs. Pattern Recognition Letters **33**(7), 934–942 (2012)
13. Frinken, V., Fischer, A., Manmatha, R., , Bunke, H.: A novel word spotting method based on recurrent neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(2), 211–224 (2012)
14. Fulkerson, B., Vedaldi, A., Soatto, S.: Localizing objects with smart dictionaries. In: Proceedings of the European Conference on Computer Vision, *Lecture Notes in Computer Science*, vol. 5302, pp. 179–192 (2008)
15. Howe, N.: Part-structured inkball models for one-shot handwritten word spotting. In: Proceedings of the International Conference on Documents Analysis and Recognition, pp. 582–586 (2013)
16. Howe, N., Rath, T., Manmatha, R.: Boosted decision trees for word recognition in handwritten document retrieval. In: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 377–383 (2005)
17. Impedovo, S., Mangini, F., Pirlo, G.: A new adaptive zoning technique for handwritten digit recognition. In: Proceedings of the International Conference on Image Analysis and Processing, pp. 91–100 (2013)
18. Khurshid, K., Faureb, C., Vincent, N.: Word spotting in historical printed documents using shape and sequence comparisons. Pattern Recognition **45**(7), 2598–2609 (2012)
19. Kołcz, A., Alspector, J., Augusteijn, M., Carlson, R., Popescu, G.: A line-oriented approach to word spotting in handwritten documents. Pattern Analysis and Applications **3**(2), 153–168 (2000)
20. Konidaris, T., Gatos, B., Ntzios, K., Pratikakis, I., Theodoridis, S., Perantonis, S.: Keyword-guided word spotting in historical printed documents using synthetic data and user feedback. International Journal of Document Analysis and Recognition **9**(2–4), 167–177 (2007)
21. Kovalchuk, A., Wolf, L., Dershowitz, N.: A simple and fast word spotting method. In: Proceedings of the International Conference on Frontiers in Handwriting Recognition (2014)
22. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, pp. 2169–2178 (2006)
23. Leydier, Y., Ouji, A., LeBourgeois, F., Emptoz, H.: Towards an omnilingual word retrieval system for ancient manuscripts. Pattern Recognition **42**(9), 2089–2105 (2009)
24. Liang, Y., Fairhurst, M., Guest, R.: A synthesised word approach to word retrieval in handwritten documents. Pattern Recognition **45**(12), 4224–4236 (2012)
25. Likforman-Sulem, L., Zahour, A., Taconet, B.: Text line segmentation of historical documents: A survey. International Journal on Document Analysis and Recognition **9**(2–4), 123–138 (2007)
26. Lladós, J., Rusiñol, 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**(5), 1263,002.1–1263,002.25 (2012)
27. Lladós, J., Sánchez, G.: Indexing historical documents by word shape signatures. In: Proceedings of the International Conference on Document Analysis and Recognition, pp. 362–366 (2007)
28. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
29. Madhvanath, S., Govindaraju, V.: The role of holistic paradigms in handwritten word recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(2), 149–164 (2001)
30. Manmatha, R., Han, C., Riseman, E.: Word spotting: a new approach to indexing handwriting. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 631–637 (1996)
31. Manmatha, R., Rothfeder, J.: A scale space approach for automatically segmenting words from historical handwritten documents. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(8), 1212–1225 (2005)

32. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: Proceedings of the European Conference on Computer Vision, *Lecture Notes in Computer Science*, vol. 3954, pp. 490–503 (2006)
33. Papandreou, A., Gatos, B., Louloudis, G.: An adaptive zoning technique for efficient word retrieval using dynamic time warping. In: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, pp. 147–152 (2014)
34. Perronnin, F., Sanchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Proceedings of the European Conference on Computer Vision, *Lecture Notes in Computer Science*, vol. 6314, pp. 143–156 (2010)
35. Plamondon, R., Srihari, S.: Online and off-line handwriting recognition: a comprehensive survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(1), 63–84 (2000)
36. Pratikakis, I., Zagoris, K., Gatos, B., Louloudis, G., Stamatopoulos, N.: ICFHR 2014 competition on handwritten keyword spotting (H-KWS 2014). In: Proceedings of the International Conference on Frontiers in Handwriting Recognition, pp. 814–819 (2014)
37. Rath, T., Manmatha, R.: Word image matching using dynamic time warping. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, pp. 521–527 (2003)
38. Rath, T., Manmatha, R.: Word spotting for historical documents. International Journal on Document Analysis and Recognition **9**(2–4), 139–152 (2007)
39. Rath, T., Manmatha, R., Lavrenko, V.: A search engine for historical manuscript images. In: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 369–376 (2004)
40. Rodríguez-Serrano, J., Perronnin, F.: Local gradient histogram features for word spotting in unconstrained handwritten documents. In: Proceedings of the International Conference on Frontiers in Handwriting Recognition, pp. 7–12 (2008)
41. Rodríguez-Serrano, J., Perronnin, F.: Handwritten word-spotting using hidden Markov models and universal vocabularies. Pattern Recognition **42**(9), 2106–2116 (2009)
42. Rodriguez-Serrano, J., Perronnin, F.: A model-based sequence similarity with application to handwritten word-spotting. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2108–2120 (2012)
43. Rodríguez-Serrano, J., Perronnin, F., Sánchez, G., Lladós, J.: Unsupervised writer adaptation of whole-word HMMs with application to word-spotting. Pattern Recognition Letters **31**(8), 742–749 (2010)
44. Rothacker, L., Rusiñol, M., Fink, G.: Bag-of-features hmms for segmentation-free word spotting in handwritten documents. In: Proceedings of the International Conference on Document Analysis and Recognition, pp. 1305–1309 (2013)
45. Rothfeder, J., Feng, S., Rath, T.: Using corner feature correspondences to rank word images by similarity. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop, p. 30 (2003)
46. Rusiñol, M., Aldavert, D., Toledo, R., Lladós, J.: Efficient Segmentation-free Keyword Spotting in Historical Document Collections. Pattern Recognition **48**(2), 545–555 (2015)
47. Rusiñol, M., Aldavert, D., Toledo, R., Lladós, J.: Browsing heterogeneous document collections by a segmentation-free word spotting method. In: Proceedings of the International Conference on Document Analysis and Recognition, pp. 63–67 (2011)
48. Rusiñol, M., Lladós, J.: Word and symbol spotting using spatial organization of local descriptors. In: Proceedings of the IAPR Workshop on Document Analysis System, pp. 489–496 (2008)
49. Sankar, P., Jawahar, C., Manmatha, R.: Nearest neighbor based collection ocr. In: Proceedings of the IAPR Workshop on Document Analysis Systems, pp. 207–214 (2010)
50. Sankar, P., Manmatha, R., Jawahar, C.: Large scale document image retrieval by automatic word annotation. International Journal on Document Analysis and Recognition **17**(1), 1–17 (2014)
51. Shekhar, R., Jawahar, C.: Word image retrieval using bag of visual words. In: Proceedings of the IAPR Workshop on Document Analysis Systems, pp. 297–301 (2012)
52. Shekhar, R., Jawahar, C.: Word image retrieval using bag of visual words. In: Proceedings of the Document Analysis Systems Workshop, pp. 297–301 (2013)
53. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proceedings of the International Conference on Computer Vision, pp. 1470–1477 (2003)
54. Srihari, S., Ball, G.: Language independent word spotting in scanned documents. In: Digital Libraries: Universal and Ubiquitous Access to Information, *Lecture Notes on Computer Science*, vol. 5362, pp. 134–143 (2008)
55. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, pp. 3360–3367 (2010)
56. Wang, P., Eglin, V., Largeron, C., Lladós, J., Fornés, A., Garcia, C.: A novel learning-free word spotting approach based on graph representation. In: Proceedings of the IAPR Workshop on Document Analysis System (2014)
57. van der Zant, T., Shoemaker, L., Haak, K.: Handwritten-word spotting using biologically inspired features. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(11), 1945–1957 (2008)
58. Zhang, X., Tan, C.: Segmentation-free keyword spotting for handwritten documents based on heat kernel signature. In: Proceedings of the International Conference on Document Analysis and Recognition, pp. 827–831 (2013)
59. Zhu, Q., Yeh, M.C., Cheng, K.T., Avidan, S.: Fast human detection using a cascade of histograms of oriented gradients. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, pp. 1491–1498 (2006)