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

Aligning Codebooks for Near Duplicate Image Detection

Sebastiano Battiato, Giovanni Maria Farinella,
Giovanni Puglisi, Daniele Ravì

Received: date / Accepted: date

Abstract The detection of near duplicate images in large databases, such as the ones of popular social networks, digital investigation archives, and surveillance systems, is an important task for a number of image forensics applications. In digital investigation, hashing techniques are commonly used to index large quantities of images for the detection of copies belonging to different archives. In the last few years, different image hashing techniques based on the Bags of Visual Features paradigm appeared in literature. Recently, this paradigm has been augmented by using multiple descriptors (e.g., Bags of Visual Phrases) in order to exploit the coherence between different feature spaces. In this paper we propose to further improve the Bags of Visual Phrases approach considering the coherence between feature spaces not only at the level of image representation, but also during the codebook generation phase. Also we introduce a novel image database specifically designed for the development and benchmarking of near duplicate image retrieval techniques. The dataset consists of more than 3300 images depicting more than 500 different scenes having at least 3 real near duplicates. The dataset has a huge variability in terms of geometric and photometric transformations between scenes and their corresponding near duplicates. Finally, we suggest a method to compress the proposed image representation for storage purposes. Experiments show the effectiveness of the proposed near duplicate retrieval technique, which outperforms the original Bags of Visual Phrases approach.

Keywords Image Forensics · Near Duplicate Images · Image Retrieval · Bags of Visual Words · Bags of Visual Phrases · Codebooks Alignment

Sebastiano Battiato, Giovanni Maria Farinella, Giovanni Puglisi, Daniele Ravì
Image Processing Laboratory, University of Catania, Department of Mathematics and Computer Science,
Viale A. Doria 6, Catania, 95125, IT
E-mail: {battiato, gfarinella, puglisi, ravi}@dmi.unict.it

1 Introduction and Motivations

Image Forensics is a science which, among the other questions, aims to answer the following one during investigation: is the image under consideration contained in a specific digital archive? The increasing use of low cost imaging devices and the availability of large databases of digital photos makes the near duplicate image retrieval (NDIR) task a common activity for a number of applications. In particular, NDIR in large databases (such as popular social networks, collections of surveillance images and videos, or digital investigation archives) is a key ingredient for different forensics activities.

During digital investigation (e.g., for copyright violation, child abuse, etc.), classic hashing techniques (e.g., MD5 [1], SHA1 [2], etc.) are commonly used to index large quantities of images in order to detect copies in different archives. However, these methods are unsuitable to find altered copies, even in case of slight modifications (e.g., near duplicates). Indeed, classic hashing techniques usually fail because just a small change in the image (even a single bit) will, with overwhelming probability, results in a completely different hash code. For example, two images depicting a scene of crime are perceptually identical under small viewpoint changes, partial occlusion, and/or low photometric distortions, but their hash code is completely different when a classic hashing approach is used to check their similarity. In order to cope with all related problems, robust hashing techniques based on image content must be developed: perceptually identical images in terms of content should have the same (or at least very similar) hash value with high probability, while perceptually different images should have independent hash values.

Most of the near duplicate detection techniques based on image content exploit the bag of visual word approach to build the image signature [6, 17, 19, 35, 53]. A problem of the bag-of-visual-word based methods is related to the ambiguity of some generated visual words [36, 37]. On the other hand, since different descriptors represent different aspects of a local region, there is no single descriptor which is superior to the others [8]. Here we propose to combine different descriptors through an alignment procedure based on clusters correspondence (i.e., number of shared local regions). The advantage of the proposed codebook alignment method is related to the enforcement of the coherence across multiple descriptors in order to capture different aspects of the considered local region (e.g., shape, texture, etc.) and hence reduce both, the visual word ambiguity and the quantization error in the visual codebook generation [6, 38]. The different aspects of a local regions are captured by the alignment during the codebook generation in the sense that the local regions falling in the intersection of two aligned clusters agree with respect to both descriptors, whereas the others agree just with one descriptor and not with the other. Taking into account such peculiarity, we split the clusters of each feature domain obtaining new codebook prototypes which consider the intersecting part of the aligned clusters, as well as the part which not intersect. Since, by using multiple descriptors there is an overhead in terms of storage of the representations of the images, and considering that image datasets are becoming more and more popular and huge (i.e., Facebook proceeds at a rate of about 22,000 uploads per minute), we also propose a method to compress the image representation by maintaining performances in terms of near duplicate image detection accuracy.

The remainder of the paper is organized as follows: Section 2 gives a brief survey of the related work. Section 3 describes the proposed model. In Section 4 the method to compress the image descriptors is suggested. The dataset built for testing purposes is described in Section 5, whereas Section 6 details the experimental settings and reports the obtained results. Finally, conclusions and avenues for further research are given in Section 7.

2 Related Works

Recently, some commercial approaches for robust content based hashing methods have been proposed for photos (PhotoDNA [3]) and videos (Videntifier [4]). These techniques make use of the recent developments in the field of Near Duplicate Image (NDI) retrieval. Note that there is no agreement on the technical definition of near-duplicates (see [40] for an in-depth discussion). The definition of near duplicate depends on the degree of variability (photometric and geometric) that is considered acceptable for each particular application. Some approaches [5] consider as NDI, images obtained by slightly modifying the original ones through common transformations such as changing contrast or saturation, scaling, cropping, etc. Other techniques [6] consider as NDI, images of the same scene but with different viewpoint and illumination. A drawback in testing near duplicate retrieval approaches is that usually near duplicate images used in the experiments are synthetically generated from a set of images or correspond to different frames of a video, hence there is an high correlation in terms of visual content, and there is no variability in terms of resolution and compression. To better evaluate the different algorithms it is needed a database composed by images depicting the same scene and/or subject whose have been acquired by different cameras, with different viewpoint, luminance condition, and variability in terms of background.

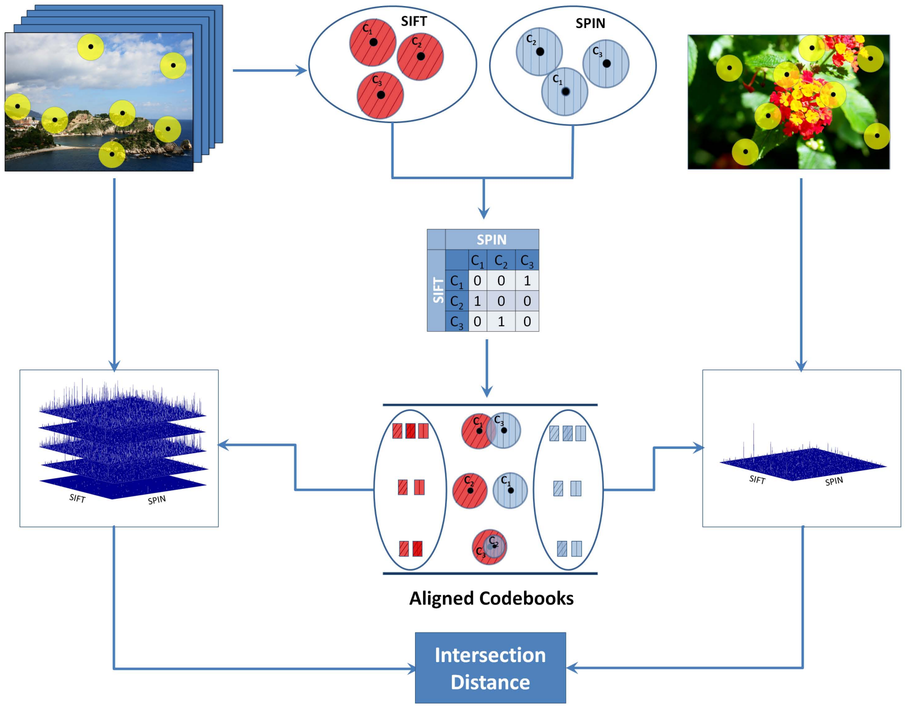
In the last few years, different image hashing techniques have been proposed in literature to cope with image retrieval and near-duplicate image detection problems. Most of these techniques are based on the Bags of Visual Words paradigm (BoVW) [7, 48] to build a holistic representation of the images.

Ke et al. [5] detected near-duplicate images by employing local descriptors [8] extracted on interest points [9] to represent and match images under several transformations. They used a hash-based indexing technique to efficiently search into the image databases, and also applied an optimized storage layout to further improve efficiency. Chum et al. [10] proposed two novel image similarity measures for image indexing through local feature descriptors and enhanced min-Hash techniques. The authors of [11] introduced a method to combine visual words with geometric information to improve hashing-based image retrieval and object detection, obtaining a novel algorithm (called Geometric min-Hash) which shows significant advantages against geometrical deformations and occlusions. Cheng et al. [12] considered local dependencies among descriptors both in scale and space, and encoded not only visual appearance but also their scale and space co-occurrence. Moreover they built SuperNodes that embody the neighbor information to speed up the retrieval execution time. Wu et al. [53] proposed a novel scheme to exploit geometrical constraints by spatially grouping features to improve the retrieval precision. Spatial verification stage to re-rank the results with the bag of-words model have been also exploited by Philbin et al. in [54]. Wang et al. [13] combined appearance-based and keypoint-based methods. The algorithm is able to extract and match keypoints from images by discarding outliers through a voting procedure based on the affine invariant ratio of normalized lengths. Later, to further validate the correspondences, the algorithm compares the color histograms of the corresponding areas which have been previously identified by the matched points.

In [14], Xu et al. proposed a two stage method based on the Spatial Pyramid Matching (SPM) technique [15] and image blocks alignment through linear programming [16]. The aim of the cascade is to deal with spatial shifts and scale variation; both transformations frequently occur between frames of a video. Since there is an increasing interest in the scalability of the Bag-of-Words based near duplicate visual search paradigm, a method to parallelize the near duplicate visual search architecture to index images over multiple servers have been proposed by Rongrong et al. in [55]. Near duplicate image retrieval based on BoVW paradigm has been also exploited for the annotation of web videos as addressed by Zhao et al. in [17,18]. Taking into account their previous work [19], the authors extract keypoints from keyframes and generate a visual dictionary by using a clustering algorithm. Each keyframe is then described by a BoVW representation. Moreover, to speed up the keyframe retrieval, inverted file indexing plus Hamming embedding is employed. A re-rank strategy based on a weak geometric consistency checking is also proposed to improve the overall performance of the system. The final similarity of a video is obtained considering both the scores of keyframes and their temporal consistency with respect to the query video. The memory usage and query-response time are two of the main issues in the retrieval task. The problem of compressing the visual codebook to better handle with storage and retrieval complexity has been studied by Rongrong et a. in [52].

In a recent study of Hu et al. [6], the BoVW paradigm has been augmented by using multiple descriptors (Bags of Visual Phrases) to exploit the coherence between different feature spaces in which local image regions are described. Specifically, to reduce the amount of false matchings in the BoVW model the authors of [6] introduced the coherent phrase model. In this model, a local image region (i.e., the patch surrounding the local interest point [9]) is described by a visual phrase of multiple descriptors instead of a visual word of a single descriptor. In the Bags of Visual Phrases approach, both feature (local regions are described by descriptors of different types) and spatial coherence (multiple descriptors are obtained from local areas at different sizes) are taken into account.

To further improve the Bags of Visual Phrases model, taking into account our preliminary work [20], we propose to exploit the coherence between feature spaces not only in the image representation, but also during the generation of codebooks. This is obtained by aligning the codebooks of different descriptors to produce a more significant quantization of the involved spaces of descriptors, which leads to a more distinctive representation. In particular, to reduce the amount of false matchings, instead of separately obtain the codebooks corresponding to the different feature spaces as proposed in [6], we generate the final codebooks taking into account the correspondence of the clusters of the involved spaces of descriptors to further enforce feature correspondence. To properly perform tests, a new image database of near duplicate images has been built by collecting images from Flickr [21] and private collections. The dataset contains 3148 images of 525 different scenes which have from 3 to 34 real near duplicates. Finally, a method to compress the image representation to be stored for near duplicate purposes is suggested. The experiments performed on the aforementioned dataset show the effectiveness of the proposed approach, which obtains a good margin of performances with respect to the approach described in [6].



	SPIN C1	SPIN C2	SPIN C3
SIFT C1	0	0	1
SIFT C2	1	0	0
SIFT C3	0	1	0

Fig. 1 Proposed Bags of Visual Phrases with codebooks alignment. First a set of keypoints are extracted from a training dataset of images by using a local detector (Hessian-Laplace in our experiments). Each local keypoints is then described by two different descriptors (SIFT [24], SPIN [25] in our experiments) and clustering is performed separately in these two feature spaces. A similarity matrix between pairs of clusters belonging to the two partitions is obtained counting the number of elements (local image regions) they share. The Hungarian algorithm is then used to find the best assignment for the cluster correspondence problem which is encoded in the similarity matrix. The obtained cluster correspondences are then used to create two novel vocabularies where visual words are generated considering the centroids relative to both common and uncommon elements between aligned clusters. The training set images are then represented by using 2D histograms of co-occurrence of visual words related to the generated codebooks. When a query is performed on the training dataset, the test image is represented by using the codebooks obtained in the training phase. Test image representation is compared with those of the training images by using the intersection distance. Finally, the training image corresponding to the lowest distance is selected as the output of the query.

3 Proposed Model

Most of the image hashing techniques for near-duplicate image detection problems typically represent images through feature vectors encoding color, texture, and/or other visual cues such as corners, edges or local interest points [8, 9, 24, 25, 41–46]. These information are automatically extracted using several algorithms and then represented by many different local descriptors. Most of these techniques are based on the Bags of Visual Words paradigm (BoVW) [7] to build a global representation of the visual content within the images. The basic idea is to consider images as visual documents composed of repeatable and distinctive visual elements, which are comparable to the words in texts. Indeed, the BoVW originates in the text categorization community [22] where it was used to describe documents by how many words (belonging to a pre-built vocabulary) occur within them. Each word embracing a semantic meaning, has an inherent set of topics where it is used more often than others. To exploit this model in computer vision and multimedia, a vocabulary of distinctive patterns, usually called "visual words", is built through a clustering approach from a set of local descriptors [8, 24, 25, 44–46] extracted in correspondence of interest points [9, 41–43] which have been previously detected on images of a training database. A local descriptor encodes properties of the region surrounding the interest point in the image from which have been generated. Hence, the "visual words" obtained by clustering the training set of local descriptor are used to identify properties, structures and textures present in the images whose are finally described as an unordered set (a bag) of "visual words". Specifically, each image is represented as a normalized histogram whose bins correspond to "visual words" of the built codebook. Since the bag of visual words description is compact, it is suitable to represent huge image databases.

The proposed approach is built by taking into account the coherent phrase model introduced in [6], where the BoVW paradigm has been augmented by using multiple descriptors (called Bags of Visual Phrases Model) to exploit the coherence between different feature spaces (i.e., local descriptors) in which the local image regions corresponding to interest points are described. To further improve the Bags of Visual Phrases model, we propose to exploit the coherence between feature spaces (i.e., local descriptors) not only in the image representation step (e.g., using a two dimensional distribution of co-occurrence of visual words of codebooks corresponding to two different feature spaces), but also during the generation of codebooks. This is obtained by aligning the codebooks of different descriptors to produce a more significant quantization of the involved spaces of descriptors, which leads to a more distinctive image representation. Differently than Hu et al. [6], we do not obtain the final codebooks corresponding to the different feature spaces separately, but we generate the final codebooks taking into account the correspondence of the clusters of the involved spaces of descriptors to further enforce feature correspondence. Specifically, the partitions obtained through the clustering procedure on each descriptor space are further analyzed with respect to the involved local regions in order to find correspondence between clusters of different features spaces. This alignment allows to further improve the Bag of Visual Phrases Model by adding the coherence of different feature spaces also during codebooks generation phase. The approach is formalized in the following.

Let I an image, and M the number of local regions extracted by making use of a local detector [9] or through dense sampling [15, 23]. Each extracted local region r_i, $i = 1, \ldots, M$, is described by H different local descriptors $\boldsymbol{\phi_{ih}}$, $h = 1, \ldots, H$. Each region r_i is then associated to a set of local descriptors [8] $\boldsymbol{\phi_i} = \{\boldsymbol{\phi_{i1}}, \boldsymbol{\phi_{i2}}, ..., \boldsymbol{\phi_{iH}}\}$. A vocabulary V_h is built for each type of local descriptor, and the different local descriptors $\boldsymbol{\phi_{ih}}$ of a region r_i are hence associated to visual words $\boldsymbol{v_h}$ belonging to the codebook V_h as in the classic BoVW paradigm [7]. This produces a H-tuple $\boldsymbol{p_i} = \{\boldsymbol{v_h} | h \in [1, 2, ..., H]\}$, called "visual phrase", which contains visual words of different feature spaces for each $\boldsymbol{\phi_i}$, $i = 1, \ldots, M$, corresponding to the M local regions detected into the considered image I. Each image is then described by the frequency distribution of visual phrases, called "Bags of Visual Phrases". This model, called "coherent phrase model" [6], incorporates the coherence across multiple descriptors in order to describe different aspects of the appearance of a local region detected within an image.

Our approach augments the coherent phrase model by improving the vocabulary generation step. In [6], H codebooks (one per local descriptor type) are generated separately and independently by using a classical clustering approach on each descriptor space. Then the images are described with a normalized multidimensional histogram in which each bin is related to a visual phrase (e.g., a 2-D distribution by considering two different local descriptors). The underlying rationale is that, although different descriptors encode different properties of a local region, they represent the same local region, hence the clustering, and the visual words belonging to different feature spaces, are in "some way" related. Hence, the coherence among different local descriptors should be exploited also in the vocabulary generation step. The main schema of the proposed approach is summarized in Fig. 1. First, the H different local descriptor spaces are clustered separately and K visual words (cluster centroids) are obtained for each vocabulary V_h (one visual vocabulary per local descriptor) as in the classic BoVW paradigm [7]. The relative ordering of cluster labels in all of the clustering are hence rearranged according to the first one. A $K \times K$ similarity matrix between pairs of clusters belonging to the two partitions is obtained by counting the number of elements (local image regions) they share. The Hungarian algorithm [26] is then used to find the best assignment for the cluster correspondence problem which is encoded in the computed similarity matrix. Therefore, the alignment between clusters of different partitions is thought as a classical resources assignment problem to be solved by a combinatorial optimization algorithm. We choose to exploit Hungarian algorithm since it has been successfully used in Computer Vision to solve different problems which can be seen as a resources assignment problem (e.g., cluster correspondence [28], feature matching [29]). By using the Hungarian method the alignment of the different vocabularies can be done in $O(K^3)$ time. Despite we have used the Hungarian algorithm in our experiments, there are more efficient algorithms that can be used to solve the same problem [27].

The obtained cluster correspondences are used to create H novel vocabularies where visual words are generated considering the centroids relative to both common and uncommon elements between aligned clusters (Fig. 1). Hence three new visual words (cluster means) per descriptor space are generated from two aligned clusters considering the operations of intersection (shared local image regions belonging to the overlap among aligned clusters) and difference (local image regions belonging to the non-overlapped parts of the aligned clusters). Notice that, although Hungarian algorithm aligns all the clusters, some of them can have no common elements (Fig. 1). If two clusters are fully separated, only two new cluster centers will be computed from individual ones. In this last case the two obtained visual words are equal to the original ones.

After building the vocabularies separately on each feature space, these are aligned (as described above) to find coherence between the different spaces based on shared keypoints. After that, each cluster of each vocabulary (which define a visual word in the considered feature space) is splitted in subclusters (defining more than one visual word if the overlap of the aligned clusters is not empty). In this way, the quantization of a descriptor space is refined by taking into account of the quantization obtained in the other feature space. So, the refinement of each vocabulary encodes also information induced from the other vocabulary. This allows to make stronger the discriminativeness of the original Bags of Visual Phrases approach [6] as empirically demonstrated by the experimental results reported in Section 6.

The algorithm described above generates a multidimensional representation of the image under analysis. In particular, starting from the original image, it extracts a set of local feature points, associates them to different descriptors and, by using a precomputed set of vocabularies, creates the final multidimensional normalized histogram.

Considering two descriptors with the associated codebooks consisting of K_1 and K_2 elements respectively, the final image representation is a matrix ($2D$ normalized histogram) of $K_1 \times K_2$ values[1].

4 Image Representation Compression

The compactness of the image representation impacts both in terms of memory storage and computational complexity of the near duplicate detection task [52]. The cost per single image query becomes a critical feature of the overall system as the number of the images stored in the dataset increases. It is then extremely useful to study some approximations of the original representation able to reduce the amount of data to be stored and used during retrieval, without considerably reduce the performance of the overall system. The analysis of the $2D$ histogram representations of the training images shows that the $K_1 \times K_2$ matrices are pretty sparse (see Section 6), hence only a limited set of elements (visual phrases) are actually used to describe the image content. Based on this analysis, we propose a simple and effective compression technique. The most representative and discriminative T visual phrases (i.e., T bins of the $K_1 \times K_2$ matrix, with $T \ll K_1 \times K_2$), together with their IDs (i.e., the number of row and column they belong into the matrix), can be selected to represent the image under analysis. This selection considers all the images of the training dataset on which image queries are performed.

To sum up, when a query is performed on the selected training dataset considering a generic test image I, the following steps are performed:

i) generate the multidimensional histogram Ψ_I of the image I;
ii) for each image J of the training dataset, select its matrix coordinates C_J relative to its most representative and discriminative T visual phrases;
iii) select the elements of the histogram Ψ_I at the coordinates C_J;
iv) for each image J compute the similarity between the compact representation of images I and J;
v) provide as output of the query the image $\widehat{J}$ belonging to the training dataset with the lowest distance from the image I.

It is worth noting that the effectiveness of the proposed approximation depends on the number of selected bins T. To obtain a satisfactory improvement in terms of memory storage and computational load this number should be considerably lower than $K_1 \times K_2$, where K_h is the dimension of the vocabulary V_h. On the other hand few visual phrases (i.e., bins) could be not able to properly discriminate the images belonging to the dataset. A smart selection strategy of the best T bins can be then useful in finding a good trade-off between compression and retrieval performance. Specifically, we employ the statistical measure TF-IDF (term frequency-inverse document frequency) [30] for the selection of the most representative and discriminative visual phrases (i.e., to select the best T bins within the representation matrix). In this way, the importance of the bin is not only related to its frequency in the image representation but also consider the frequency of the bin (i.e., the discriminativeness of the visual phrase composed by a pair of descriptors) with respect to the entire training dataset. In other words, for each image of the training dataset, we select the most representative and discriminative T

[1] Note that at this stage other encoding methods can be used starting from the aligned vocabulary [39].

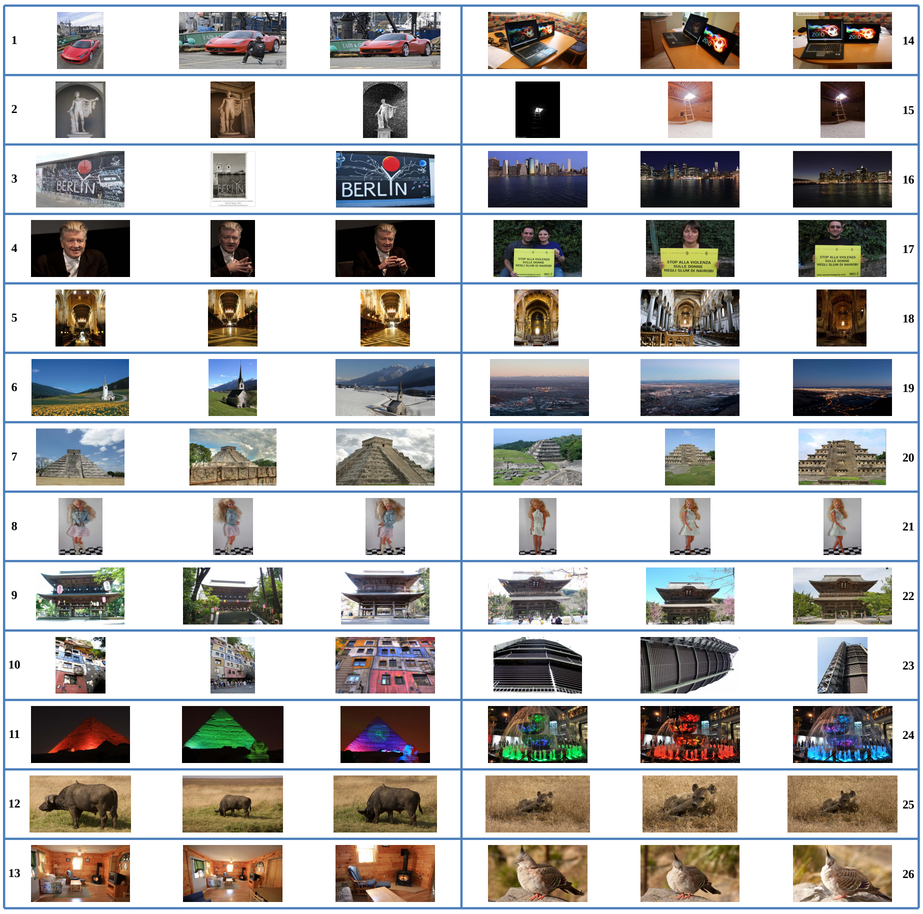


Fig. 2 Examples of 26 different scenes belonging to the considered dataset. For each scene three near duplicates are shown.

visual phrases (e.g., bins) as indicated by the TF-IDF measure. During a comparison of an query image I with an image of the training dataset J only the T visual phrases of the image J which have been selected taking into account the TF-IDF measure are considered.

5 The Experimental Datasets

An image I is considered a near-duplicate of another image J if its content is "similar", according to some defined similarity measures, to the image J. So, the definition of a near duplicate image changes accordingly with the allowed photometric and geometric variations. As in [10], we consider an image I a near-duplicate of another image J if it contains the same scene of J with possibly different photometric and/or geometric variations (e.g., viewpoints changes, illumination and color variations, partial scene, occlusion, different compression and camera acquisition, etc.). The problem addressed here is hence the one to enumerate all the near duplicates of a given query image in a dataset.

In order to test and compare different algorithms for near duplicate image retrieval, a representative dataset should be used. Despite different datasets have been employed in literature for testing purposes, most of them are synthetic [2] [5] or obtained taking into account keyframes of videos [6]. Although synthetic datasets are compliant with the definition of near duplicate given above, they aren't representative of the real variation that can be observed in real near duplicate images (see Fig. 2). On the other hand, datasets built by collecting frames of videos contain near duplicates with no variability in terms of resolution and compression factor. The classic datasets used for image retrieval testing purposes (e.g., CBIR task), such as the one introduced in [32], are not compliant with the aim of near duplicate image retrieval, where the problem is to search for the same scene with possibly different photometric and/or geometric variations, given an image as query.

The above motivations induced us in building and using a new representative dataset for the problem under consideration. In this way we can properly test and compare the proposed augmented version of Bag of Visual Phrase model with respect to the original one [6]. Specifically, a dataset with images acquired by different cameras, in different conditions (e.g., viewpoint, scale, illumination, distance from the subjects, etc.), and high content variability (indoor, outdoor, object, natural scenes, etc.), has been collected from Flickr [21] and from private collections. To this aim, 525 different keywords (e.g., New York, Animal, Car, Church, Computer, Mountains, Landscape, etc.) have been chosen. Each keyword has been then used to retrieve images from Flickr. From the retrieved images a set of near duplicates have been hence manually sampled. Each specific set corresponding to a keyword contains from 3 to 34 near duplicates. The whole dataset contains 3148 images. In Fig. 2 some of the images belonging to the built dataset are shown. Specifically, in the figure are reported three near duplicates of 26 different scenes. As evident by visual inspection, there is a high variability in terms of scenes (outdoor, indoor, close up objects, portraits, archeological sites, buildings, animals, open scenes, etc.) as well as a high variability in terms of geometric and photometric characteristics among near duplicates of the same scene (different point of view, luminance and color variation, zoom, rotation, background variation, etc.). Moreover, different scenes have regions with similar appearances, such as in the case of the scenes with animals (see images of the scenes with number 12 and 25 in Fig. 2) and the ones with Japanese buildings (see images of the scenes with number 9 and 22 in Fig. 2). Differently than classic content based image retrieval task in which, for instance, given an image of the scene numbered as 9 in Fig. 2 all the images of the the scene numbered as 22 are acceptable in terms of visual similarity, in the context of near duplicate image detection this become an unacceptable error. The database was hence built to properly test the challenging task under consideration.

Since near duplicate image detection techniques are usually tested on datasets used in the context of object recognition [6], we have performed tests also considering the UKBench dataset which contains a total of 10200 images of 2550 different objects

[2] We consider a dataset as synthetic when the near duplicates are generated from a set of images (or frames of videos) by using transformations typically available on image manipulation software (e.g., ImageMagick [31]), such as colorizing, contrast changing, cropping, despeckling, downsampling, format changing, framing, rotating, scaling, saturation changing, intensity changing, shearing. To generate near duplicates the basic transformations are usually applied changing the different involved parameters and/or making combination of them.

with four near duplicate images (photometric and/or geometric variations) for each object [33].

6 Experimental Results

In this section the effectiveness of the proposed approach is demonstrated through a number of experiments and comparisons. A first test, conducted on the dataset we built (see previous section), compares our method with respect to the coherent visual phrase model described in [6] and the technique proposed by Zhao et al. in [17, 18]. Note that both [6] and the classic BoVW approaches have been reimplemented at the best of our knowledge whereas the original code provided by the related authors has been used for [17, 18].

To properly evaluate the different methods, the experiments have been repeated three times. At each run the different approaches are executed on the same training and test sets. To this purpose, at each run we have built training and test sets by selecting images at random. Specifically, we have randomly selected one image per each set of near duplicates to build a training set with 525 different scenes, whereas two images per each set of near duplicates have been randomly selected to build the test set. All the parameters involved in the experiments have been learned from the corresponding training sets for each method. The results presented in the following are obtained by averaging the results of all three runs.

For every run, training images have been used for the generation of codebooks. First, local interest points have been detected (Hessian-Laplace [9]). Afterward, two different descriptors have been extracted on each interest point: SIFT [24] and SPIN [25]. Since these descriptors are extracted considering different image properties (gradient orientation (SIFT) and intensity distribution at different distance from the center (SPIN)), they are somewhat complementary, hence can be fruitfully combined. K-means algorithm (K=500 in our tests) has been then used to produce the two independent codebooks corresponding to the two involved descriptors. The two obtained partitions have been aligned with the Hungarian algorithm to generate the new codebooks (see Section 3). Finally, training images have been represented by visual phrases (with a 2D histogram) by considering the new aligned codebooks.

It is worth noting that the proposed procedure for codebook generation creates two novel vocabularies (one for each type of descriptor involved in the experiment) with a higher number of elements with respect to the original ones. Considering, as example, two codebooks of 500 elements, the alignment procedure will produce, in the worst case, two novel vocabularies of 1500 elements for each type of descriptor. In order to reduce the dimension of the final image representation maintaining at the same time good performance, analysis and tests have been performed. In particular, useful hint can be derived from the analysis of the degree of similarity between the clusters associated in the alignment procedure performed through the Hungarian algorithm. As reported in Fig. 3, which have been obtained sorting the aligned clusters with respect to their dissimilarity, after a certain threshold, the aligned clusters cannot be considered “similar”. This means that after a given value the aligned clusters share only few keypoints (or nothing at all) and hence there is not too much coherence among these aligned clusters. The threshold imposed on cluster dissimilarity is chosen taking into account the gradient of the dissimilarity curve. At some point, the gradient of the curve starts to be very small and this fact can be used to set the threshold. Moreover, given a

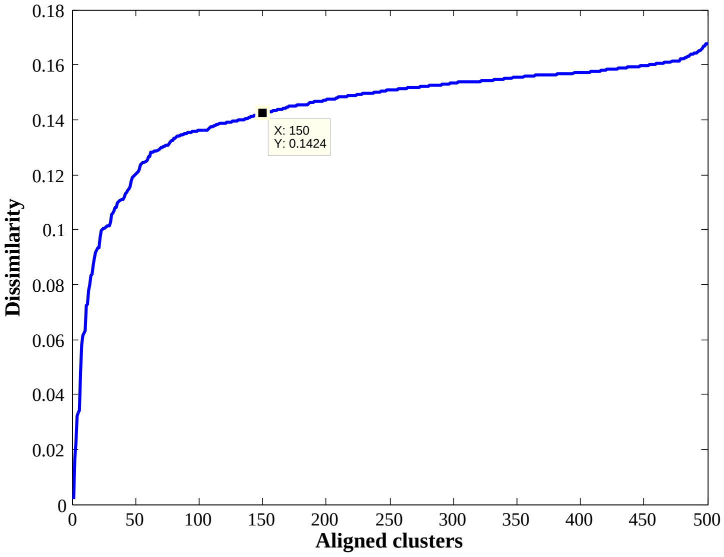


Fig. 3 Sorted dissimilarity of aligned vocabularies. The first 150 aligned pairs of clusters can be considered "similar" in terms of shared keypoints, whereas the others are "dissimilar".

threshold, the number of elements of the aligned vocabulary is propery established. For example, in the case reported in Fig. 3, where a threshold which consider 150 aligned cluster is selected in correspondence of a small gradient, the final number of employed centroids is equal to $150 \times 3 + 350$ for each feature space. The cluster intersections produce 150×3 new visual words for each feature space, whereas the other not aligned 350 clusters produce 350 visual word for each feature space. So considering both, the gradient of the curve and the dimension of the final codebook, the threshold can be fixed. Taking into account the previous analysis, a more compact vocabulary can be hence generated performing the procedure for the generation of aligned codebooks (see Section 3) only for the aligned clusters having a high degree of similarity; for all other "dissimilar" clusters will be retained only the original centroids on the corresponding feature space. The analysis of the dissimilarity curves related to the different three training set considered in our tests, pointed out that the first 150 aligned pairs of clusters can be considered properly aligned (i.e., "similar" in terms of shared keypoints). In this way a final codebook of 800 visual words per descriptor (SIFT and SPIN) has been generated instead of one of 1500. To be fair, the comparisons with the other approaches (Hu et al. [6], BoVW SIFT and BoVW SPIN) have been performed considering codebooks with 800 elements per descriptor independently generated through K-means clustering.

At each run, test images are used to perform queries on the related training dataset. Each test image is represented by a visual phrase histogram obtained considering the aligned codebooks (see Fig. 1). This representation is then used to retrieve images in the training dataset, by means of a similarity function between Bag of Phrases histograms. To cope with partial matching, we use the intersection distance τ defined as follows [34, 47]:

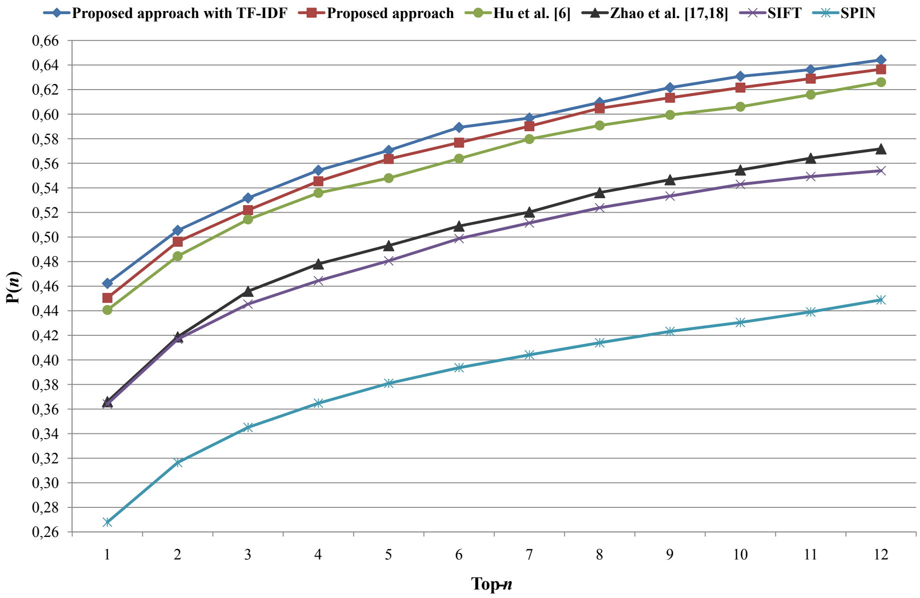


Fig. 4 Top-n NDI retrieval performances comparison on the proposed dataset.

$$\tau(\Psi_I, \Psi_J) = \sum_{p=1}^{P} min(\Psi_p(I), \Psi_p(J)) \tag{1}$$

where Ψ_I, Ψ_J are two visual phrase histograms and $\Psi_p(.)$ is the p^{th} bin of the histogram. Both representation, with and without TF-IDF weighting scheme have been considered and compared.

Each query image has been associated to a list of training images. The retrieval performance has been evaluated with the probability of the successful retrieval $P(n)$ in a number of test queries [6, 49–51]:

$$P(n) = \frac{Q_n}{Q} \tag{2}$$

where Q_n is the number of successful queries according to top-n criterion, i.e., the correct NDI is among the first n retrieved images, and Q is the total number of queries.

The proposed approach has been compared with the original Bags of Phrases approach [6], with the approach proposed in [17, 18], as well as with respect to the classic BoVW approach considering both SIFT and SPIN descriptors. The obtained results are reported in Fig. 4. Both proposed strategy, with and without TF-IDF, outperforms the original Bags of Visual Phrases, the approach proposed in [17, 18], and the classic BoVW model. We also show the precision/recall values at top-n=1 in Table 1. Note that the precision and recall for top-n=1 are equivalent because there is only one correct match for each query. Some visual examples of the first retrieved image on a specific query are reported in Fig. 5. Specifically, for some query images reported in the first column of Fig. 5, the first retrieved images obtained with the proposed approach and the method described in [6] are shown respectively in the second and third

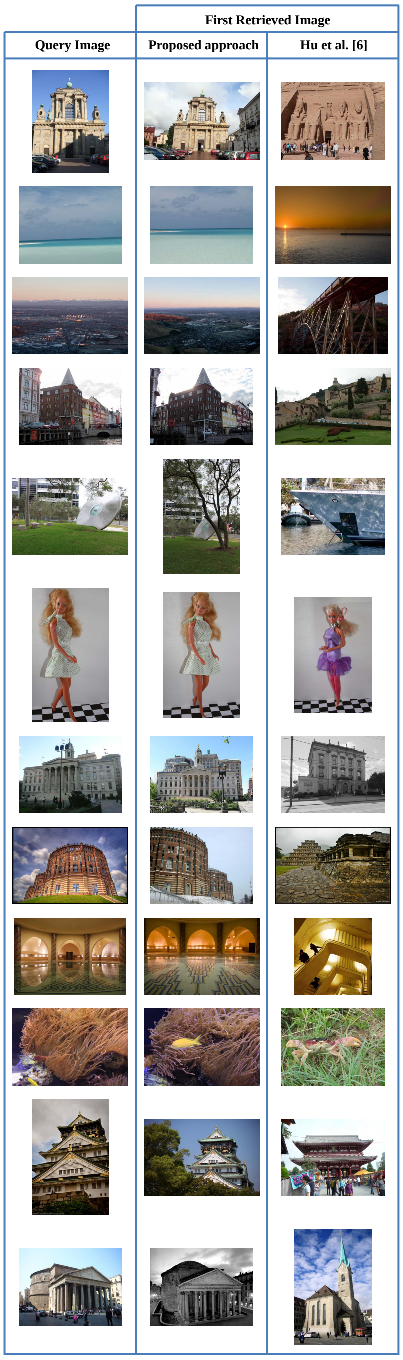


Fig. 5 Some visual examples of the first retrieved image on a specific query. In these examples the proposed approach outperforms the method proposed by Hu et al. [6].

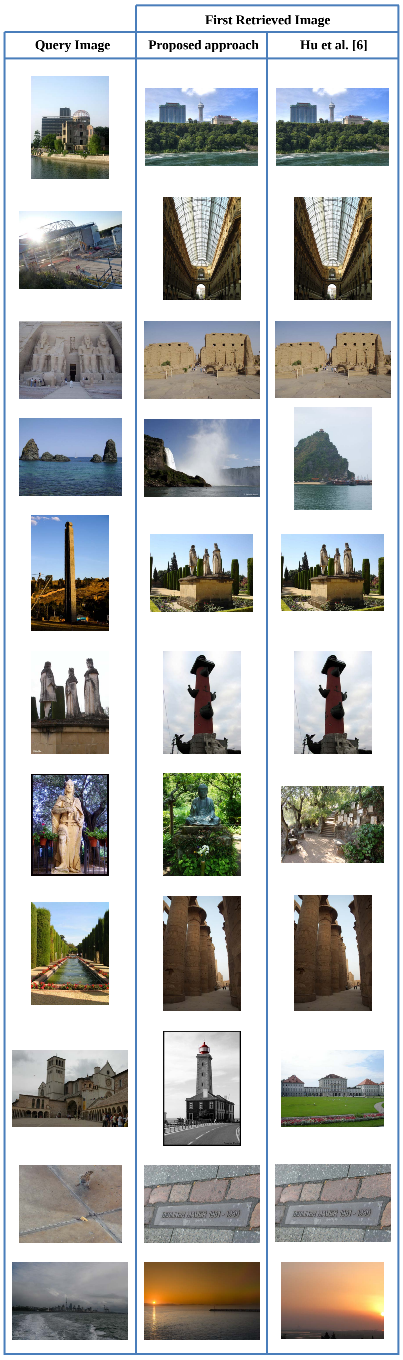


Fig. 6 Some visual examples of the first retrieved image on a specific query. In these examples both, the proposed approaches and the one described in [6] fail.

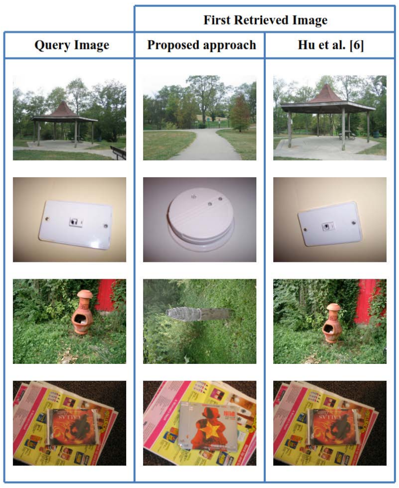


Fig. 7 Some visual examples of the first retrieved image on a specific query. In these examples the method proposed by Hu et al. [6] outperforms the proposed approach.

Table 1 Precision/Recall values on the proposed dataset.

Method	**Precision/Recall**
Proposed approach with TF-IDF	0.4622
Proposed approach	0.4505
Hu et al. [6]	0.4406
Zhao et al. [17, 18]	0.3660
SIFT	0.3641
SPIN	0.2679

columns of Fig. 5. The proposed method is able to detect the corresponding near duplicate within the training set, whereas the technique proposed in [6] retrieves images which aren't a near duplicate of the queries and hence fail the aim. For completeness, further visual examples in which both approaches fail, as well as some examples in which the method of Hu et al. [6] outperforms our approach are reported respectively in Fig. 6 and Fig. 7. As evident from Fig. 6, often both approaches fail in the same way (i.e. selecting the same wrong image).

As already stated in Section 4, some analysis and tests have been performed to compress the image representation in order to speed up the retrieval process and to reduce the amount of data to be stored. Although the image representation is based on a 2D histogram of 800 × 800 elements, only a limited number of bins are actually different from zero. Specifically, from the performed analysis we have observed that the training images are described, on average, by 1700 non-zero elements (with a standard deviation of 1039). This analysis motivated the compression strategy described in Section 4. Taking into account the number of non-zero elements it is possible to guess the number of bins to be used in the image representation. Several tests have been performed to validate the proposed compression strategy and to find a good trade-off between compression and retrieval performance. As shown by Fig. 8, the retrieval performance increases at increasing of the elements used for image representation. Moreover, the results obtained considering 1600 elements are comparable with the ones of the proposed approach without compression in which the overall 800 × 800 elements are involved. The considered number of bins (i.e., 1600) is very close to the number of the non-zero element computed during the aforementioned analysis (i.e., 1700). In this way we are able to obtain a compact image representation without sacrificing the retrieval performance.

It is worth noting that the built dataset is really challenging; in some cases even an human observer could have some difficulties in finding the correct near duplicate image (e.g., compare the scenes marked with number 9 and 22 in Fig. 2). Moreover, in the described near duplicate image retrieval system each of the 525 classes are described by only one image, a design choice that could limit the overall performance of the proposed method, but is realistic, for instance, in the context of forensic science where investigators have only one image example of a criminal scene.

To further confirm the effectiveness of the proposed approach, additional experiments have been performed on the UKBench dataset [33]. This dataset, usually used for object recognition tasks, contains 10200 images of 2550 different objects. Specifically, there are four near duplicate images with photometric and/or geometric variations for each object. In our test, the training dataset has been built randomly selecting one image per class. The remaining images have been then used for testing purposes. The test has been repeated three times and the final results are obtained by averaging the results obtained on each test. As can be easily seen from Fig. 9 and Table 2, also considering this dataset, the proposed approach obtains satisfactory results. Also in this case the proposed approach outperforms the original Bag of Phrases approach [6] obtaining a good margin in terms of performances. The approach proposed in [17, 18] results worst than the original Bag of Phrases method and it is not reported in Fig. 9.

As pointed out in Section 3, in terms of computational complexity the proposed approach has an additional cost due to the alignment of clusters during the codebook generation. On the other hand, this allows a richer description of the images which is reflected in the increasing of the performances with respect to the original Bag of Visual Phrases paradigm [6]. Moreover, the alignment procedure is performed just once during the vocabulary generation and it does not affect the retrieval step in terms of extra costs. Considering the computational complexity during the retrieval task, the description compression proposed in Section 4 helps to reduce both, space and time with to respect the original paradigm [6] by maintaining the performances of the proposed codebooks alignment framework. Moreover, regarding the retrieval task, the proposed technique is comparable with the one proposed in [17, 18] in terms of computational complexity. Indeed, considering vocabularies with size K for the different descriptors, the mapping of each image with M local regions to the related visual vocabularies has computational complexity $\mathrm{O}(MK)$. The time needed to build the visual phrases distribution is $\mathrm{O}(M)$, whereas the compression of the image representation described in Section 4 takes time $\mathrm{O}(T)$. Finally, the similarity between the query and an image belonging to the training dataset has computational complexity $\mathrm{O}(T)$. Hence the overall computational complexity to represent and check a query image with to respect an image into the training dataset is $\mathrm{O}(MK)+\mathrm{O}(M)+\mathrm{O}(T)$. It is worth noting that by employing the compression strategy the complexity in terms of computational power as well as the one related to the memory usage, have been considerably reduced. Specifically, considering a simple 2D matrix representation (without optimized data structure such as sparse matrix) the complexity of the comparisons is $\mathrm{O}(K^2)$ instead of $\mathrm{O}(T)$ where $K^2 \gg T$. Hence the final complexity of the algorithm without compression is $\mathrm{O}(MK)+\mathrm{O}(M)+\mathrm{O}(K^2)$.

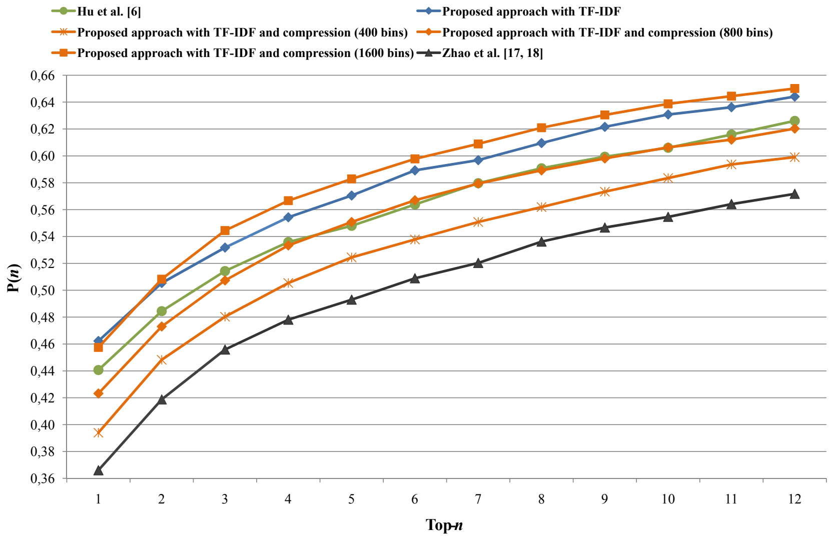


Fig. 8 Top-n NDI retrieval performances of proposed approach with compression on the built dataset. Results are reported at varying of the number of elements involved into the image representation.

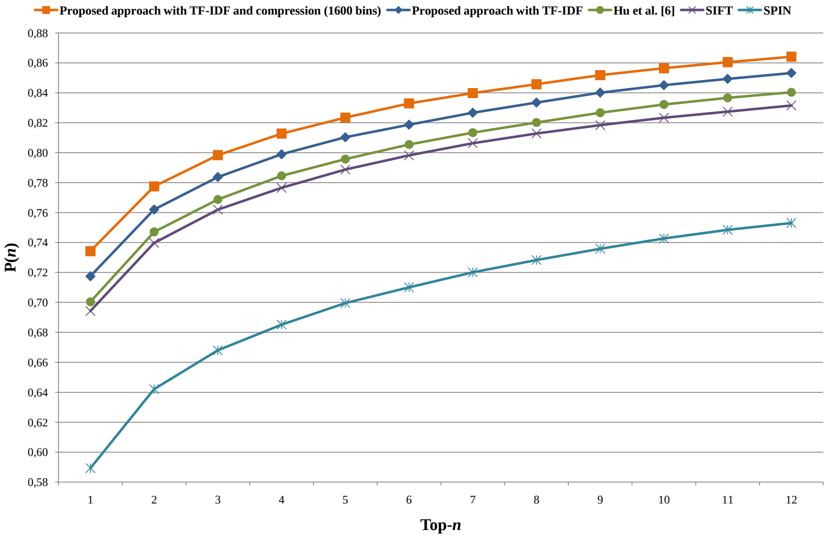


Fig. 9 Top-n NDI retrieval performances of the proposed approach with compression on the UKBench dataset.

Table 2 Precision/Recall values on the UKBench dataset.

Method	Precision/Recall
Proposed approach with TF-IDF (1600 bins)	0.7342
Hu et al. [6]	0.7003

7 Conclusions and Future Works

In this paper we proposed an improvement of the coherent phrase model (Bags of Phrases) originally proposed in [6]. The main contribution of the presented approach is in augmenting the original paradigm by exploiting coherence between different feature spaces also during the codebook generation step. This is achieved through alignment of the feature space partitions obtained from independent clustering. Moreover, a method based on TF-IDF statistical measure to compress the proposed image representation for storage purposes is suggested. Experiments show the effectiveness of the described method on both, a novel and challenging near duplicate image database and a classic benchmark one. Future works will be devoted to extend the proposed alignment methodology to consider multiple ($h > 2$) types of descriptors.

Acknowledgments

The authors would like to thank Giuseppe Claudio Guarnera, Tony Meccio and Rosetta Rizzo for their invaluable help. Part of this work has been performed in the project PANORAMA, co-funded by grants from Belgium, Italy, France, the Netherlands, the United Kingdom, and the ENIAC Joint Undertaking.

References

1. R. L. Rivest, “RFC 1321,” http://tools.ietf.org/html/rfc1321.
2. D. Eastlake and P. Jones, “RFC 3174,” http://tools.ietf.org/html/rfc3174.
3. “Photodna,” http://www.microsoftphotodna.com/.
4. H. Lejsek, H. ÃđormóÃřsdóttir, F. Ásmundsson, K. DaÃřason, Á. Ãđ. Jóhannsson, B. Ãđ. Jónsson, L. Amsaleg, “Videntifier Forensic: large-scale video identification in practice,” *In Proceeding of ACM workshop on Multimedia in forensics, security and intelligence*, pp. 1–6, 2010.
5. Y. Ke, R. Sukthankar, and L. Huston, “Efficient near-duplicate detection and sub-image retrieval,” in *In Proceeding of ACM Multimedia*, pp. 869–876, 2004.
6. Y. Hu, X. Cheng, L.-T. Chia, X. Xie, D. Rajan, and A.-H. Tan, “Coherent phrase model for efficient image near-duplicate retrieval,” *IEEE Transactions on Multimedia*, vol. 11, no. 8, pp. 1434–1445, 2009.
7. R. Szeliski, *Computer Vision: Algorithms and Applications.* Springer, 2010. Available online at http://szeliski.org/Book/
8. K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *IEEE Transactions on Pattern Analysis & Machine Intelligence (PAMI)*, vol. 27, no. 10, pp. 1615–1630, 2005.
9. K. Mikolajczyk and C. Schmid, “Scale & affine invariant interest point detectors,” *International Journal of Computer Vision (IJCV)*, vol. 60, no. 1, pp. 63–86, 2004.
10. O. Chum, J. Philbin, and A. Zisserman, “Near duplicate image detection: min-hash and tf-idf weighting,” in *Proceeding of BMVC*, 2008.
11. O. Chum, M. Perdoch, and J. Matas, “Geometric min-hashing: Finding a (thick) needle in a haystack,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 17–24, 2009.
12. X. Cheng, Y. Hu, and L.-T. Chia, “Exploiting local dependencies with spatial-scale space (s-cube) for near-duplicate retrieval,” *Computer Vision and Image Understanding*, vol. 115, no. 6, pp. 750–758, 2011.
13. Y. Wang, Z. Hou, and K. Leman, “Keypoint-based near-duplicate images detection using affine invariant feature and color matching,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, pp. 1209–1212, 2011.

14. D. Xu, T. J. Cham, S. Yan, L. Duan and S.-F. Chang, "Near Duplicate Identification with Spatially Aligned Pyramid Matching," *IEEE Trans. on Circuits Systems for Video Technology (TCSVT)*, vol. 20, no. 8, pp. 1068-1079, 2010.
15. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2169–2178, 2006.
16. D. Xu and S.-F. Chang, "Visual Event Recognition in News Video using Kernel Methods with Multi-Level Temporal Alignment," in *Proceeding of IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.
17. W.-L. Zhao, X. Wu, and C.-W. Ngo, "On the annotation of web videos by efficient near-duplicate search," *IEEE Transactions on Multimedia*, vol. 12, no. 5, pp. 448–461, 2010.
18. W.-L. Zhao, X. Wu, and C.-W. Ngo, "SOTU: A Toolkit for Efficient Near-duplicate Image/Video & Retrieval/Detection," *Manual for SOTU Version 1.06*, 2011 `http://www.cs.cityu.edu.hk/~wzhao2/sotu.htm`
19. W.-L. Zhao, C.-W. Ngo, "Scale-Rotation Invariant Pattern Entropy for Keypoint-Based Near-Duplicate Detection," *IEEE Transactions on Image Processing*, vol. 18, no. 2, pp. 412–423, 2009
20. S. Battiato, G. M. Farinella, G. C. Guarnera, T. Meccio, G. Puglisi, D. RavÃň, R. Rizzo, "Bags of Phrases with Codebooks Alignment for Near Duplicate Image Detection," in *Proceedings of the International ACM Workshop on Multimedia in Forensics, Security and Intelligence (MiFor 2010), in conjunction with International ACM Multimedia Conference*, pp. 65–70, 2010.
21. "Flickr," http://www.flickr.com/.
22. G. Salton and M. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.
23. S. Battiato, G. M. Farinella, G. Gallo, and D. RavÃň, "Exploiting Textons Distributions on Spatial Hierarchy for Scene Classification," *Eurasip Journal on Image and Video Processing*, Article ID 919367, doi:10.1155/2010/919367, pp. 1–13, 2010.
24. D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
25. A. E. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3d scenes," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 21, no. 5, pp. 433–449, 1999.
26. C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity.* Prentice-Hall, Inc., 1982.
27. R. Jonker and A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems," *Computing*, vol. 38, no. 4, pp. 325–340, 1987.
28. A. Saffari, H. Bischof, "Clustering in a Boosting Framework," *Computer Vision Winter Workshop*, pp. 75–82, 2007.
29. S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. 24, no. 24, pp. 509-âĂŞ521, 2002.
30. G. Salton and C. Buckley,"Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.
31. "ImageMagick," http://www.imagemagick.org.
32. M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval.* New York, NY, USA: ACM, 2008.
33. D. Nistèr and H. Stewènius, "Scalable recognition with a vocabulary tree," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2161–2168, 2006.
34. K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1458–1465, 2005.
35. S. Battiato, G. M. Farinella, E. Messina, G. Puglisi, "Robust Image Alignment for Tampering Detection," *IEEE Transaction on Information Forensics and Security*, Vol. 7, no. 4, pp. 1105–1117, 2012.
36. J.C. van Gemert, C.J. Veenman, A.W.M. Smeulders, and J.-M. Geusebroek,"Visual Word Ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, no. 7, pp. 1271–1283, 2010.

37. J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman,"Discovering Object Categories in Image Collections," in *Proceedings of the International Conference on Computer Vision*, 2005.
38. S. Lazebnik, M. Raginsky, "Supervised Learning of Quantizer Codebooks by Information Loss Minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.31, no.7, pp.1294–1309, 2009.
39. K. Chatfield, V. Lempitsky, A. Vedaldi and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *Proceedings of the British Machine Vision Conference*, 2011.
40. R. De Oliveira, M. Cherubini, and N. Oliver, "Looking at near-duplicate videos from a human-centric perspective," *ACM ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 6, no. 3, pp. 15:1–15:22, 2010.
41. H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *International Journal on Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.
42. E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proceedings of the European Conference on Computer Vision*, pp. 430–443, 2006.
43. J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," in *Proceedings of the British Machine Vision Conference*, pp. 384–393, 2002.
44. S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, no. 4, pp. 509–522, 2002.
45. W. Freeman and E. Adelson, "The Design and Use of Steerable Filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 9, pp. 891–906, Sept. 1991.
46. J. Koenderink and A. van Doorn, "Representation of Local Geometry in the Visual System," *Biological Cybernetics*, vol. 55, pp. 367–375, 1987
47. M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.
48. T. Leung, J. Malik, J., "Recognizing surfaces using three-dimensional textons," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1010–1017 1999.
49. D.-Q. Zhang and S.-F. Chang, "Detecting image near-duplicate by stochastic attributed relational graph matching with learning," in *Proceedings of the ACM Multimedia Conference*, pp. 877–884, 2004.
50. W.-L. Zhao, C.-W. Ngo, H.-K. Tan, and X. Wu, "Near-duplicate keyframe identification with interest point matching and pattern learning," *IEEE Transactions Multimedia*, vol. 9, no. 5, pp. 1037–1048, 2007.
51. J. Zhu, S. C. Hoi, M. R. Lyu, and S. Yan, "Near-duplicate keyframe retrieval by nonrigid image matching," in *Proceedings of the ACM Multimedia Conference*, pp. 41–50, 2008.
52. J. Rongrong, Y. Hongxun, L. Wei, S. Xiaoshuai, T. Qi Tian, "Task-Dependent Visual-Codebook Compression," *IEEE Transactions on Image Processing*, vol.21, no.4, pp. 2282–2293, 2012.
53. Z. Wu, Q. Ke, M. Isard, J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pp. 25–32, 2009.
54. J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2007.
55. J. Rongrong, L.-Y. Duan, J. Chen, L. Xie, H. Yao, W. Gao, "Learning to Distribute Vocabulary Indexing for Scalable Visual Search," *IEEE Transactions on Multimedia*, vol.15, no.1, pp. 153–166, 2013.