Leveraging semantic segmentation for hybrid image retrieval methods

Achref Ouni, Eric Royer, Marc Chevaldonné, Michel Dhome

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

Leveraging semantic segmentation for hybrid image retrieval methods

Achref Ouni · Eric Royer · Marc Chevaldonné · Michel Dhome ·

Received: date / Accepted: date

Abstract Content Based Image Retrieval (CBIR) is the task of finding images in a database that are the most similar to the input query based on its visual characteristics. Several methods from the state of the art based on visual methods (Bag of visual words, VLAD, ...) or recent deep leaning methods try to solve the CBIR problem. In particular, Deep learning is a new field and used for several vision applications including CBIR. But, even with the increase of the performance of deep learning algorithms, this problem is still a challenge in computer vision. In this work, we propose three different methodologies combining deep learning based semantic segmentation and visual features. We show experimentally that by exploiting semantic information in the CBIR context leads to an increase in the retrieval accuracy. We study the performance of the proposed approach on eight different datasets (Wang, Corel-10k, Corel-5k, GHIM-10K, MSRC V1, MSRC V2, Linnaeus, NUS-WIDE)

Keywords Deep learning · Semantic Segmentation · CBIR · Image classification

1 Introduction

The exponential increase in image acquisition and information technology has enabled the creation of large image datasets. Therefore, it is important to create data frameworks to effectively deal with these collected images. In particular, Content-Based Image Retrieval (CBIR) systems offer a solution to quickly find an image in a large amount of data.

CBIR is a fundamental step in many applications and can be used to solve a large variety of tasks. For example, when searching on the web or in a large

Université Clermont Auvergne, CNRS, SIGMA Clermont, Institut Pascal, F-63000 CLERMONT-FERRAND, FRANCE
E-mail: Achref.EL_OUNI@uca.fr

image database, a user can have difficulties to express his need. With a CBIR algorithm, this problem called intention gap can be solved by providing an example image instead of a textual description. CBIR can also be very useful in robotics, where an image from an on board camera can be used for visual localization. The same applies in augmented reality systems and in many other applications.

CBIR is the task of retrieving the images similar to the input query from a dataset based on their contents. A CBIR system (see figure 1) is often based on three main steps: (1) Features extraction (2) Signature construction (3) Retrieved images. The performance of any proposed approach depends on the way an image signature is constructed. Therefore, the construction of image signatures is a key step and the core of a CBIR system. The state of the art mentions two main approaches used to retrieve the closest images: BoVW [57] (Bag of Visual Words) and CNN [58] (Convolutional Neural Networks) descriptors for image retrieval.

Those methods make use of information such as colour, shape and texture. A few authors propose to explicitly take into consideration the semantic information that can be extracted from the images. For example [78, 56, 80, 55] use classical semantic segmentation based on K-MEANS. We think that thanks to the development of modern CNN architectures and training datasets for semantic segmentation, this information can be incorporated in an effective way to CBIR algorithms. The output of a segmentation network is a 2D-map that associates a semantic label (class) to each pixel. This is a high level representation suitable for building an image signature invariant to viewpoint and illumination.

Based on the semantic content given from the semantic segmentation output and the bag of visual words model [9], we propose in this work three different ways of constructing the image signature to improve the CBIR task and image classification. We show that the use of semantic information offers potential for improvement over standard approaches with benefits in terms of accuracy and computation time. It is an extension of the framework we initially proposed in [79], with a new image representation proposal and a complete study of the framework through extensive experiments. Three methodologies are proposed to build the image signature as well as a semantic filter to obtain our final image representation. Our contributions are as follows:

- a signature combining interest points and semantic information
- a signature combining visual features and semantic information
- a signature depending only on semantic information
- a semantic filter able to neutralize and penalize the images which are semantically different from an input query.

Our experimental results highlight the potential of our proposals with considerably more retrieved images than current state-of-the-art techniques on eight retrieval datasets.

The rest of the paper is structured as follows: we provide a brief overview of convolutional neural networks descriptors and bag of visual words related

works in Sect. 2. We explain our proposals in Sect. 3. We present the experimental part on different datasets and discuss the results of our work in Sect. 4. Section 5 is the conclusion.

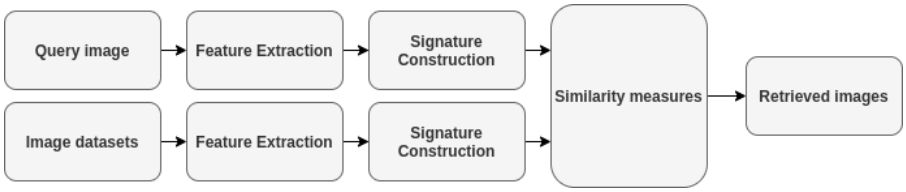

Fig. 1 Cbir system

2 State of the art

Many CBIR systems have been proposed during the last years [10,1,46,34]. In the literature there are three main methods for retrieving the images by similarity: (1) methods based on visual features extracted from the image using visual descriptors (2) learning methods based on deep learning architecture for constructing a global signature extracted from the features layer (3) end-to-end CNN based methods. Let's start by describing the methods based on visual features. Bags of visual words (BoVW) or Bags of visual features (BoF) [9] is the popular model used for image classification and similarity (see Figure 2). BoVW is treated as following. For each image, the visual features are detected and extracted using a visual descriptors such as SIFT [25].

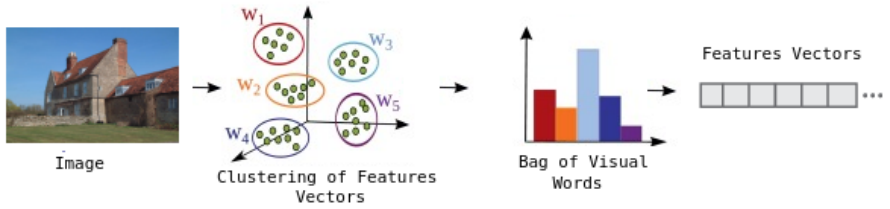

Fig. 2 Bag of visual words model

This step is repeated on all the images until all the visual descriptors in the dataset have been collected. Then a clustering step using K-MEANS [19] is applied on the descriptors to build the visual vocabulary (visual words) from the center of each cluster. In order to obtain the bags of visual words, the features extracted from the query image are replaced by the index of the visual words which are the nearest using euclidean distance. Finally, the image is described as a histogram of the frequency of the visual words. Inspired by BoVW, vector of locally aggregated descriptors (VLAD) [17] improves over BoVW by

assigning to each visual feature its nearest visual word and accumulate this difference for each visual word. Fisher Vector encoding [32] uses GMM [35] to construct a visual word dictionary. VLAD and Fisher are similar but VLAD does not store second order information about the features and use K-MEANS instead of GMM. Many descriptors have been proposed to encode the local image features into a vector. Scale Invariant Feature Transform (SIFT) [81] and Speeded-up Robust Features (SURF) [82] are the most used descriptors in CBIR. Interesting work from Arandjelović and Zisserman [83] introduces an improvement by upgrading SIFT to RootSift. In [84] a novel multi-scale 2D feature detection and description algorithm is presented. Inspired by LBP [85] descriptors, [86] proposes a novel method for image description with multichannel decoded LBPs and [75] propose a novel descriptor algorithm using local tetra patterns (LTrPs). [77] present a new image feature description based on the local wavelet pattern (LWP) for medical image retrieval. The authors in [76], propose a robust and invariant descriptor to rotation and illumination. Another method inspired by BoVW is the Bag of visual phrase (BoVP) [30,2,31]. BoVP describe the image as a matrix of visual phrase occurrence instead of a vector in BoVW. The idea is to link two or more visual words by a criterion. Then the phrase can be constructed in different ways (sliding windows, k-nearest neighbors, graph). In [2], local regions are grouped by the method of clustering (single-linkage). [31] group each key point with its closest spatial neighbors using L_2 distance. [27] proposes a framework between local and global histograms of visual words. The image features can be encoded then extracted based on color [38,7,87,14], texture [29,18] or shape [13,12]. In [45], a framework based on color features and texture analysis is presented. [33] introduces an effective image indexing technique where the features are extracted from discrete cosine transform (DCT) coefficients. [42] proposes a discriminative EODH descriptor with strong rotation-invariant and scale-invariant feature. [15] present a discriminative image descriptor dependent on both contour and color information. Inspired by bag of visual features, [43] proposes an image signature using spatial information. [62] proposes a a combination between HSV color moments and Gray Level Co-occurrence Matrix for a robust CBIR system. In [61], the proposed technique applies the texton layout to distinguish then extract the consistent zone of an image. Therefore, it computes the dominant color descriptor feature on the pixels in this consistent zone.

On another side, deep learning has proven very useful in computer vision applications. In particular, convolutional neural networks (CNN or ConvNet) are commonly applied to analyze image content. The architecture of CNN is composed of a set of layers. The major layers are: the input layer, hidden layers and the output layer. At the beginning of CNN networks, The CBIR problem was solved based on the classification model. Many CNN architectures have been proposed, including AlexNet [20], VGGNet [37], GoogleNet [41] and ResNet [40]. The fully connected layer (feature layer) is usually found towards the end of the CNN architecture with a 4096 dimensional floating point vector and it describes the image features (color, shape, texture, ...). The works in

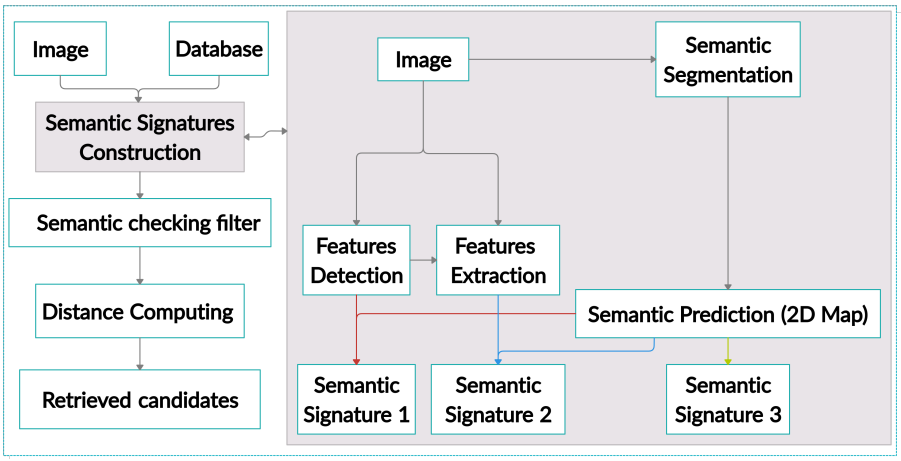

Fig. 3 Global framework

[88–90] present CNN for multiple image categorisation instead of assigning a single label by image. The computing of the similarity between two images is based on the L_2 metric between the features vector from the feature layer and the evaluation is based on the mean average precision (MAP). NetVLAD [3] inspired by VLAD is a CNN architecture used for image retrieval. [4] reduces the training time and provides an improvement in accuracy. Using ACP is frequent in the CBIR application thanks to its ability to reduce the descriptor dimension without losing its accuracy. [36] uses convolutional neural network (CNN) to train the network and support vector machine (SVM) to train the hyerplane then computes the distance between the features image and the trained hyper-plane. [64], introduced a novel neural network which use the heterogeneous superpixel to facilitate image object relational analysis. Based on neural network architecture for content based image retrieval [16] proposes an efficient feature extraction method. Recently, convolutional neural networks (CNN) become more efficient for image retrieval tasks. In [60], the introduced model uses a ResNet50 with co-occurrence matrix (RCM) model for CBIR. The authors in [59], propose an image signature based on VGG16 model.
Recently, several authors [91,92,48,49,74,73,72] have proposed new detectors and descriptors based on deep learning which can replace classical local features. Their utilization is getting increasingly frequent in computer vision applications. Moreover, CNN can give a global descriptor of an image such as LBP [85]. The proposed works [70,69,68,67], transform an input image into a global representation. Descriptors based on deep learning have been shown to be more robust against rotation and illumination changes than classical descriptors.

3 Contributions

The majority of CBIR systems describe the image as a N-dimensional vector. Bag of visual words [9] represent the image as a frequency histogram of vocabulary that are in the image. In deep learning approaches, the image signature is a vector of N floats extracted from the feature layer.

In this section, we present a signature construction framework. Our aim is to improve the image representation without prior knowledge on the images. The efficiency of any CBIR system depends on the robustness of the image signature. Figure 3 presents the different steps of our global framework. Motivated by the recent successes of deep learning in particular the convolutional neural networks (CNN), we propose three different methodologies to construct the images signature: (1) signature combining interest points and semantic information denoted bag of semantic visual words (BoSW) (2) signature combining visual features and semantic information denoted bag of semantic labels (BoSL) (3) signature depending only on semantic information denoted bag of semantic proportion (BoSP). After building the image signatures, we improve the CBIR algorithm with a semantic filter. First, we classify the images based on their semantic content. We check that the candidate shares the same classes labels with the query. If this is the case, we consider the candidate as a true candidate and proceed with the distance computation step. Otherwise, the images are semantically different and we can prune the candidate. This semantic filter decreases the CBIR computation time and increases the CBIR accuracy. Finally, we compute the distance between the query and the selected candidates using the L_2 distance.

3.1 Bag of semantic visual words: BoSW

In this section, we present a new idea to construct an image signature. We need two main elements to successfully construct the image signature: visual feature descriptors and semantic segmentation (2D map). By incorporating these information, we build a robust features description for an image taking into account both semantic and visual description.

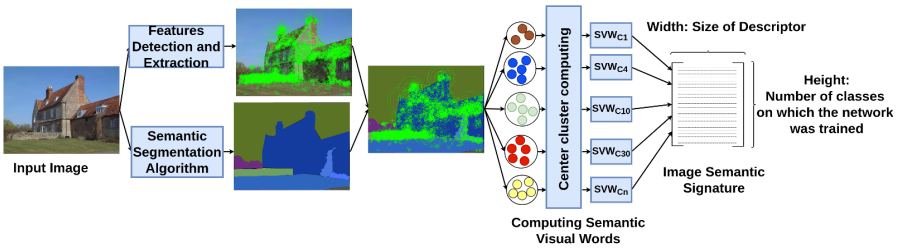

Fig. 4 Different steps for constructing the bag of semantic visual words (BoSW)

We define the signature as a $M \times N$ matrix where the width N corresponds to the size of the descriptor (128 for SIFT) and the height M corresponds to the number of classes on which the network was trained. Figure 4 describes the different steps of our approach. The process of construction is composed of three different steps: (i) detection and extraction of the visual features (ii) extraction of semantic information (iii) clustering the keypoints by class label and computing the center of the clusters. To compute the center, for each class label on the image we select the set of keypoints that belong to it and we apply the clustering algorithm (K-MEANS) with K=1 (average of keypoints). Consequently, each class label will be represented by a vector of N float denoted semantic visual words Svw_i. Finally the image signature is composed of N semantic visual words that represent the existing class labels in the image. It is not necessary that the image contains all classes. In this case, we assign a null vector for the missing classes.

$$BoSW_i = \begin{bmatrix} Svw_1 \\ Svw_2 \\ ..... \\ Svw_{n-1} \\ Svw_n \end{bmatrix}$$

3.2 **Bag of semantic labels: BoSL**

Inspired by bag of visual words [9], we propose a bag of semantic labels that describes the occurrence of semantic label within an image. Our idea applies on two main steps. First, we start by detecting the interest points from an input image. Then, we project the pixel coordinate (x, y) of the detected points on the segmented image given by the semantic segmentation network (figure 5). As a result, we obtain for each image a frequency histogram of labels that are in the image. The vector size corresponds to the number of classes on which the network was trained. It is not necessary that the image contains all semantic classes given by the network. In this case, we assign a null value in the cells of the missing classes.

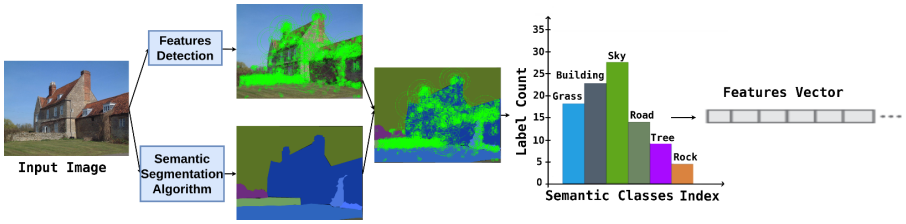

Fig. 5 Different steps for constructing the bag of semantic labels (BoSL)

3.3 Bag of semantic proportions: BoSP

Deep Learning based semantic segmentation networks output a 2D-map that associates a semantic label (class) to each pixel. From this output we can know the objects in the image and their proportion. So, fully depending on the CNN output, we exploit the semantic segmentation information to build a semantic signature "bag of semantic proportions" for image similarity. Since the term similar means here "with the same semantic content", our signature compare the images according to their semantic content. The construction process needs only the 2D-map from the output of the deep semantic architecture. As shown in figure 7, given a segmented image we divide it into N sub-images where each one represent a semantic object in the image. In the next step, we create a feature vector whose size corresponds to the number of classes on which the network was trained and each element contains the proportion of a semantic object in percentage. Also here, we assign a null value in the element of the missing classes.

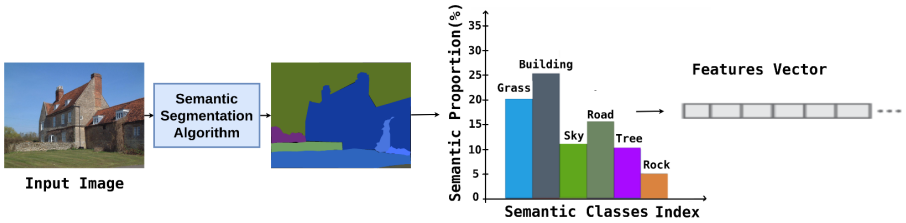

Fig. 6 Different steps for constructing the bag of semantic proportions (BoSP)

3.4 Semantic Filter

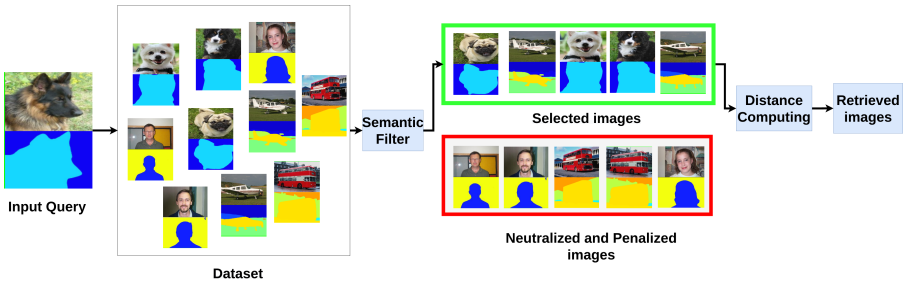

Fig. 7 Different steps for the retrieved images using the semantic filter

Semantic segmentation indicates which objects exist in the image. Using this information, we propose a semantic test to check the semantic similarity be-

tween two images. This means, we check if two images (the candidate and the query) share the same semantic classes. If this is the case, we proceed with the computing distance step. Otherwise, the images are semantically different and we can prune the candidate. The checking phase leads to decrease the CBIR time by keeping only the images that have the same semantic content as the query. This step reduces the complexity of computing from $O(n^2)$ to $O(nlog(n))$ and also increases the CBIR accuracy. According to the experiments the semantic filter offers us an increase in MAP score between 4% and 6%. For more explanation, figure 7 shows how we exploit the semantic filter to neutralize and penalize the images considered semantically different. In our example, the predicted classes in the input query are grass and dog. We exploit this information to find the images that share the same semantic classes with the query. Once we get the list of images considered as true candidates, we compute the distance between them. For the dissimilar images, we assign a negative score in order to neutralize them in the retrieval step.

4 Experimental results

4.1 Benchmark datasets for retrieval.

In this section, we present the potential of our approach on eight different datasets. Our goal is to increase the CBIR accuracy and reduce the execution time. To evaluate our proposition we test on the following datasets :
• Corel 1K or Wang [44] is a dataset of 1000 images divided into 10 categories (see figure 8) and each category contains 100 images. The evaluation is computed by the average precision of the first 100 nearest neighbors among 1000.

Fig. 8 Example of images from corel dataset

• Corel 10K [22] is a dataset of 10000 images divided into 100 categories and each category contains 100 images. The evaluation is computed by the average precision of the first 100 nearest neighbors among 10000.
• Corel 5K [22] is the first 5000 images from Corel-10K composed of 50 categories and each category contains 100 image. The evaluation is computed by

the average precision of the first 100 nearest neighbors among 10000.
• GHIM-10K [22] is a dataset of 10000 images divided into 20 categories and each category contains 500 images. The evaluation is computed by the average precision of the first 500 nearest neighbors among 10000.

Fig. 9 Example of images from Linnaeus dataset

• MSRC v1 (Microsoft Research in Cambridge) which has been proposed by Microsoft Research team. MSRC v1 contains 241 images divided into 9 categories. The evaluation on MSRC v1 is based on MAP score (Mean Average Precision)
• MSRC v2 contains 591 images including MSRC v1 dataset and divided into 23 categories. The evaluation on MSRC v2 is based on MAP score (Mean Average Precision)
• Linnaeus [6] is a new dataset composed of 8000 images of 4 categories (berry, bird, dog, flower) and a set of negative images. The evaluation on Linnaeus [6] is based on MAP score (Mean Average Precision)
• NUS-WIDE dataset [65] is a large dataset for multi-label recognition with 81 categories, which contains 269,648 images with associated tags from Flickr. NUS-WIDE dataset is composed of 161,789 images for training and 107,859 images for testing. The evaluations are based on precision, recall and F1-measure score.

4.2 **Training datasets for semantic segmentation.**

Many semantic segmentation datasets have been proposed in the last years such as Cityscapes [8], Mapillary [28], CoCo [24], ADE20K [47], CoCo-stuff [5], Mseg [54] and others. The semantic datasets are composed by two main objects: stuff and things. Things objects have characteristic shapes like vehicle, dog, computer... . Stuff is the description of amorphous objects like sea, sky, tree, ...

To segment an image we use the recent architecture named High-Resolution Net (HRNet) [53] with HRNetV2-W18 as backbone. We choose this network because of its superior results compared to older networks and its ability to produce a high-resolution representation of an input image. This architecture is trained on a large collection of datasets cited in table 1.

Dataset	Images	Categories	Obj Cls	Obj inst	Scene	Stuff /Things	Year
CoCo-stuff [5]	164K	172	-	-	Indoor/ Outdoor	91/ 80	2018
CoCo [24]	123K	91	-	889,284	Indoor/ Outdoor	- / -	2014
Mseg [54]	220K	194	-	-	Indoor/ Outdoor	-	2020
ADE20K [47]	25k	150	2,693	434,826	Indoor/ Outdoor	35 / 135	2017
Mapillary[28]	25k	66	-	-	Outdoor	28/38	2017
Pascal Context [52]	10.5k	59	540	104,398	Indoor/ Outdoor	152/388	2012

Table 1 Details about semantic datasets used to train the segmentation network

4.3 **Results on Benchmark datasets for retrieval**

In table 2, we present the results obtained for the Bag of semantic visual Words approach (BoSW). We conducted our experiments by training the segmentation network on six semantic datasets (Table 1). We then tested BoSW on eight benchmarking datasets. We extract the visual features with floating point based descriptors (Kaze, Hog, Surf). Three different extensions of KAZE descriptors have been used in this work (K_{region}, K_{edge}, $K_{sharpedge}$). Moreover, we have additionally tested our methodology utilizing the local detector and descriptor *SuperPoint* based on deep learning [48].

Descriptors/ Retrieval Dataset	SIFT	SURF	HOG	K_{region}	K_{edge}	$K_{sharpedge}$	SuperPoint
MAP evaluations using Mseg dataset							
MSRC v1	0.78	0.77	0.70	0.82	0.81	0.83	**0.85**
MSRC v2	0.58	0.56	0.53	0.61	0.61	0.62	**0.65**
Wang [44]	0.70	0.71	0.57	0.77	**0.78**	0.77	0.75
Corel-5K [22]	0.48	0.52	0.39	0.54	0.55	0.55	**0.57**
Corel-10K [22]	0.38	0.41	0.30	0.44	0.45	0.45	**0.48**
GHIM-10K [22]	**0.34**	0.32	0.31	0.32	0.33	0.33	0.33
Linnaeus [6]	**0.65**	0.62	0.64	0.61	0.62	0.62	0.61
NUS-WIDE[65]	0.71	0.72	0.75	0.70	0.74	0.72	**0.76**
MAP evaluations using CoCo-stuff dataset							
MSRC v1	0.78	0.76	0.79	0.86	0.84	0.85	**0.89**
MSRC v2	0.64	0.61	0.63	0.70	0.70	0.71	**0.73**
Wang [44]	0.74	0.72	0.71	0.82	**0.83**	0.82	**0.83**
Corel-5K [22]	0.65	**0.66**	0.57	0.62	0.63	0.63	0.65
Corel-10K [22]	0.55	**0.60**	0.47	0.52	0.53	0.53	0.59
GHIM-10K [22]	0.34	0.31	0.32	0.30	0.30	0.31	**0.41**
Linnaeus [6]	0.77	**0.80**	0.79	0.78	0.79	0.79	0.79
NUS-WIDE[65]	0.55	**0.57**	0.54	0.51	0.54	0.51	0.53
MAP evaluations using ADE20K dataset							
MSRC v1	0.71	0.70	0.64	0.77	0.78	0.77	**0.79**
MSRC v2	0.54	0.52	0.45	0.58	**0.59**	0.58	0.55
Wang [44]	0.64	0.65	0.53	0.71	0.71	**0.72**	0.71
Corel-5K [22]	0.46	0.49	0.36	0.52	0.51	0.52	**0.55**
Corel-10K [22]	0.36	0.38	0.28	0.41	**0.42**	0.41	0.40
GHIM-10K [22]	0.33	0.32	0.29	0.31	**0.34**	**0.34**	0.31
Linnaeus [6]	0.47	**0.48**	0.47	0.46	0.45	0.45	0.47
NUS-WIDE[65]	0.56	0.59	0.61	0.60	0.61	0.61	**0.63**
MAP evaluations using Mapillary dataset							
MSRC v1	0.65	0.60	0.56	0.74	0.73	0.74	**0.75**
MSRC v2	0.43	0.40	0.37	0.49	0.48	0.50	**0.53**
Wang [44]	0.50	0.49	0.34	0.55	0.56	**0.57**	0.55
Corel-5K [22]	0.35	0.36	**0.40**	0.33	0.34	0.30	0.31
Corel-10K [22]	**0.41**	0.40	0.32	0.35	0.37	0.35	**0.41**
GHIM-10K [22]	0.25	0.26	0.20	0.22	0.23	0.23	**0.31**
Linnaeus [6]	0.46	**0.49**	0.47	0.46	0.46	0.47	0.45
NUS-WIDE[65]	0.45	0.46	0.41	0.44	0.45	0.44	**0.47**
MAP evaluations using Pascal Context dataset							
MSRC v1	0.86	0.85	0.87	0.86	0.85	0.85	**0.89**
MSRC v2	0.65	0.64	0.64	0.60	0.61	0.62	**0.69**
Wang [44]	0.60	0.61	0.50	0.62	0.61	0.61	**0.65**
Corel-5K [22]	0.48	0.51	0.38	0.51	**0.52**	**0.52**	0.50
Corel-10K [22]	0.43	0.46	0.33	0.46	**0.47**	0.46	**0.47**
GHIM-10K [22]	0.37	0.36	0.33	0.37	**0.38**	**0.38**	0.37
Linnaeus [6]	0.56	**0.57**	0.54	0.56	0.55	0.53	0.55
NUS-WIDE[65]	0.45	0.46	0.41	0.46	0.45	0.44	**0.49**
MAP evaluations using CoCo dataset							
MSRC v1	0.83	0.81	0.78	0.86	0.86	0.87	**0.88**
MSRC v2	0.62	0.58	0.60	0.66	0.66	**0.68**	0.66
Wang [44]	0.74	0.73	0.64	0.80	0.81	0.80	**0.83**
Corel-5K [22]	0.65	0.65	0.57	0.63	0.62	0.63	**0.67**
Corel-10K [22]	0.39	0.41	0.30	0.44	**0.45**	**0.45**	0.44
GHIM-10K [22]	**0.36**	0.34	0.33	0.34	0.34	0.35	0.33
Linnaeus [6]	0.75	0.72	0.69	0.70	0.67	0.72	**0.77**
NUS-WIDE[65]	0.69	0.72	0.72	0.74	0.74	0.75	**0.76**

Table 2 MAP evaluations scores for bag of semantic visual words (BoSW)

Retrieval Dataset \ Detectors	SURF	KAZE	Harris	FAST	Min Eigen	MSER	Super Point
MAP evaluations using Mseg dataset							
MSRC v1	0.82	0.81	**0.89**	0.82	0.79	0.84	0.88
MSRC v2	0.61	0.63	0.63	0.64	0.58	**0.66**	0.65
Wang [44]	0.76	0.74	0.57	0.53	0.75	0.60	**0.77**
Corel-5K [22]	0.35	**0.38**	0.33	0.32	0.38	0.34	**0.41**
Corel-10K [22]	0.26	**0.30**	0.25	0.25	0.29	0.26	0.28
GHIM-10K [22]	0.39	0.41	0.44	0.45	0.45	**0.46**	0.45
Linnaeus [6]	0.60	0.60	**0.64**	0.58	0.62	0.56	0.60
NUS-WIDE[65]	0.71	0.70	0.68	0.65	0.66	0.60	**0.72**
MAP evaluations using CoCo-stuff dataset							
MSRC v1	0.84	0.83	0.83	0.82	0.83	0.86	**0.89**
MSRC v2	0.69	0.69	0.70	0.70	0.65	**0.73**	0.71
Wang [44]	0.85	0.85	0.83	0.83	0.82	0.84	**0.87**
Corel-5K [22]	0.68	**0.69**	0.65	0.58	0.68	0.65	0.68
Corel-10K [22]	0.57	0.59	0.58	0.54	0.57	0.55	**0.61**
GHIM-10K [22]	0.47	0.52	**0.53**	0.51	0.45	0.54	0.52
Linnaeus [6]	0.76	0.78	**0.80**	0.76	0.74	0.78	0.79
NUS-WIDE[65]	**0.73**	0.71	**0.73**	0.68	0.69	0.61	0.72
MAP evaluations using ADE20K dataset							
MSRC v1	0.77	0.76	0.74	0.75	0.76	0.79	**0.81**
MSRC v2	0.54	0.54	0.55	0.54	0.52	**0.56**	0.55
Wang [44]	**0.73**	0.72	0.68	0.66	0.69	0.68	**0.73**
Corel-5K [22]	0.30	0.31	**0.34**	0.30	**0.34**	0.30	0.32
Corel-10K [22]	0.30	**0.34**	0.29	0.28	**0.34**	0.29	0.33
GHIM-10K [22]	0.41	0.43	0.43	0.43	0.38	**0.44**	0.43
Linnaeus [6]	0.44	**0.45**	0.44	0.44	0.42	0.44	0.44
NUS-WIDE[65]	0.48	0.51	0.55	0.35	0.41	0.44	**0.57**
MAP evaluations using Mapillary dataset							
MSRC v1	0.71	**0.72**	0.65	0.67	0.71	**0.72**	0.71
MSRC v2	0.55	0.57	0.52	0.54	0.59	0.58	**0.60**
Wang [44]	**0.56**	0.55	0.50	0.40	0.52	0.41	0.55
Corel-5K [22]	0.22	0.25	0.21	0.20	0.25	0.21	**0.31**
Corel-10K [22]	0.17	0.20	0.16	0.15	0.19	0.16	**0.23**
GHIM-10K [22]	0.34	**0.35**	0.32	0.33	0.32	0.34	0.34
Linnaeus [6]	0.28	0.29	0.28	0.28	0.30	0.29	**0.33**
NUS-WIDE[65]	0.48	0.51	**0.55**	0.35	0.41	0.44	0.54
MAP evaluations using Pascal Context dataset							
MSRC v1	0.84	0.83	0.79	0.81	0.83	0.82	**0.87**
MSRC v2	0.49	0.50	0.46	0.43	**0.52**	0.47	0.51
Wang [44]	0.52	0.53	0.50	0.50	0.52	0.55	**0.57**
Corel-5K [22]	0.34	0.33	0.30	**0.35**	0.33	0.30	0.34
Corel-10K [22]	**0.29**	**0.29**	0.28	0.27	0.23	0.27	0.28
GHIM-10K [22]	0.34	0.32	**0.35**	0.34	0.32	0.34	0.33
Linnaeus [6]	0.54	0.55	0.55	**0.56**	0.55	0.53	0.55
NUS-WIDE[65]	0.44	0.45	0.41	0.39	0.36	0.41	**0.49**
MAP evaluations using CoCo dataset							
MSRC v1	0.84	0.84	0.83	0.84	0.84	0.89	**0.91**
MSRC v2	0.68	0.67	0.69	0.68	0.65	**0.73**	0.71
Wang [44]	0.82	0.81	0.78	0.78	0.78	0.80	**0.86**
Corel-5K [22]	0.35	**0.38**	0.33	0.32	0.38	0.33	**0.40**
Corel-10K [22]	0.26	0.58	0.54	0.51	**0.59**	0.55	0.57
GHIM-10K [22]	0.42	0.45	0.46	0.45	0.39	**0.47**	0.45
Linnaeus [6]	0.73	0.73	0.68	0.69	0.65	0.70	**0.75**
NUS-WIDE[65]	0.68	**0.70**	0.65	0.67	0.66	0.62	0.69

Table 3 MAP evaluations scores for bag of semantic labels method (BoSL)

In table 3, we present the results obtained for the bag of semantic labels method (BoSL). With the same setup, we conducted our experiment by using six semantic datasets for training the segmentation network (Table 1). Then we tested on eight retrieval datasets. We detect the interest points using seven different detectors (SURF, KAZE, Harris, FAST, MinEigen, MSER, SuperPoint).

In table 4, we present the results obtained for the full semantic signature (BoSP). What distinguishes this method is that it can quickly classify the image depending only on the semantic segmentation without any additional information. Among the three proposed signatures, we obtained the best score with the BoSW method. However, BoSL and BoSP have shwown close results just between 1% and 3% below those of BoSW (and in a few cases the results are better than BoSW). Also, these two methods are faster and easier to implement compared to BoSW.

Semantic dataset / Retrieval Datasets	CoCo-stuff	Mseg	Pascal Context	Mapillary	ADE20K	CoCo
MSRC v1	0.84	0.82	**0.85**	0.72	0.77	0.84
MSRC v2	**0.68**	0.60	0.55	0.47	0.54	0.66
Wang [44]	0.83	**0.88**	0.53	0.56	0.72	0.80
Corel-5K [22]	**0.68**	0.62	0.34	0.41	0.55	0.62
Corel-10K [22]	**0.57**	0.51	0.29	0.33	0.43	0.52
GHIM-10K [22]	**0.45**	0.38	0.34	0.32	0.38	0.40
Linnaeus [6]	**0.73**	0.57	0.56	0.28	0.48	0.65
NUS-WIDE[65]	**0.77**	0.75	0.45	0.53	0.66	0.64

Table 4 MAP evaluations scores for bag of semantic proportion method (BoSP)

Table 5 summarizes the executions time for different steps of our framework. We conclude that BoSW is a method which takes more time for signature construction than BoSP and BoSL due to the computation of the semantic words. The BoSW method depends on descriptors, so when the descriptor size increases, the signature creation time also increases. For each query from Corel-5K/Corel-10K datasets the evaluation is computed by the average precision of the first 100 nearest neighbors among all images in the datasets. In figure 10, the experiments are made on different image sizes (10, 20,... 100).

Proposed Methods	Features Detection	Features Extraction	Semantic Segmentation	Signature construction
Bag of semantic visual words	0.31	0.09	0.30	0.070
Bag of semantic labels	0.31	-	0.30	0.003
Bag of semantic proportions	-	-	0.30	0.004

Table 5 Execution time of the proposed methods in seconds per image

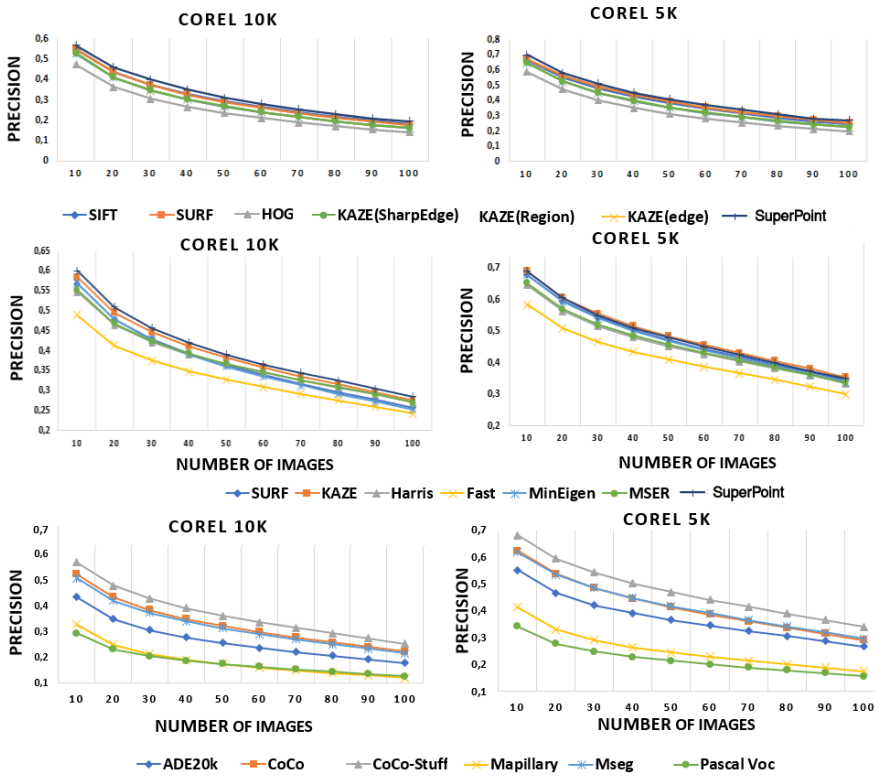

Fig. 10 Precision graph for Corel-10K and Corel-5K datasets

In the top part of figure 10, we present the MAP score obtained by the bag of semantic visual words (BoSW) method with the segmentation network trained on CoCo-Stuff. The results are obtained using five different descriptors (SIFT, SURF, KAZE, HOG, SuperPoint). For KAZE descriptors, we complete the evaluations with three extensions where the points are detected with three different detectors (edge, region, sharp-edge). The best results were obtained with the SuperPoint descriptor. In the middle part of figure 10, we present the map score obtained by the bag of semantic label (BoSL) method with the segmentation network trained on CoCo-Stuff. The results are shown for sev-

eral detectors (SURF, KAZE, Harris, FAST, MinEign, MSER, SuperPoint). We notice that the best results were obtained with the KAZE detector. On the other hand, the worst results were obtained with the Fast detector. In the bottom part of figure 10, we present the MAP scores obtained by the bag of semantic proportion (BoSP) method. We show the results with a large collection of semantic datasets used to train the segmentation network (ADE20K, CoCO, CoCO-Stuff, Mseg, Pascal Context, Mappilary). We notice that the best results were obtained with the CoCo-Stuff dataset. On the other hand, the worst results were obtained with both Pascal Context and Mapillary datasets. In figure 11, we present the average precision AP for each class in the NUS-WIDE dataset using BoSP.

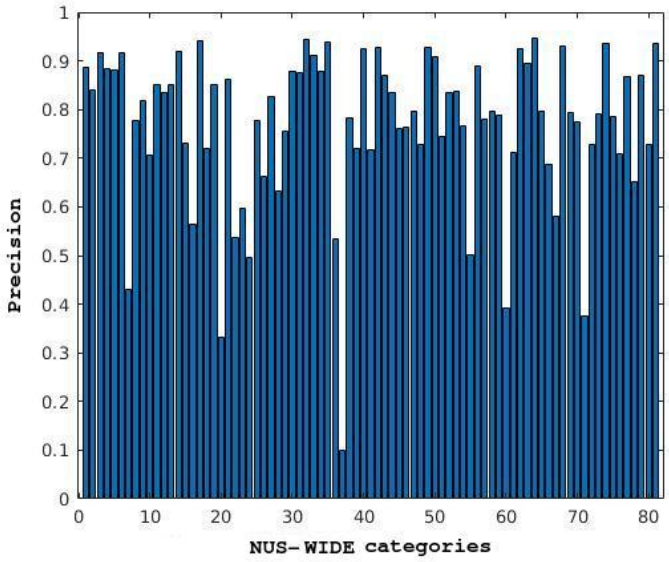

Fig. 11 Precision per class of the 81 tags in the NUS-WIDE dataset with k = 3.

We have also done experiments with a support vector machine (SVM) with linear kernel using the histogram computed using BoSP method. For a detailed comparison, the confusion matrix and ROC curve for Corel dataset for 10 categories is displayed in figure 12. As shown the results for the experiment with the SVM show the robustness of our method.

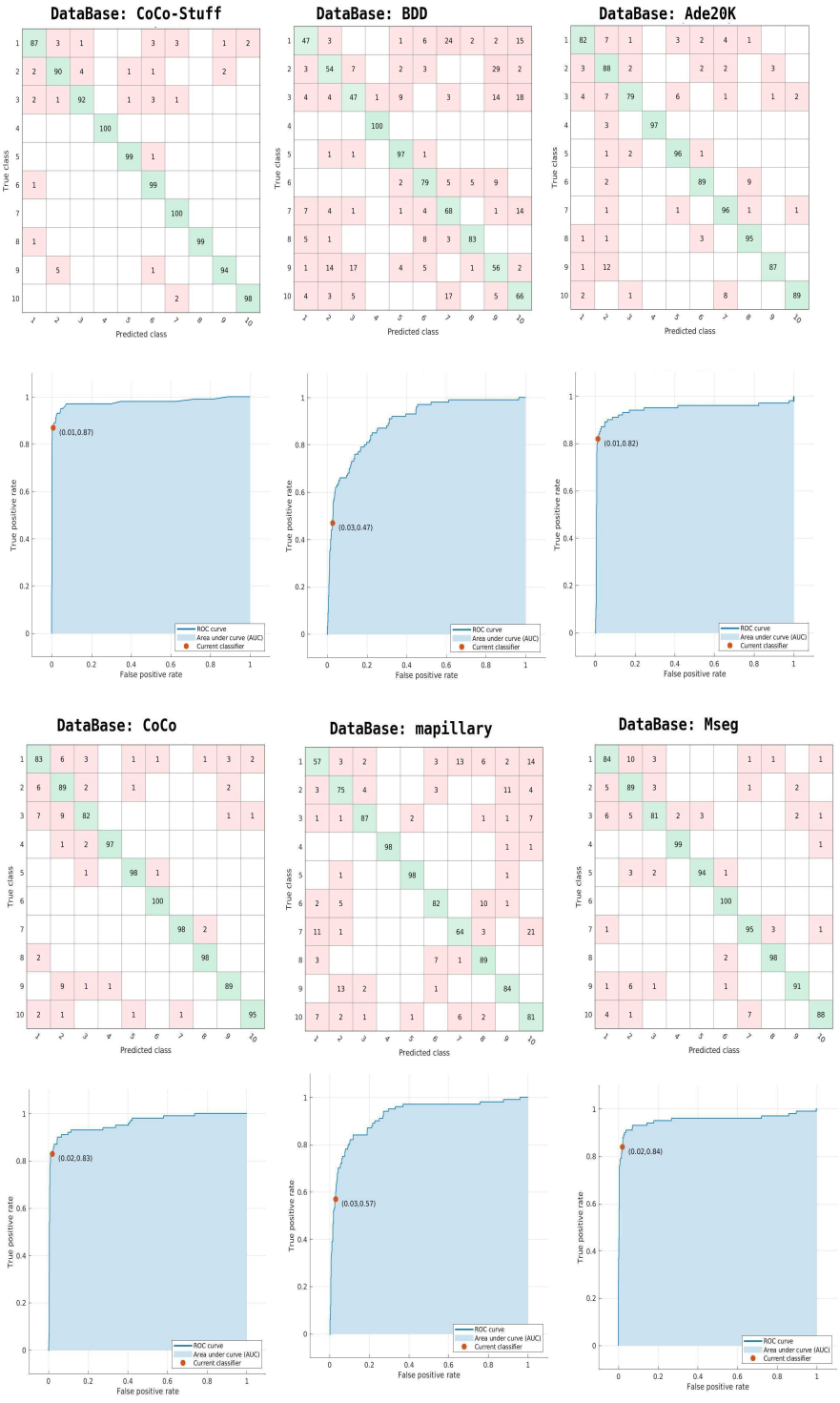

Fig. 12 ROC curves and confusions matrices for BoSP method on Corel 1000 dataset using several datasets for training the segmentation network.

4.4 **Comparisons with the state of the art**

In order to test the efficiency of our methods, we conducted the experimentation on eight retrieval datasets. We divided the state of the art into two main categories : (i) local visual feature approaches: methods that are based on local features (texture, color, shape) including the inherited methods such as BoVW, Vlad, Fisher... (ii) deep learning approaches: methods based on learning the features using deep learning algorithms. In Table 6 we compare our results with a large selection of state of the art methods.

Methods	$MSRC_1$	$MSRC_2$	Linnaeus	Wang	Corel10K	GHIM
Local visual feature approaches						
BoVW [9]	0.48	0.30	0.26	0.48	0.30	0.39
n-BoVW [30]	0.58	0.39	0.31	0.60	0.34	0.40
VLAD [17]	0.79	0.41	-	0.74	0.38	0.44
N-Gram [31]	-	-	-	0.37	-	-
SaCoCo[15]	-	-	-	0.54	0.17	0.15
Shiv [86]	-	-	-	0.75	-	-
Mary [62]	-	-	-	0.83	0.56	-
DCD-HM [61]	-	-	-	0.78	0.53	-
Deep learning approaches						
AlexNet [20]	0.81	0.58	0.47	0.87	0.41	0.52
VGGNet [37]	0.76	0.63	0.48	0.76	0.45	0.57
ResNet [40]	0.83	0.70	0.69	0.82	-	**0.62**
Li [51]	-	-	-	0.76	-	-
Anna [16]	-	-	-	0.75	-	-
Yang [64]	-	-	**0.81**	-	-	-
Ruigang [36]	-	-	0.70	-	-	-
MariP[60]	-	-	-	-	0.56	-
Ours (best)	**0.91**	**0.73**	0.80	**0.88**	**0.60**	0.53

Table 6 Comparison of the accuracy of our approach with methods from the state of the art

Methods	F1	P	R
KNN [65]	0.47	0.42	0.53
WARP [88]	0.54	0.48	0.60
SRN [90]	**0.62**	0.56	**0.69**
CNN-RNN [89]	0.55	0.50	0.62
Ours(best)	**0.62**	**0.77**	0.52

Table 7 Comparison results between state-of-the-art methods on NUS-WIDE dataset.

Following conventional settings [89,90] for NUS-WIDE, we report the following statistics: the average overall precision (P), the average overall recall (R) and F1-measure (F1). For each query, the 3 closest images are retrieved. Those with a distance to the query higher than 0.5 are eliminated.

In table 7, we present the quantitative results obtained by our method and compared with state of the art methods on the NUS-WIDE dataset.

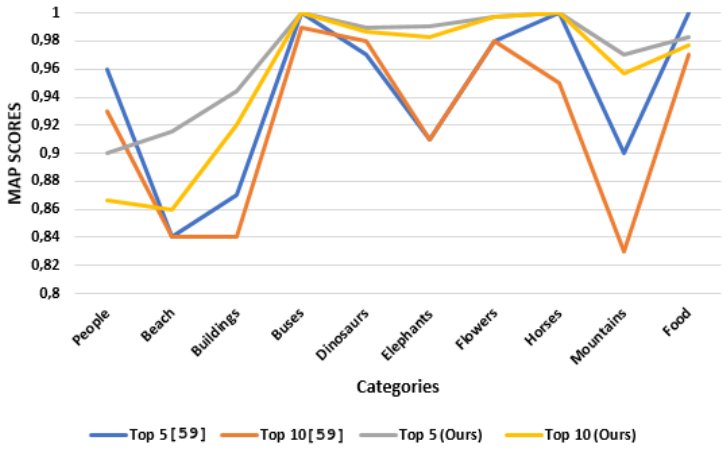

Fig. 13 Comparison of mean average precision (MAP) of top 10, or top 5 for the Wang dataset.

In figure 13, we compare the mean average precision (MAP) of the top 5 and top 10 retrieved images for all categories for the Wang [44] dataset between our method (BoSP) and a CNN based method [59]. Our method presents a good performance in almost all categories. In table 8 and figure 14, similar comparisons for the top 20 retrieved images are shown with a wide selection of state of the art methods [10, 1, 46, 34, 27, 45, 16, 33, 42]. For these comparisons, our method is the BoSW method whose signature was built on the basis of a CNN trained on CoCo-Stuff.

Class	BoSW	[10]	[1]	[46]	[34]	[27]	[45]	[16]	[33]	[42]
Africa	**0.85**	0.69	0.72	0.73	0.85	0.73	0.63	0.65	0.70	0.74
Beach	**0.85**	0.55	0.59	0.69	0.71	0.74	0.64	0.60.	0.44	0.37
Buildings	0.80	0.66	0.55	0.72	0.73	**0.81**	0.69	0.62	0.70	0.53
Buses	**0.99**	0.85	0.84	0.85	0.87	0.95	0.91	0.85	0.76	0.96
Dinosaurs	0.98	0.96	0.96	0.95	0.96	0.97	0.99	0.93	**1**	0.99
Elephants	**0.98**	0.36	0.73	0.72	0.76	0.87	0.78	0.65	0.63	0.66
Flowers	**0.96**	0.87	0.91	0.92	0.80	0.85	0.94	0.94	0.92	0.92
Horses	**0.99**	0.84	0.91	0.85	0.91	0.86	0.95	0.77	0.94	0.87
Mountains	**0.84**	0.44	0.47	0.55	0.70	0.82	0.73	0.73	0.56	0.58
Food	**0.96**	0.70	0.65	0.75	0.66	0.78	0.80	0.81	0.74	0.62

Table 8 Comparison of precision for top 20 retrieved images(Wang [44] dataset)

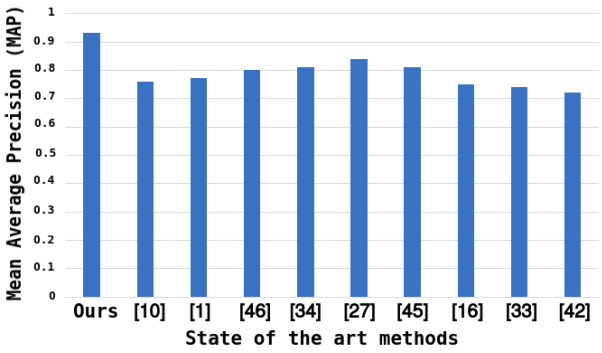

Fig. 14 Comparison of mean average precision score (MAP) for the Wang dataset between our method and the state of the art methods

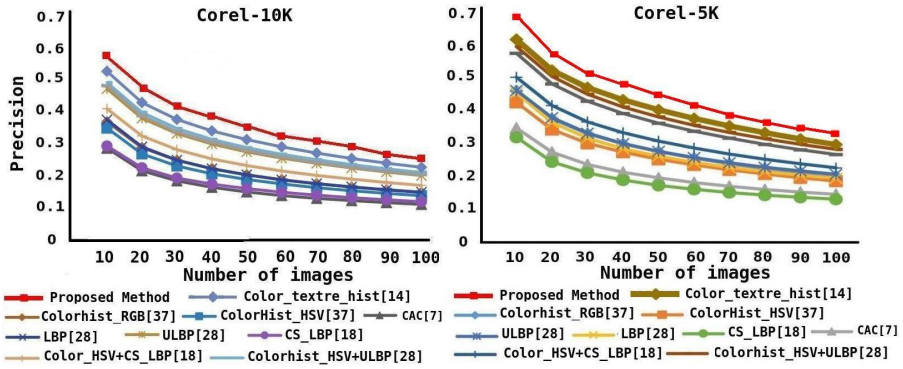

Fig. 15 Comparison of precision score for Corel-10K and Corel-5K datasets with the state of the art methods

In figure 15, we compare our approach (BoSW) with methods based on color [38,7,14] or texture [29,18] histogram on different image sizes (10, 20,..., 100). We show experimentally on Corel5K and Corel-10K datasets that the BoSW method offers potential for improvement over standard approaches with benefits in terms of accuracy.

It is important to compare the runtime of the proposed methods with both deep learning and local visual state of the art methods. In table 9, we compare only the time taken by each method for signature construction. Our BoSP, BoSL signatures are more than 135 times smaller than those built on VGG-16 architecture and 90 times smaller than those extracted with the ResNet architecture. In addition, we obtain a vector length smaller than all cited methods in table 9. This is an advantage in terms of searching time and memory. We consider the vector length of the proposed methods is 64 if during the build-

ing of the signature the descriptor used is SURF or KAZE. For VLAD [17], N-BoVW [30], BoVW [9] the length of the vector depends on the number K which is used to calculate the visual words using the K-means algorithm.

Methods	Vector Length	Signature Construction Time (ms)
SaCoCo[15]	120	18
VGG-16 [37]	4096	354
ResNet [40]	2048	270
AlexNet [20]	4096	193
BoVW [9]	dynamic	39
N-BoVW [30]	dynamic	48
VLAD [17]	dynamic	250
BoSW	$64 \times N$	70
BoSL	64	3
BoSP	64	4

Table 9 Comparing the properties of the proposed methods with the state of the art methods. N is the number of classes for which the segmentation network was trained.

Based on semantic content, we show some examples (figure 16) of bag of semantic proportion (BoSP) output. From different categories selected from different datasets (Corel 1K, MSRC V1), we show for each query the three nearest neighbors predicted on CoCo-Stuff dataset.

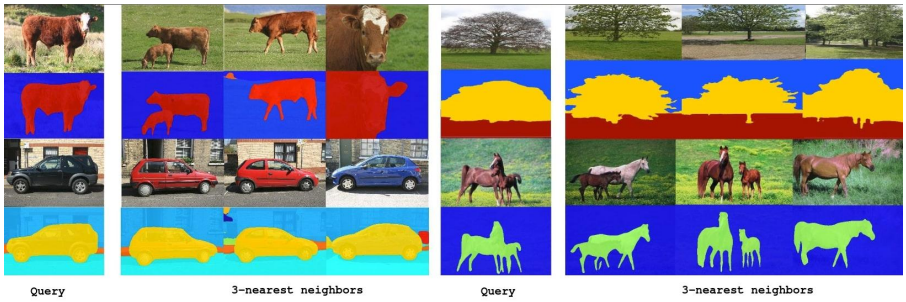

Fig. 16 Example of Bag of semantic proportion (BoSP) output

Classes	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Africa (1)	75	5	4	0	0	5	0	3	1	7
Beach (2)	2	72	6	0	0	8	0	0	12	0
Buildings (3)	4	7	64	6	0	0	0	6	7	6
Buses (4)	5	3	10	75	0	0	0	1	4	2
Dinosaurs (5)	0	0		0	100	0	0	0	0	0
Elephants (6)	10	3	3	0	0	75	0	1	2	6
Flowers (7)	0	0	1	0	0	0	93	0	0	6
Horses (8)	1	3	0	0	0	5	0	87	0	4
Mountains (9)	3	6	3	6	0	5	0	0	73	4
Food (10)	8	0	0	0	0	2	9	0	0	81

Table 10 Confusion matrix of method [87].

Classes	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Africa (1)	66	6	10	0	0	4	0	0	2	12
Beach (2)	2	64	10	0	2	6	0	2	12	2
Buildings (3)	6	4	68	2	0	6	2	2	6	4
Buses (4)	2	2	4	86	0	0	0	0	2	4
Dinosaurs (5)	0	0		0	92	0	0	0	0	8
Elephants (6)	0	8	0	0	0	82	0	0	8	2
Flowers (7)	0	0	0	0	0	0	98	0	0	2
Horses (8)	0	0	0	0	0	2	0	98	0	0
Mountains (9)	2	10	8	4	0	2	0	0	72	2
Food (10)	10	2	2	2	0	2	9	0	4	80

Table 11 confusion matrix of method [80].

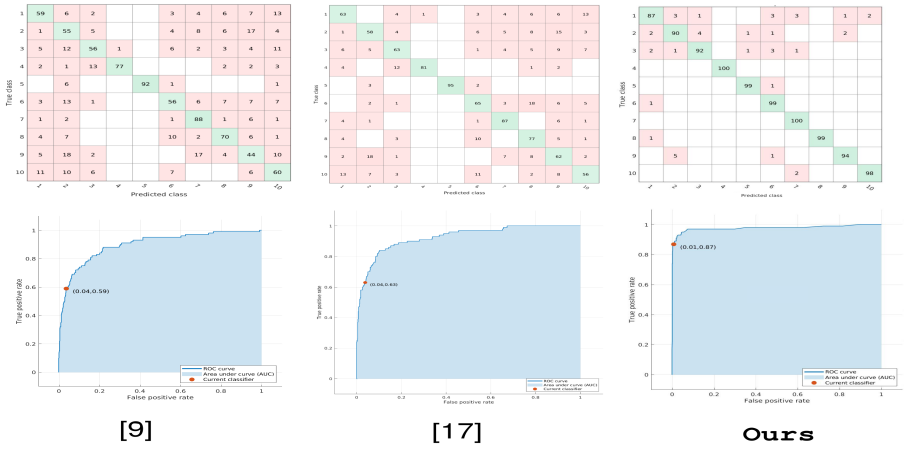

Fig. 17 Comparison of confusion matrices and ROC curves between [9] [17] and bag of semantic proportion (BoSP) method

In Tables 11,10, we present the confusion matrix obtained by [87] and [9] methods for Corel 1000 dataset. Then, we compared our confusion matrix obtained by BoSP (see figure 17) against four methods from the state of the art [87] [80] [9] [17]. Our confusion matrix shows the best results in all categories except the dinosaur class in method [87]. In addition, we compare the roc curves and confusion matrices between our method (BoSP) and [9] [17] in figure 17. We show experimentally that our proposals present better outcomes against standard image retrieval approaches.

5 Conclusion

We have presented in this paper an efficient framework for CBIR tasks and image classification. We leverage the discriminative information provided by a semantic segmentation CNN in the retrieval context to propose three different methodologies. Based on semantic content combined with local visual features, our propositions have shown that the use of the semantic content improves the retrieval accuracy. Another contribution of this paper is the proposed semantic filter. It allows the proposed framework to reduce the error rate and speed up the comparison between images. Using different descriptors, detectors and semantic datasets our approach achieves better results in terms of accuracy and computation time compared to the state of the art methods.

Compliance with ethical standards

Conflict of interest: Authors have no conflict of interest in this work.

References

1. Nandkumar S Admile and Rekha R Dhawan. Content based image retrieval using feature extracted from dot diffusion block truncation coding. In 2016 International Conference on Communication and Electronics Systems (ICCES), pages 1–6.IEEE, 2016.
2. Rami Albatal, Philippe Mulhem, and Yves Chiaramella. Visual phrases for automatic images annotation. In 2010 International Workshop on Content Based Multimedia Indexing (CBMI), pages 1–6. IEEE, 2010.
3. R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In IEEE Conference on Computer Vision and Pattern Recognition, 2016.
4. Thanasekhar Balaiah, Timothy Jones Thomas Jeyadoss, Sri Sainee Thirumurugan, and Rahul Chander Ravi. A deep learning framework for automated transfer learning of neural networks. In 2019 11th International Conference on Advanced Computing (ICoAC), pages 428–432. IEEE, 2019.
5. Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1209–1218, 2018.
6. G Chaladze and L Kalatozishvili. Linnaeus 5 dataset for machine learning. Technical report, Tech. Rep, 2017.

7. Young Deok Chun, Nam Chul Kim, and Ick Hoon Jang. Content-based image retrieval using multiresolution color and texture features. IEEE Transactions on Multimedia, 10(6):1073–1084, 2008.
8. Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3213–3223, 2016.
9. Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In Workshop on statistical learning in computer vision, ECCV, volume 1, pages 1–2. Prague, 2004.
10. M Esmel ElAlami. A new matching strategy for content based image retrieval system Applied Soft Computing, 14:407–418, 2014.
11. Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep, 8, 2011.
12. Rong-Xiang Hu, Wei Jia, Haibin Ling, Yang Zhao, and Jie Gui. Angular pattern and binary angular pattern for shape retrieval. IEEE Transactions on Image Processing, 23(3):1118–1127, 2013.
13. Rui Hu, Mark Barnard, and John Collomosse. Gradient field descriptor for sketch based retrieval and localization. In 2010 IEEE International Conference on Image Processing, pages 1025–1028. IEEE, 2010.
14. Jing Huang, S Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih. Image indexing using color correlograms. In Proceedings of IEEE computer society conference on Computer Vision and Pattern Recognition, pages 762–768. IEEE, 1997.
15. Chryssanthi Iakovidou, Nektarios Anagnostopoulos, Mathias Lux, Klitos Christodoulou, Y Boutalis, and Savvas A Chatzichristofis. Composite description based on salient contours and color information for cbir tasks. IEEE Transactions on Image Processing, 28(6):3115–3129, 2019.
16. Aun Irtaza, M Arfan Ja ar, Eisa Aleisa, and Tae-Sun Choi. Embedding neural networks for semantic association in content based image retrieval. Multimedia tools and applications, 72(2):1911-1931, 2014.
17. Herv Jegou, Matthijs Douze, Cordelia Schmid, and Patrick Perez. Aggregating lo-cal descriptors into a compact image representation. In 2010 IEEE computer society conference on computer vision and pattern recognition, pages 3304-3311. IEEE, 2010.
18. Suresh Kumar Kanaparthi, USN Raju, P Shanmukhi, G Khyathi Aneesha, and Mohammed Ehsan Ur Rahman. Image retrieval by integrating global correlation of color and intensity histograms with local texture features. Multimedia Tools and Applications, pages 1-37, 2019.
19. K krishna1999genetic and M Narasimha Murty. Genetic k-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 29(3):433-439, 1999.
20. Alex Krizhevsky, Ilya Sutskever, and Geo rey E Hinton. Imagenet classi cation with deep convolutional neural networks. In Advances in neural information processing sys-tems, pages 1097-1105, 2012.
21. John Lambert, Liu Zhuang, Ozan Sener, James Hays, and Vladlen Koltun. MSeg: A composite dataset for multi-domain semantic segmentation. In Computer Vision and Pattern Recognition (CVPR), 2020.
22. Jia Li and James Ze Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Transactions on pattern analysis and machine intelligence, 25(9):1075-1088, 2003.
23. Jun Li, Bo Yang, Wankou Yang, Changyin Sun, and Jianhua Xu. Subspace-based multi-view fusion for instance-level image retrieval. The Visual Computer, pages 1-15, 2020.
24. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740-755. Springer, 2014.
25. David G Lowe. Object recognition from local scale-invariant features. In Proceedings of the seventh IEEE international conference on computer vision, volume 2, pages 1150-1157. Ieee, 1999.

26. Achref Ouni Eric Royer Marc Chevaldonne, Michel Dhome. Hybrid approach for improved image similarity using semantic segmentation. In to appear in ISVC2020. Springer, 2020.
27. Zahid Mehmood, Syed Muhammad Anwar, Nouman Ali, Ha z Adnan Habib, and Muhammad Rashid. A novel image retrieval based on a combination of local and global histograms of visual words. Mathematical Problems in Engineering, 2016, 2016.
28. Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In Proceedings of the IEEE International Conference on Computer Vision, pages 4990-4999, 2017.
29. Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classi cation with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence, 24(7):971-987, 2002.
30. Achref Ouni, Thierry Urruty, and Muriel Visani. A robust cbir framework in between bags of visual words and phrases models for speci c image datasets. Multimedia Tools and Applications, 77(20):26173-26189, 2018.
31. Glauco V Pedrosa and Agma JM Traina. From bag-of-visual-words to bag-of-visual-phrases using n-grams. In 2013 XXVI Conference on Graphics, Patterns and Images, pages 304-311. IEEE, 2013.
32. Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In 2007 IEEE conference on computer vision and pattern recognition, pages 1-8. IEEE, 2007.
33. P Poursistani, Hossein Nezamabadi-pour, R Askari Moghadam, and Masoud Saeed. Image indexing and retrieval in jpeg compressed domain based on vector quantization. Mathematical and Computer Modelling, 57(5-6):1005-1017, 2013.
34. Jitesh Pradhan, Sumit Kumar, Arup Kumar Pal, and Haider Banka. Texture and color visual features based cbir using 2d dt-cwt and histograms. In International Conference on Mathematics and Computing, pages 84-96. Springer, 2018.
35. Carl Edward Rasmussen. The in nite gaussian mixture model. In Advances in neural information processing systems, pages 554-560, 2000.
36. Ruigang Fu, Biao Li, Yinghui Gao, and Ping Wang. Content-based image retrieval based on cnn and svm. In 2016 2nd IEEE International Conference on Computer and Communications (ICCC), pages 638-642, 2016.
37. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
38. Manimala Singha and K Hemachandran. Content based image retrieval using color and texture. Signal and Image Processing, 3(1):39, 2012.
39. Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for-sun2019highand regions. arXiv preprint arXiv:1904.04514, 2019.
40. Christian Szegedy, Sergey Io e, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty- rst AAAI conference on articial intelligence, 2017.
41. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1-9, 2015.
42. Xiaolin Tian, Licheng Jiao, Xianlong Liu, and Xiaohua Zhang. Feature integration of eodh and color-sift: Application to image retrieval based on codebook. Signal Processing: Image Communication, 29(4):530-545, 2014.
43. Chuanqian Wang, Baochang Zhang, Zengchang Qin, and Junyi Xiong. Spatial weighting for bag-of-features based image retrieval. In International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, pages 91-100. Springer, 2013.
44. James Ze Wang, Jia Li, and Gio Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. IEEE Transactions on pattern analysis and machine intelligence, 23(9):947-963, 2001.
45. Sherin M Youssef. Ictedct-cbir: Integrating curvelet transform with enhanced dominant colors extraction and texture analysis for e cient content-based image retrieval. Computers and Electrical Engineering, 38(5):1358-1376, 2012.

46. Shan Zeng, Rui zeng2016imageWang, and Zhen Kang. Image retrieval using spatiograms of colors quantized by gaussian mixture models. Neurocomputing, 171:673-684, 2016.
47. Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 633-641, 2017.
48. DeTone, D., Malisiewicz, T., Rabinovich, A. (2018). Superpoint: Self-supervised interest point detection and description. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 224-236).
49. Paulin, M., Douze, M., Harchaoui, Z., Mairal, J., Perronin, F., Schmid, C. (2015). Local convolutional features with unsupervised training for image retrieval. In Proceedings of the IEEE international conference on computer vision (pp. 91-99).
50. Dubey, S. R., Singh, S. K., Singh, R. K. (2016). Multichannel decoded local binary patterns for content-based image retrieval. IEEE transactions on image processing, 25(9), 4018-4032.
51. Jun Li, Bo Yang, Wankou Yang, Changyin Sun, and Jianhua Xu. Subspace-based multiview fusion for instance-level image retrieval. The Visual Computer, pages 1–15, 2020.
52. Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep, 8, 2011.
53. Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High- resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514, 2019.
54. John Lambert, Liu Zhuang, Ozan Sener, James Hays, and Vladlen Koltun. MSeg: A composite dataset for multi-domain semantic segmentation. In Computer Vision and Pattern Recognition (CVPR), 2020.
55. Ville Viitaniemi et al. Image segmentation in content-based image re- trieval. Helsini University of Technology, Department of Electrical and Communications Engineering, Master Thesis, Finland 24th May, 2002.
56. Mustafa Ozden and Ediz Polat. A color image segmentation approach for content-based image retrieval. Pattern recognition, 40(4):1318–1325, 2007.
57. Mukherjee, Jit, Jayanta Mukhopadhyay, and Pabitra Mitra. "A survey on image retrieval performance of different bag of visual words indexing techniques." In Proceedings of the 2014 IEEE Students' Technology Symposium, pp. 99-104. IEEE, 2014.
58. Dubey, Shiv Ram. "A Decade Survey of Content Based Image Retrieval using Deep Learning." arXiv preprint arXiv:2012.00641 (2020).
59. Ramanjaneyulu, K., K. Veera Swamy, and CH Srinivasa Rao. "Novel CBIR System using CNN Architecture." In 2018 3rd International Conference on Inventive Computation Technologies (ICICT), pp. 379-383. IEEE, 2018.
60. Anandababu, P., and M. Kamarasan. "An Novel Framework For Content Based Image Retrieval With Quality Assessment System using Optimal Deep Convolution Neural Network. 2019"
61. Xie, Guangyi, Baolong Guo, Zhe Huang, Yan Zheng, and Yunyi Yan. "Combination of Dominant Color Descriptor and Hu Moments in Consistent Zone for Content Based Image Retrieval." IEEE Access 8 (2020): 146284-146299.
62. Bella, Mary I. Thusnavis, and A. Vasuki. "An efficient image retrieval framework using fused information feature." Computers and Electrical Engineering 75 (2019): 46-60.
63. Robey, Dennis, Wesley Thio, Herbert Iu, and Jason Eshraghian. "Naturalizing Neuromorphic Vision Event Streams Using GANs." arXiv preprint arXiv:2102.07243 (2021).
64. Yang, Alex, Charlie T. Veal, Derek T. Anderson, and Grant J. Scott. "Recognizing Image Objects by Relational Analysis Using Heterogeneous Superpixels and Deep Convolutional Features." arXiv preprint arXiv:1908.00669 (2019).
65. Chua, Tat-Seng, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. "Nus-wide: a real-world web image database from national university of singapore." In Proceedings of the ACM international conference on image and video retrieval, pp. 1-9. 2009.
66. Gordo, Albert, Jon Almazán, Jerome Revaud, and Diane Larlus. "Deep image retrieval: Learning global representations for image search." In European conference on computer vision, pp. 241-257. Springer, Cham, 2016.

67. Gordo, Albert, Jon Almazán, Jerome Revaud, and Diane Larlus. "Deep image retrieval: Learning global representations for image search." In European conference on computer vision, pp. 241-257. Springer, Cham, 2016.
68. Revaud, Jerome, Jon Almazán, Rafael S. Rezende, and Cesar Roberto de Souza. "Learning with average precision: Training image retrieval with a listwise loss." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5107-5116. 2019.
69. Ng, Tony, Vassileios Balntas, Yurun Tian, and Krystian Mikolajczyk. "SOLAR: second-order loss and attention for image retrieval." In European Conference on Computer Vision, pp. 253-270. Springer, Cham, 2020.
70. Brown, Andrew, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. "Smooth-AP: Smoothing the path towards large-scale image retrieval." In European Conference on Computer Vision, pp. 677-694. Springer, Cham, 2020.
71. Pautrat, Rémi, Viktor Larsson, Martin R. Oswald, and Marc Pollefeys. "Online Invariance Selection for Local Feature Descriptors." In European Conference on Computer Vision, pp. 707-724. Springer, Cham, 2020.
72. Tolias, Giorgos, Tomas Jenicek, and Ondřej Chum. "Learning and aggregating deep local descriptors for instance-level recognition." In European Conference on Computer Vision, pp. 460-477. Springer, Cham, 2020.
73. Tolias, Giorgos, Tomas Jenicek, and Ondřej Chum. "Learning and aggregating deep local descriptors for instance-level recognition." In European Conference on Computer Vision, pp. 460-477. Springer, Cham, 2020.
74. Tyszkiewicz, Michał J., Pascal Fua, and Eduard Trulls. "DISK: Learning local features with policy gradient." arXiv preprint arXiv:2006.13566 (2020).
75. Murala, Subrahmanyam, R. P. Maheshwari, and R. Balasubramanian. "Local tetra patterns: a new feature descriptor for content-based image retrieval." IEEE transactions on image processing 21, no. 5 (2012): 2874-2886.
76. Dubey, Shiv Ram, Satish Kumar Singh, and Rajat Kumar Singh. "Rotation and illumination invariant interleaved intensity order-based local descriptor." IEEE Transactions on Image Processing 23, no. 12 (2014): 5323-5333.
77. Dubey, Shiv Ram, Satish Kumar Singh, and Rajat Kumar Singh. "Local wavelet pattern: a new feature descriptor for image retrieval in medical CT databases." IEEE Transactions on Image Processing 24, no. 12 (2015): 5892-5903.
78. Liu, Yang, Lei Huang, Siqi Wang, Xianglong Liu, and Bo Lang. "Efficient segmentation for region-based image retrieval using edge integrated minimum spanning tree." In 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 1929-1934. IEEE, 2016.
79. Ouni, Achref, Eric Royer, Marc Chevaldonné, and Michel Dhome. "A Hybrid Approach for Improved Image Similarity Using Semantic Segmentation." In International Symposium on Visual Computing, pp. 647-657. Springer, Cham, 2020.
80. Vinayak, Vandana, and Sonika Jindal. "CBIR system using color moment and color auto-Correlogram with block truncation coding." International Journal of Computer Applications 161, no. 9 (2017): 1-7.
81. Lindeberg, Tony. "Scale invariant feature transform." (2012): 10491.
82. Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool. "Surf: Speeded up robust features." In European conference on computer vision, pp. 404-417. Springer, Berlin, Heidelberg, 2006.
83. Arandjelovic, Relja, and Andrew Zisserman. "All about VLAD." In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1578-1585. 2013.
84. Alcantarilla, Pablo Fernández, Adrien Bartoli, and Andrew J. Davison. "KAZE features." In European conference on computer vision, pp. 214-227. Springer, Berlin, Heidelberg, 2012.
85. Banerji, Sugata, Abhishek Verma, and Chengjun Liu. "Novel color LBP descriptors for scene and image texture classification." In Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV), p. 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2011
86. Dubey, Shiv Ram, Satish Kumar Singh, and Rajat Kumar Singh. "Multichannel decoded local binary patterns for content-based image retrieval." IEEE transactions on image processing 25, no. 9 (2016): 4018-4032..

87. Islam, SM Mohidul, and Rameswar Debnath. "An RST invariant image retrieval approach using color moments and wavelet packet entropy." In 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), pp. 747-752. IEEE, 2016.
88. Gong, Yunchao, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. "Deep convolutional ranking for multilabel image annotation." arXiv preprint arXiv:1312.4894 (2013).
89. Wang, Jiang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. "Cnn-rnn: A unified framework for multi-label image classification." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2285-2294. 2016.
90. Zhu, Feng, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. "Learning spatial regularization with image-level supervisions for multi-label image classification." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5513-5522. 2017.
91. Tian, Yurun, Bin Fan, and Fuchao Wu. "L2-net: Deep learning of discriminative patch descriptor in euclidean space." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 661-669. 2017.
92. Mishchuk, Anastasiya, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. "Working hard to know your neighbor's margins: Local descriptor learning loss." arXiv preprint arXiv:1705.10872 (2017).