Multimedia Tools and Applications (2023) 82:10145–10163
https://doi.org/10.1007/s11042-022-14297-7

1225: SENTIENT MULTIMEDIA SYSTEMS AND UNIVERSAL VISUAL LANGUAGES

An efficient ir approach based semantic segmentation

Achref Ouni[1] · Thierry Chateau[1] · Eric Royer[1] · Marc Chevaldonné[1] · Michel Dhome[1]

Received: 15 July 2021 / Revised: 14 November 2022 / Accepted: 3 December 2022 /
Published online: 28 December 2022

Abstract

Content Based Image Retrieval (CBIR) is the task of finding similar images from a query one. The state of the art mentions two main methods to solve the retrieval problem: (1) Methods dependent on visual description, for example, bag of visual words model (BoVW), Vector of Locally Aggregated Descriptors (VLAD) (2) Methods dependent on deep learning approaches in particular convolutional neural networks (CNN). In this article, we attempt to improve the CBIR algorithms with the proposition of two image signatures based on deep learning. In the first, we build a fast binary signature by utilizing a CNN based semantic segmentation. In the second, we combine the visual information with the semantic information to get a discriminative image signature denoted semantic bag of visual phrase. We study the performance of the proposed approach on six different public datasets: Wang, Corel 10k, GHIM-10K, MSRC-V1,MSRC-V2, Linnaeus. We significantly improve the mean of average precision scores (MAP) between 10% and 25% on almost all the datasets compared to state-of-the-art methods. Several experiments achieved on public datasets show that our proposal leads to increase the CBIR accuracy.

Keywords CBIR · Deep learning · Semantic segmentation · Image retrieval

1 Introduction

Content Based Image Retrieval (CBIR) is the task of retrieving in a dataset the images similar to an input query based on their contents. In addition CBIR is a fundamental step in

✉ Achref Ouni
Achref.ELOUNI@uca.fr

[1] CNRS, SIGMA Clermont, Institut Pascal, Université Clermont Auvergne, F-63000 Clermont-Ferrand, France

many computer vision applications such as pose estimation, virtual reality, remote sensing, crime detection, video analysis and military surveillance. In the medical field and more specifically in medical imaging, the search for content through the image can help to make a diagnosis by comparing an x-ray with previous cases being close to it. Current methods for image retrieval are efficient but can be further improved to have a quick search on large databases.

The state of the art mentions two main contributions used for image similarity: BoVW [15] (Bag of visual words) and CNN descriptors [28]. For retrieval, the images must be represented as numeric values. Both contributions represents images as vectors of valued features. This vector encodes the primitive image such as color, texture, and shape. BoVW encode each image by a histogram of the frequency of the visual words in the image. Deep learning is a set of machine learning methods attempting to model with a high level of data abstraction. Deep learning, learn features from input data (images in our case) using multiple layers for a specified task. Furthermore, deep learning has been used to solve many computer vision problems such as image and video recognition, image classification, medical image analysis, natural language processing... . Particularly Convolutional Neural Networks (CNN) have yielded an improvement on several image processing tasks.

In CNN-based CBIR approaches, the image signature is a vector (feature map) of N floats extracted from the feature layer (for example, the Fc7 layer for AlexNet [28]). The similarity between images is computed according the L2 distance between their signatures. When the dataset is large, the approximate nearest neighbor (ANN) search is used to speed up the computation. CNN based features used in existing CBIR works have been trained for classification problems. It is therefore invariant to the spatial position of objects. However CBIR applications should take care of the spatial position of semantic objects.

Semantic segmentation is a key step in many computer vision applications such as traffic control systems, video surveillance, video object co-segmentation and action localization, object detection and medical imaging. In CBIR models, the raw image should be transformed in a high level presentation. We argue that semantic segmentation networks, originally designed for other application can also be used for CBIR. We propose, in this paper to study how recent semantic segmentation networks can be used in CBIR context. Deep Learning based semantic segmentation networks output a 2D-map that associates a semantic label (class) to each pixel. This is a high level representation suitable for encoding a feature vector for CBIR that also encodes roughly spatial positions of objects. Two methodologies based on semantic segmentation are proposed in this paper with the aim to improve image representation. It is an extension of the work we initially proposed in [38], with a new image signature and a comprehensive study using extensive experience. Our contributions are as follows:

- The first approach transforms the semantic output (2D-map) into binary semantic descriptor. The descriptor which encodes the image integrates at the same time the semantic proportions of objects and their spatial positions.
- The second approach builds a semantic bag of visual phrase by combining the visual vocabulary with semantic information from the output of CNN network.

To test the performance of our framework we conducted the experimentation on six different databases. This article is structured as follows: we provide a brief overview of convolutional neural networks descriptors and bag of visual words related works in Section 2. We explain our proposals in Section 3. We present the experimental part on six different datasets and discuss the results of our work in Section 4.

2 State of the art

Many CBIR systems have been proposed in the last years [1, 9, 13, 19, 44, 64]. The content based image retrieval system (Fig. 1) receives as input a query image and returns a list of the most similar images in the database. The framework starts with the detection and extraction of the features and the signature construction step. Finally, the closest images to the input query found by the similarity measures between the images signature using L2 distance. We present a brief overview of approaches based on either visual and learning features.

2.1 Local visual features

Bag of Visual Words proposed by [15] is the most utilized model to classify the images by content. This methodology is made out of three principle steps: (I) Detection and Feature extraction (ii) Codebook generation (iii) Vector quantization. Recognition and extraction of features in an image can be performed utilizing extractor algorithms. Numerous descriptors have been proposed to encode the image into a vector. Scale Invariant Feature Transform (SIFT) [33] and Speeded-up Robust Features (SURF) [7] are the most utilized descriptors in image retrieval. From another point, parallel descriptors have demonstrated to be efficient. These descriptors based binary encoding of the features images. Rublee et al. [49] proposes ORB (Oriented FAST and Rotated BRIEF) to speed up the search. An other work [30] combines two aspects: accuracy and speed because of the BRISK (Binary Robust Invariant Scalable Keypoints) descriptor. Iakovidou et al. [23] presents a discriminant descriptor for image dependent on a mix of contour and color information.

In an offline stage, the codebook is generated from the collection off all descriptors from a training dataset. To do this, the K-MEANS approach is applied on the set of descriptors to obtain the visual words. The center of each cluster will be used as the visual word. Finally, for each image the histograms of the frequency of vocabularies or visual words, i.e. the image signature is created. Because of the limits of visual words approach numerous upgrades have been proposed for more accuracy. Bag of visual phrases (BoVP) is a significant level utilizing more than single word for representing an image. In [41], the proposed approach formed the phrases using a sequence of n-consecutive words regrouped by L2 metric. In [22], the authors proposed to link the visual words based on sliding window algorithm. Ren et al. [47] build an initial graph then split it into a fixed number of sub-graphs using the N-Cut algorithm. Every histogram of visual words in a sub-graph is a visual phrase. Chen et al. [12] link the visual words in pairs using the neighbourhood of each point of interest. Perronnin and Dance [43] apply Fisher Kernels to visual words represented by means of a Gaussian Mixture Model (GMM) and introduced a simplification for Fisher kernel. Similar to BoVW model, the vector of locally aggregated descriptors (VLAD) [24] affect to each feature or keypoint its nearest visual word and accumulates this

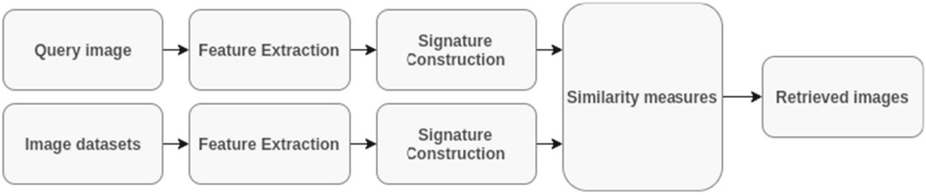


Fig. 1 General CBIR system architecture

difference for each visual word. Using ACP is frequent in CBIR applications thanks to its ability to reduce the descriptor dimension without losing its accuracy.

2.2 Learning-based features

First CNN models are interested in extracting the vector features or feature descriptor from the fully connected layer (AlexNet [28], VGGNet [51], GoogleNet [53] and ResNet [54]). For example, in AlexNet the size of the descriptor from the fc7 layer is 4,096. Similar to Local visual Feature approaches, after extracting all descriptors the retrieval is achieved using Euclidean distance between the signatures. Before being used to extract features, the CNN must be trained on large-scale datasets like ImageNet [16]. Inspired from VLAD, NetVLAD [3] is a CNN architecture used for image retrieval. Balaiah et al. [5] reduce the training time and provides an average improvement in accuracy. Fu et al. [21] use at the same time a convolutional neural network (CNN) and support vector machine (SVM) to solve the CBIR problem.

Recently, some networks have been developed especially for the CBIR task. Different models have been proposed such as generative adversarial networks [25, 52, 57], auto-encoder networks [20, 50, 59] and reinforcement learning networks [62, 63]. In [8] the VGG16 architecture is used for extricating significant elements and utilizing these components for the image retrieval task. The method proposed in [26] consists in exploring the use of CNN to determine a high dynamic range (HDR) image descriptor. In [46], a method is proposed to fine-tune CNNs for image retrieval on a large collection of unordered images in a fully automated manner. In [45] a CBIR system has been proposed based on transfer learning from a CNN trained on a large image database. The authors in [61] have used PCA Hashing method in combination with CNN features extracted by the fine-tuned model to improve the performance of CBIR tasks. In addition local detectors and descriptors [17, 34, 40, 55] based on CNN for CBIR task can also replace the classical features detection where each interest point is described by a vector.

Despite the speed of the approaches based on visual features and their good results on small datasets, they are still unable to find an image on a large scale database. Approaches based on deep learning have proven useful for both large and small datasets in term of accuracy and precision. While Deep Learning has many advantages, it also has its limits, including a huge need for computing power to ensure the maintenance of artificial neural networks, but also to process the very large amount of data required. In this article, we have tried to combine the discriminative power of two approaches in order to obtain more relevant results.

3 Methods

Encoding is the process of converting the data into a specified format for a specific task. In CBIR, encoding image content has met with great success. In addition, encoding images offers many advantages and benefits in terms of searching, retrieving and increasing the accuracy of CBIR system. Many approaches based on encoding such as BoVW [15], Fisher vector encoding [43], VLAD [24], CNN [28] achieve excellent performance. Consequently, encoding image content is a key element which leads to increase the CBIR system performance. Inspired by recent successes of deep learning, we propose two image signatures based on the use of CNN. In the following sections we will explain each one in more depth.

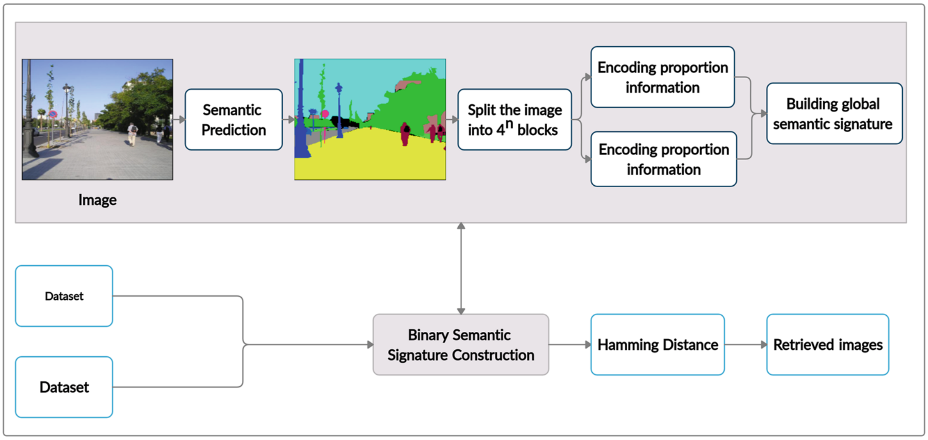


Image
Semantic Prediction
Split the image into 4^n blocks
Encoding proportion information
Encoding proportion information
Building global semantic signature
Dataset
Dataset
Binary Semantic Signature Construction
Hamming Distance
Retrieved images

Fig. 2 Different steps for building semantic binary signature

3.1 Semantic binary signature : SBS

Since the term similar means here with the same semantic content, we propose to explore in this section, an image signature that uses semantic segmentation networks, coupled with a binary spatial encoding. Such simple representation has several relevant properties: 1) It takes advantage of the state of the art semantic segmentation networks and 2) the proposed binary encoding allows a Hamming distance that requests a very low computation budget resulting to a fast CBIR method.

Given a semantic 2D-map, our method (Fig. 2) transforms the semantic prediction into a semantic binary signature. The signature construction comprises two main unsupervised processing units: (i) Encoding of spatial information (ii) Encoding of proportion. As shown in the upper part of the Fig. 2, given a query image I_q, we obtain the prediction I_{seg} using the semantic segmentation algorithm described in [58] in an offline stage. Then, we split the predicted I_{seg} into 4^n blocks I_{sub}. For each block, we encode both spatial and proportion information into a binary matrix. In order to obtain the two main components, we concatenate them to perform a discriminative semantic signature. The similarity between the images signatures are computed by $Hamming$ metric because this distance is fast for the comparison of binary data.

3.1.1 Encoding of spatial information

We propose to encode spatial information using a binary encoding. In a first stage, the image is divided in a recursive way (see Fig. 3). For level one, the image is split into 2 x 2 spatial areas without overlap that are denoted as blocks. The same operation is then achieved for each block (level 1), and so on. It results that for n levels, the recursive splitting process generates a set of $n_b = 4^n$ blocks. In a second stage, a binary vector is associated to each block. It is a simple way to encode spatial statistics and has been used for histogram based features for example. The binary vector we propose should provide information from existing semantic classes in the block: if a semantic class is present in the block, it is assigned a 1, otherwise a 0. We thus obtain a binary vector for each block that indicates the presence of semantic classes.

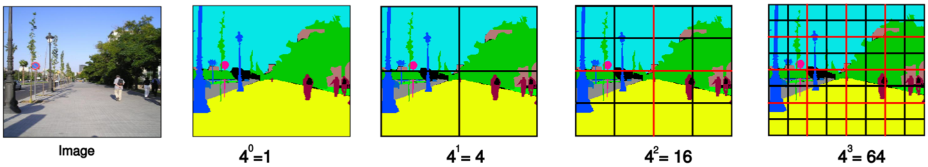


Fig. 3 Illustration of the spatial division. The semantic image divided into 4^n blocks

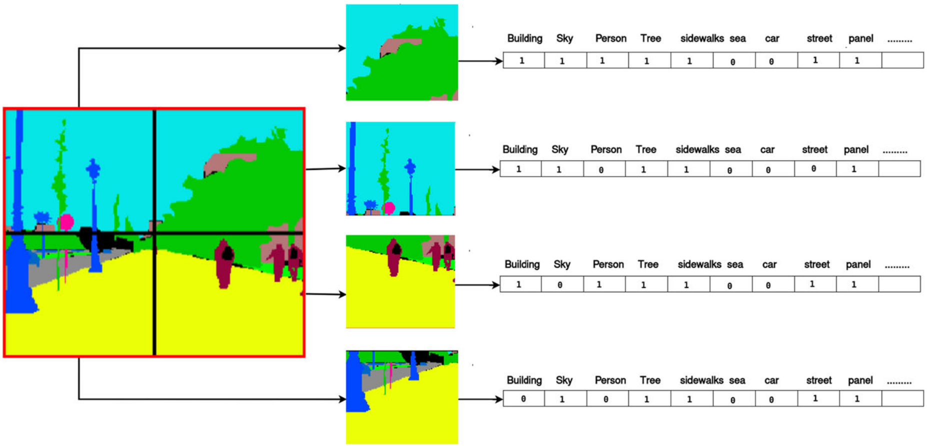
Figure 4 shows a spatial division into four blocks of the semantic image. A binary vector is assigned to each block to indicate the presence of semantic classes. Our example here shows by value 1 the presence of semantic classes such as sky, building, person, ... and by 0 the missing classes. The process of creating binary vectors stops when we obtain four vectors corresponding to the four blocks. Finally, we concatenate the binary vectors of all blocks to obtain the global signature S_s from an input image.

3.1.2 Encoding of proportion information

In the second step, we complete the binary spatial presentation with information on the proportion of each semantic class (Fig. 5). To do this, we propose to encode the proportion of semantic classes from the segmented image using the same spatial division used when encoding spatial information.

Given a segmented image I_{seg}, we detect the semantic classes present in each block using the neural network. Then, for each semantic class C_i we calculate its proportion as a percentage P_{C_i} in the block. After assigning the percentages of all the classes, a binary conversion process is applied to each P_{C_i} indicated in the (2):

$$\begin{cases} \text{if} \quad 0 < P_{C_i} <= 0.25 & \text{then} \quad BP_{C_i} = [0001] \\ \text{if} \quad 0.25 < P_{C_i} <= 0.5 & \text{then} \quad BP_{C_i} = [0011] \\ \text{if} \quad 0.5 < P_{C_i} <= 0.75 & \text{then} \quad BP_{C_i} = [0111] \\ \text{if} \quad P_{C_i} > 0.75 & \text{then} \quad BP_{C_i} = [1111] \end{cases} \tag{1}$$



Building	Sky	Person	Tree	sidewalks	sea	car	street	panel	....
1	1	1	1	1	0	0	1	1	

Building	Sky	Person	Tree	sidewalks	sea	car	street	panel	....
1	1	0	1	1	0	0	0	1	

Building	Sky	Person	Tree	sidewalks	sea	car	street	panel	....
1	0	1	1	1	0	0	1	1	

Building	Sky	Person	Tree	sidewalks	sea	car	street	panel	....
0	1	0	1	1	0	0	1	1	

Fig. 4 An example of converting a semantic block to a semantic binary vector

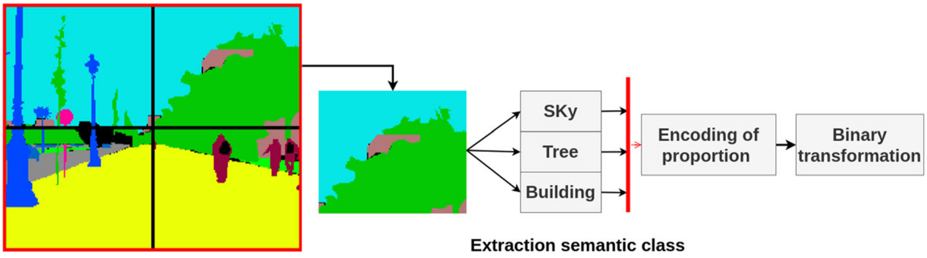


Fig. 5 Example of encoding the proportion information. Given an image divided into 4 blocks, we iteratively select each block to calculate the proportion of the semantic class inside

For cases where the semantic class C_i is not present in the block, the bit string $BP_{C_i} =$ [0000] is automatically assigned. In order to keep all the scores, we collect them together in the bit string named $BSub_j$ which is a binary description of the proportion of the classes in the block number j:

$$BSub_j = [BP_{C_1}\ BP_{C_2} \ldots BP_{C_M}]$$

where M is the number of classes that the network has learned to detect. Finally, we concatenate all the bit strings $BSub_j$ to obtain a signature of global proportion S_P corresponding to the segmented input image I_{seg} where $S_P = [BSub_1\ BSub_2 \ldots BSub_{n_b}]$. We start the tests with large blocks, then we repeat them with smaller blocks. When $n_b = 1$ it means that no spatial division was applied on the image. Therefore, we only encode the semantic proportion information of the whole image. Finally, the binary signature S of an image is the bit string $[S_S\ S_P]$.

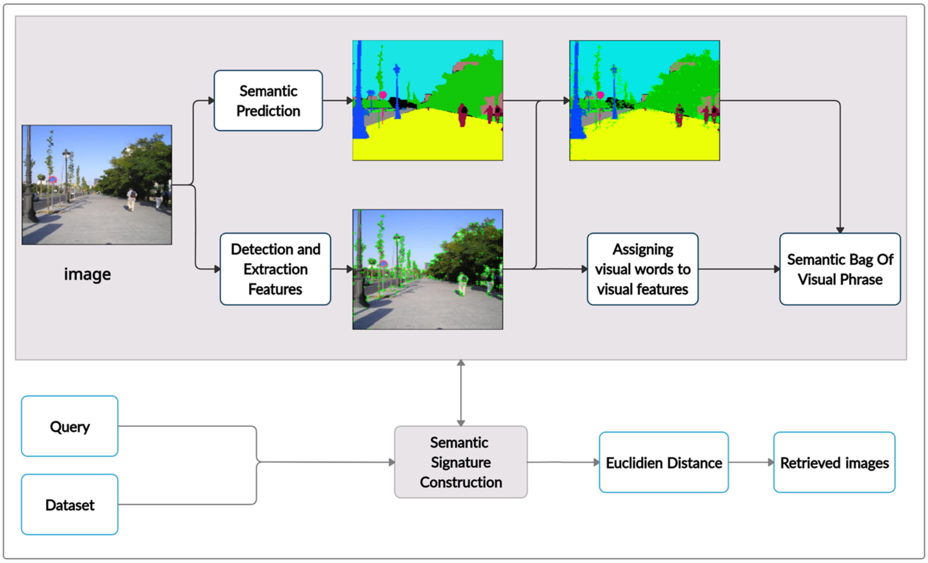

Semantic Prediction
image
Detection and Extraction Features
Assigning visual words to visual features
Semantic Bag Of Visual Phrase
Query
Dataset
Semantic Signature Construction
Euclidien Distance
Retrieved images

Fig. 6 Different steps for building semantic bag of visual phrase signature

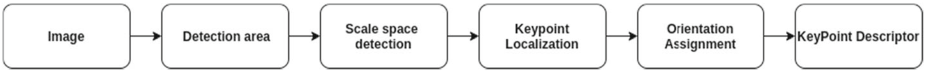


Fig. 7 Flow-chart of features extraction

3.2 Semantic bag of visual phrase: SBOVP

Based on the semantic segmentation output (2D-map), we propose in this part an efficient image signature combining the bag of visual phrase and semantic segmentation. As shown in Fig. 6, we start by constructing the images signatures for both query and dataset. Our signature join semantic data with visual features to improve the image representation without prior knowledge. We compute the similarity between the signature of the query and the signature of each image in the dataset utilizing the euclidean distance (DL2). Then, the candidates with lowest distance are considered the most similar to the input query. We will clarify our methodology in detail in the following.

Bag of visual phrase is an improved version of bag of visual words model and a high-level image description utilizing more than one word. Therefore, a visual phrase is a set of words linked together. Various methods have been proposed [22, 39, 47] in the state of the art that are able to construct visual phrases by different manners (Clustering, Graphs, Regions, KNN by metric, etc). The primary burden of the proposed strategies is that they do not take into consideration the spatial position of semantic objects.

The proposed bag of visual phrase algorithm uses deep learning, in particular semantic segmentation, to link the visual words in the image. We attempt by the Fig. 9 to clarify in detail the signature construction steps. We start by two parallel processes: (1) Extraction of semantic information (2D-map) using the semantic segmentation algorithm and features detection then extraction (see Fig. 7) using visual descriptors (KAZE/SURF in our case). Next, we project the location of keypoints on the 2D-map to assign a class label to each keypoint (see Fig. 8). In parallel, we assign to each keypoint from an image the visual word VW_j with the lowest distance using (2) (Fig. 9).

$$\|d_{kp_i} - VW_j\|_{L_2} = \sqrt{\sum_{d=1}^{dim}(d_{kp_i}(d) - VWj(d))^2} \tag{2}$$

where dim is the dimension of descriptor (64 in our case) and d_{kp_i} is the descriptor of the keypoint number i.

The visual vocabulary or visual words are computed in an offline stage using the K-MEANS [27] algorithm trained on Pascal Voc dataset. At this point each keypoint is described by two main components: class label C_i and visual words VW_j. We obtain at this stage a discriminative keypoint description combining visual and semantic information.

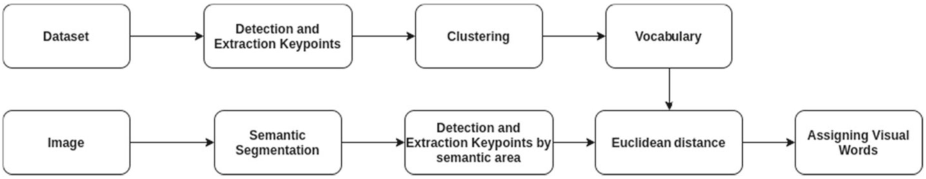


Fig. 8 Flow-chart of assigning semantic visual words

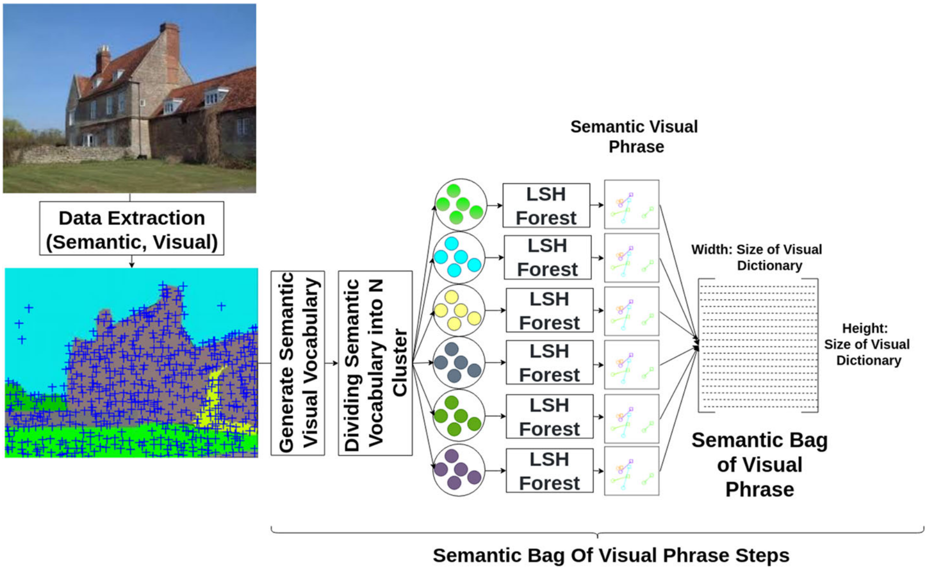

Data Extraction (Semantic, Visual)
Generate Semantic Visual Vocabulary
Dividing Semantic Vocabulary into N Cluster
LSH Forest
LSH Forest
LSH Forest
LSH Forest
LSH Forest
LSH Forest
Semantic Visual Phrase
Width: Size of Visual Dictionary
Height: Size of Visual Dictionary
Semantic Bag of Visual Phrase
Semantic Bag Of Visual Phrase Steps

Fig. 9 Flow-chart of semantic bag of visual phrase

Next step, we divide the obtained semantic visual words into N regions corresponding to the semantic classes predicted by the trained CNN. We confirm here that the keypoints are grouped by semantic criteria in which each region represents an object in the image. For each region, we construct the visual phrases based on the semantic visual words inside. Then, for each visual word VW_i in the region, we link it to its nearest neighbor using approximate nearest neighbor (ANN) algorithm (LSH forest [6]) to obtain a visual phrase (VW_i, VW_j). The main gain of using ANN algorithm in the signature construction process is to reduce the complexity of searching time compared to brute force algorithm especially when the region contains an exponential number of visual words VW_i.

In the bag of visual phrase approach, the image signature is an upper triangular matrix H of dimensions $L \times L$ with L the number of visual words in the codebook. This matrix plays a role similar to the histogram in the bag of words approach. The

Table 1 Database used to evaluate of approach

Name	Size DB / Queries	Ground truth	Query mode
Corel 1K [56] (Wang)	1000 / 1000	100	query-in-ground Truth
Corel 10K [56]	10.000 / 10.000	100	query-in-ground Truth
GHIM-10K [56]	10.000 / 10.000	500	query-in-ground Truth
Linnaeus [11]	6000 / 2000	400	queries/ dataset are disjoint
MSRC v1	241 / 241	–	query-in-ground Truth
MSRC v2	591 / 591	–	query-in-ground Truth

matrix H is initialized at zero. Then for each visual phrase composed of a set of visual words $S = \{VW_{i_1}, VW_{i_2}, ..., VW_{i_n}\}$, we increment the values $H(i_{k_1}, i_{k_2})$ for each pair of $\{VW_{k_1}, VW_{k_2}\} \subset S$. The last step is to select the candidates from the dataset that are similar to an input query depending on the distance between their signatures according to (3):

$$D(H_1, H_2) = \sqrt{\sum_{i=1}^{L} \sum_{j=i}^{L} (H_1(i, j) - H_2(i, j))^2} \tag{3}$$

4 Experimental protocol

4.1 Benchmark datasets for retrieval

In this section, we present the potential of our approach on six different datasets (Table 1). 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 [56] or Wang is a dataset of 1000 images divided into 10 categories and each category contains 100 images. The evaluation is done by computing the average precision of the first 100 nearest neighbors among 1000.
- Corel 10K [31] is a dataset of 10000 images divided into 100 categories and each category contains 100 images. The evaluation is done by computing the average precision of the first 100 nearest neighbors among 10000.
- GHIM-10K [31] is a dataset of 10000 images divided into 20 categories and each category contains 500 images. The evaluation is done by computing the average precision of the first 500 nearest neighbors among 10000.
- 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 (Microsoft Research in Cambridge) 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 [11] is a new dataset composed of 8000 images of 4 categories (berry, bird, dog, flower). The evaluation on Linnaeus is based on MAP score (mean average precision)

4.2 Deep learning methodology for semantic segmentation

In the last years, many architectures have been proposed for image segmentation such as Hourglass [36], SegNet [4], DeconvNet [37], U-Net [48], SimpleBaseline [60] and encoder-decoder [42]. The existing approaches propose to encode the input image as a low-resolution representation by connecting high to low resolution convolutions in series and then recover the high-resolution representation from the encoded low-resolution representation. In this

Table 2 Details about semantic dataset used to train the network

Dataset	Images	Merged classes	All classes	Stuff / Thing classes	Year
Coco-stuff [10]	164K	172	172	92 / 80	2018
Mseg [29]	220K	194	316	102 / 94	2020

work, we use a trained architecture, namely High-Resolution Net (HRNet) [58]. The advantage of using HRNet is that the resulting representation is semantically richer and spatially more exact which allows to maintain high-resolution representations. Then, to segment an image we apply the HRNet model pretrained on multiple datasets cited in Table 2 to obtain the class label of each pixel in the image.

HRNet [58] use the SGD optimizer [18] with the base learning rate of 0.01, the momentum of 0.9 and the weight decay of 0.0005. The poly learning rate policy with the power of 0.9 is used for dropping the learning rate. All the models are trained for 120K iterations (epochs) with the batch size of 12 on 4 GPUs and syncBN. As stated in [58], the inference time cost is around 0.15 s per batch for an input size 1024 × 2048 and a batch size $bs = 1$ on a V100 GPU card, which is 2 to 3 times faster than competing models.

4.3 Training datasets for semantic segmentation

Many semantic segmentation datasets have been proposed in last years such as Cityscapes [14], Mapillary [35], COCO [32], ADE20K [65], Coco-stuff [10], Mseg [29] and others. In this work, we use the recent implementation HRNet-W48 [58] architecture trained on Coco-stuff [10] and Mseg [29] datasets. The main advantage of using Coco-stuff [10] and Mseg [29] datasets (Table 2) is that they are able to handle both thing and stuff objects. Thing objects have characteristic shapes like vehicle, dog, computer... and stuff objects is the description of amorphous objects like sea, sky, tree,... .

5 Results

5.1 Results on benchmark datasets for retrieval

We conducted our experimentation on two different semantic prediction datasets [10, 29] and six retrieval datasets (Table 1). Table 3 presents the mean average precision (MAP) [2] scores for dataset per size of blocks for the semantic binary signature. We conduct the tests by starting with large blocks then going to small blocks. When the parameter $n = 1$, the encoding of semantic spatial information is not done and we encode only the semantic proportion information. We notice that the performance (MAP) increase in Table 3, as the number of blocks increases.

$$MAP = \frac{1}{n} \sum_{k=1}^{n=k} AP_k \tag{4}$$

where AP_k= the AP of class K and n= the number of classes.

The $Hamming$ distance is the similarity metric used to compute the similarity between the query and dataset for the semantic binary signature (Table 4).

Table 3 MAP evaluations for semantic binary signature using Mseg and Coco-stuff datasets (best scores in bold)

Number of blocks Retrieval dataset	$4^0 = 1$	$4^1 = 4$	$4^2 = 16$	$4^3 = 64$
Semantic Dataset : Mseg [29]				
MSRC v1	0.79	0.83	0.81	**0.89**
MSRC v2	0.64	0.71	0.67	**0.73**
Linnaeus [11]	0.71	**0.78**	0.73	0.77
Corel 1K(Wang) [56]	0.77	0.81	0.80	**0.86**
Corel 10K [31]	0.53	0.56	0.56	**0.57**
GHIM-10K [31]	0.53	0.53	0.54	**0.55**
Semantic Dataset : Coco-stuff [10]				
MSRC v1	0.75	0.81	0.79	**0.83**
MSRC v2	0.61	0.62	0.61	**0.66**
Linnaeus [11]	0.58	0.66	0.64	**0.68**
Corel 1K(Wang) [56]	0.76	0.78	0.75	**0.82**
Corel 10K [31]	0.48	0.47	0.46	**0.49**
GHIM-10K [31]	0.47	**0.48**	0.46	**0.48**

In Table 5, we present the quantitative MAP results utilizing the semantic bag of visual phrase signature (SBOVP) on the retrieval dataset (see Table 1). We test after training the semantic segmentation network on two different semantic datasets (Mseg, Coco-Stuff). In addition, for each semantic dataset we have utilized two different visual descriptors to test our mixture approach among visual and semantic data. The first descriptor used is SURF descriptor. It is a feature detection algorithm and also a descriptor. SURF is partly inspired by the SIFT descriptor, which it surpasses in speed and, according to its authors, in robustness for different image transformations. The second descriptor is Kaze. It is a new method inspired by the SIFT descriptor. The KAZE method is a multi-scale algorithm for detecting and describing 2D features. We notice that the prediction obtained using Mseg dataset is better in terms of score than Coco-stuff.

We study in Fig. 10 the effect of increasing the number of visual words during the process of visual phrase construction and its impact on MAP score. In the definition, the visual phrase is built by at least two visual words. In the experiments, we test the effect of visual

Table 4 Execution time in milliseconds (**ms**) per image (using a single thread) for all datasets

Number of blocks Retrieval dataset	$4^0 = 1$	$4^1 = 4$	$4^2 = 16$	$4^3 = 64$
MSRC v1	8.8	9.1	12.8	28.1
MSRC v2	8.5	9.8	13.6	30.6
Linnaeus [11]	9.1	11.3	18.6	37.6
Corel 1K(Wang) [56]	10.1	14.3	29.5	41.6
Corel 10K [31]	10.4	14.5	28.9	42.1
GHIM-10K [31]	11.2	15.4	30.1	44.2

Table 5 MAP evaluations for semantic bag of visual phrase signature using Mseg and Coco-stuff datasets (best scores in bold)

Method / Retrieval dataset	SBOVP(Mseg [29])		SBOVP(Coco-stuff [10])	
	SURF	KAZE	SURF	KAZE
MSRC v1	0.90	**0.91**	0.89	0.88
MSRC v2	0.71	0.73	0.74	**0.75**
Linnaeus [11]	0.79	**0.80**	0.77	0.78
Corel 1K(Wang) [56]	**0.89**	0.88	0.87	0.85
Corel 10K [31]	0.61	**0.63**	0.60	0.61
GHIM-10K [31]	**0.57**	0.55	0.56	0.54

phrase made when the number n of visual words is comprised between 2 and 4. There is little difference in the MAP score between $n = 2$ and $n = 3$. However, augmenting the value of n to 4 produces noise and negatively affects the robustness of the constructed visual phrase.

5.2 Comparison with state-of-the-art

We compare our results with two main categories of approaches: (i) Local visual Feature: methods based on local features like Surf, Sift included the inherited methods such as BoVW, Vlad, Fisher. (ii) Learning based features: methods based on learning the features using deep learning algorithms. In Table 6 we compare our results with several state of the art methods and we highlight in bold the best MAP score. As can be seen, our proposed method present good performance on nearly all datasets. Except on Corel 10K dataset in which working in ResNet [54] gives a better result than us because of training the model on one million image (ImageNet). In addition, our approach combine between visual and semantic information which gives us good performance.

For methods which use deep learning, signatures are built on the basis of information provided by the various layers from the architecture. In other side, the methods that use visual features, the signatures are constructed based on the position of the interest points

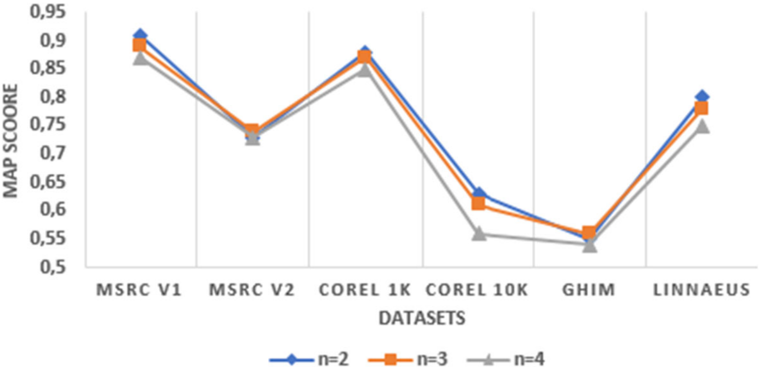


Fig. 10 Investigation of the impact of parameters n, number of visual words in phrases

Table 6 Comparison of the accuracy of our approach with methods from the state of the art (best scores in bold)

Methods	MSRC v1	MSRC v2	Linnaeus	Wang	Corel-10K	GHIM-10K
BoVW [15]	0.48	0.30	0,26	0.48	0.30	0.39
n-BoVW [39]	0.58	0.39	0.31	0.60	0.34	0.41
VLAD [24]	0.78	0.41	–	0.74	0.38	0.44
N-Gram [41]	–	–	–	0.37	–	–
AlexNet [28]	0.81	0.58	0.47	0.68	0.40	0.52
VGGNet [51]	0.76	0.63	0.48	0.76	0.45	0.57
ResNet [54]	0.83	0.70	0.69	0.82	**0.59**	0.62
SaCoCo [23]	–	–	–	0.54	0.17	0.15
Ruigang [21]	–	–	0.70	–	–	–
Ayan [9]	–	–	–	0.79	0.52	–
Chu [13]	–	–	–	0.80	0.45	0.51
Ours(best)	**0.91**	**0.75**	**0.80**	**0.89**	0.57	**0.63**

and the robustness of the visual descriptors used for extracting the keypoints. In our work, we achieve the signature image using the semantic information in the first proposal and combining the visual features with the semantic information for the second proposal. Thus, through discriminative information, we have successfully built two robust image signatures.

For any CBIR system the execution time depends on the time needed for the signature construction. The main desired objective of the semantic binary signature is its ability to reduce and minimize the execution time of CBIR. We compare only the time taken by each method to build its signature. We want to highlight here that the extraction, detection and semantic segmentation time are not taken into consideration for all the compared methods. Figure 11 presents a comparison of the time needed for the signature construction for the state of the art methods and our semantic binary signature method. The low computation time is a strong advantage of our method. Moreover, the time required for the computation of the distance between signatures is also very low because we use the fast Hamming distance.

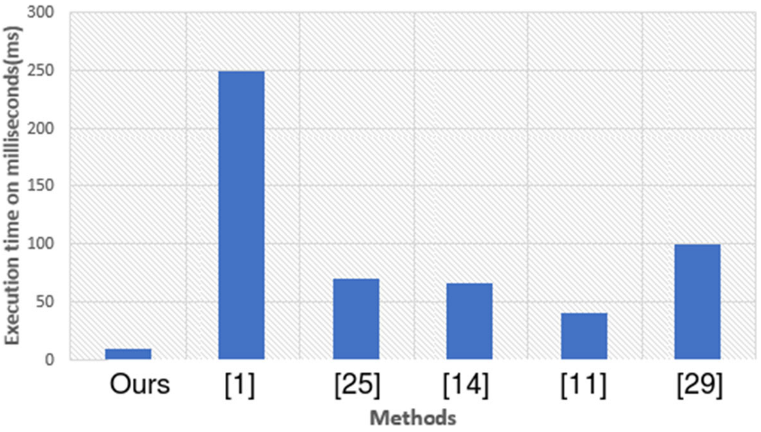


Fig. 11 Comparison of execution time between semantic binary signature and the state of the art

Table 7 Comparison of MAP for the top 20 retrieved images on the Wang dataset with the state of the art methods

Methods	Top 20
ElAlami [19]	0.76
Guo and Prasetyo [1]	0.77
Zeng et al. [64]	0.80
Jitesh Pradhan [44]	0.81
Proposed method	0.94

Our method is the semantic bag of visual phrase

In Table 7, we compare based semantic bag of visual phrase the MAP score of the top 20 retrieved image for all categories for Wang dataset. In Figure 12 we show the mean precision (AP) performance of top 20 retrieved image for 10 category compared to [1, 19, 44, 64] methods.

Figure 13 clearly indicates how the semantic binary signature is able to select the similar images to the input query based on the semantic content. The selection is based on the hamming distance between the query and the image dataset. Experiments with a single thread for each image, the descriptor requires 9 ms on average (Table 4).

6 Discussion

The main benefit of our framework is its ability to construct an image signature quickly and with a low complexity. The computation time of the retrieval of binary signatures is clearly less than the semantic bag of visual phrase due to the binary encoding. The histogram construction process takes 5 times less than the state of the art methods and 3 times less than the second proposed method. As we obtain an image signature combining semantic and visual features (SBOVP), the increase in complexity of the construction of the image signature has an effect in the global retrieval process in terms of time. In addition, the results are good because of the mix between semantic and visual data. The time of the

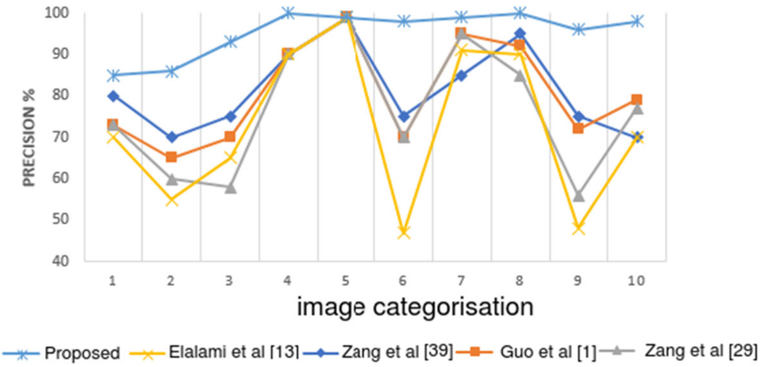


Fig. 12 Comparison of precision for the top 20 retrieved images for all categories (Corel 1K (Wang) dataset) using semantic bag of visual phrase method

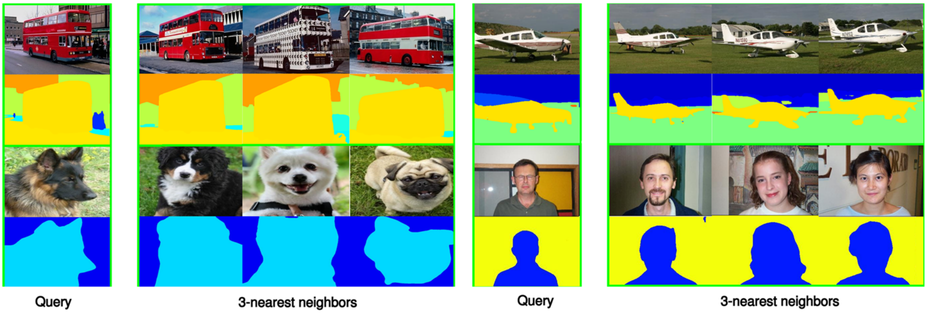


Fig. 13 From different categories selected from different datasets, we show the queries with their corresponding segmentation and the three nearest neighbors selected by our method using HRNet-W48 [58] trained on Mseg dataset

signature construction step depends on the number of visual words that are linked together for obtaining the visual phrase. The ideal case is when the visual phrase is constructed with two visual words.

In our work the image signatures mainly depend on the semantic objects detected by the neural network trained on thousands of labelled objects in images from a large dataset. Then the key step is converting the semantic output to numeric values for the retrieval step. In some cases the retrieval process cannot find the exact results because the test images contain new objects which were not present in the training dataset. The main disadvantage here is that most libraries of deep neural networks such as Pytorch, tensorflow are implemented for GPUs. It is an expensive graphics processor that performs fast mathematical calculations, mainly for image rendering.

7 Conclusion

We have presented in this paper two different image signatures based on deep learning. We exhibit that the use of semantic segmentation in the CBIR subject can improve the recovery of images. In the first contribution, we have shown that by encoding the image information as binary leads to improve the CBIR accuracy and reduce the execution time. In the second, we combined the visual information with the semantic information to build a discriminative signature. Indeed, even the second signature is better in terms of precision, the first signature is faster to classify the images based on semantic content. The experimental evaluation indicates that our approach achieve a better results in terms of accuracy and time compared to the state of the art methods.

Data Availability The datasets analysed during the current study are available on these web pages :
MSRC-v2 : https://figshare.com/articles/dataset/MSRC-v2imagedataset/6075788
MSRC-v1 : https://mldta.com/dataset/msrc-v1/home/
Linnaeus : http://chaladze.com/l5/
Wang : https://sites.google.com/site/dctresearch/Home/content-based-image-retrieval
Corel-10K : http://wang.ist.psu.edu/docs/related/
GHIM-10K : https://www.kaggle.com/datasets/guohey/ghim10k

Declarations

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

References

1. Admile NS, Dhawan RR (2016) Content based image retrieval using feature extracted from dot diffusion block truncation coding. In: International conference on communication and electronics systems (ICCES), IEEE, pp 1–6
2. Angelopoulou E, Boutalis YS, Iakovidou C, Chatzichristofis SA (2014) Mean normalized retrieval order (mnro) : a new content-based image retrieval performance measure
3. Arandjelović R, Gronat P, Torii A, Pajdla T, Sivic J (2016) NetVLAD : CNN architecture for weakly supervised place recognition . In: IEEE conference on computer vision and pattern recognition
4. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet : a deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence 39(12):2481–2495
5. Balaiah T, Jeyadoss TJT, Thirumurugan SS, Ravi RC (2019) A deep learning framework for automated transfer learning of neural networks. In: 2019 11th international conference on advanced computing (ICoAC), IEEE, pp 428–432
6. Bawa M, Condie T, Ganesan P (2005) Lsh forest : self-tuning indexes for similarity search. In: Proceedings of the 14th international conference on World Wide Web, pp 651–660
7. Bay H, Tuytelaars T, Gool LV (2006) Surf : speeded up robust features. In: European conference on computer vision, Springer, pp 404–417
8. Bhandi V, Devi KS (2019) Image retrieval by fusion of features from pre-trained deep convolution neural networks . In: 2019 1st international conference on advanced technologies in intelligent control, environment, computing & communication engineering (ICATIECE), IEEE, pp 35–40
9. Bhunia AK, Bhattacharyya A, Banerjee P, Roy PP, Murala S (2019) A novel feature descriptor for image retrieval by combining modified color histogram and diagonally symmetric co-occurrence texture pattern. Pattern Anal Applic, 1–21
10. Caesar H, Uijlings J, Ferrari V (2018) Coco-stuff : thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1209–1218
11. Chaladze G, Kalatozishvili L (2017) Linnaeus 5 dataset for machine learning. Technical Report Tech. Rep
12. Chen T, Yap K-H, Zhang D (2014) Discriminative soft bag-of-visual phrase for mobile landmark recognition. IEEE Trans Multimedia 16(3):612–622
13. Chu K, Liu G-H (2020) Image retrieval based on a multi-integration features model. Math Probl Eng, 2020
14. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3213–3223
15. Csurka G, Dance C, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV, vol 1, Prague, pp 1–2
16. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet : a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 248–255
17. 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
18. Duda J (2019) Sgd momentum optimizer with step estimation by online parabola model. arXiv:1907.07063
19. Esmel ElAlami M (2014) A new matching strategy for content based image retrieval system. Appl Soft Comput 14:407–418
20. Feng F, Wang X, Li R (2014) Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the 22nd ACM international conference on multimedia, pp 7–16
21. Fu R, Li B, Gao Y, Wang P (2016) Content-based image retrieval based on cnn and svm. In: 2016 2nd IEEE international conference on computer and communications (ICCC), pp 638–642
22. Ginn D, Mendes A, Chalup S, Chen Z (2018) Sliding window bag-of-visual-words for low computational power robotics scene matching. In: 2018 4th international conference on control, automation and robotics (ICCAR), IEEE, pp 88–93
23. Iakovidou C, Anagnostopoulos N, Lux M, Christodoulou K, Boutalis Y, Chatzichristofis SA (2019) Composite description based on salient contours and color information for cbir tasks. IEEE Trans Image Process 28(6):3115–3129

24. Jégou H, Douze M, Schmid C, Pérez P (2010) Aggregating local descriptors into a compact image representation. In: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE, pp 3304–3311
25. Jin S, Zhou S, Liu Y, Chen C, Sun X, Yao H, Hua X-S (2020) Ssah : semi-supervised adversarial deep hashing with self-paced hard sample generation. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 11157–11164
26. Khwildi R, Zaid AO, Dufaux F (2021) Query-by-example hdr image retrieval based on cnn. Multimed Tools Appl 80(10):15413–15428
27. Krishna K, Murty MN (1999) Genetic k-means algorithm. IEEE Trans Syst Man Cybern , Part B (Cybernetics) 29(3):433–439
28. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
29. Lambert J, Zhuang L, Sener O, Hays J, Koltun V (2020) MSeg : a composite dataset for multi-domain semantic segmentation. In: Computer vision and pattern recognition (CVPR)
30. Leutenegger S, Chli M, Siegwart RY (2011) Brisk : binary robust invariant scalable keypoints. In: 2011 IEEE International conference on computer vision (ICCV), IEEE, pp 2548–2555
31. Li J, Wang JZ (2003) Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans Pattern Anal Mach Intell 25(9):1075–1088
32. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco : common objects in context. In: European conference on computer vision, Springer, pp 740–755
33. Lowe DG (1999) Object recognition from local scale-invariant features. In: Proceedings of the seventh IEEE international conference on computer vision, vol 2. IEEE, pp 1150–1157
34. Mishchuk A, Mishkin D, Radenovic F, Matas J (2017) Working hard to know your neighbor's margins : local descriptor learning loss. In: Advances in neural information processing systems, pp 4826–4837
35. Neuhold G, Ollmann T, Bulo SR, Kontschieder P (2017) The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE international conference on computer vision, pp 4990–4999
36. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: European conference on computer vision, Springer, pp 483–499
37. Noh H, Hong S, Han B (2015) Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 1520–1528
38. Ouni A, Chateau T, Royer E, Chevaldonné M, Dhome M (2022) A new cbir model using semantic segmentation and fast spatial binary encoding. In: Conference on computational collective intelligence technologies and applications, Springer, pages 437–449
39. Ouni A, Urruty T, Visani M (2018) A robust cbir framework in between bags of visual words and phrases models for specific image datasets. Multimed Tools Appl 77(20):26173–26189
40. 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
41. Pedrosa GV, Traina AJ (2013) From bag-of-visual-words to bag-of-visual-phrases using n-grams. In: XXVI conference on graphics, patterns and images, IEEE, pp 304–311
42. Peng X, Feris RS, Wang X, Metaxas DN (2016) A recurrent encoder-decoder network for sequential face alignment. In: European conference on computer vision, Springer, pp 38–56
43. Perronnin F, Dance C (2007) Fisher kernels on visual vocabularies for image categorization. In: IEEE conference on computer vision and pattern recognition, IEEE, pp 1–8
44. Pradhan J, Kumar S, Pal AK, Banka H (2018) Texture and color visual features based cbir using 2d dt-cwt and histograms. In: International conference on mathematics and computing, Springer, pp 84–96
45. Putzu L, Piras L, Giacinto G (2020) Convolutional neural networks for relevance feedback in content based image retrieval. Multimed Tools Appl 79(37):26995–27021
46. Radenović F, Tolias G, Chum O (2018) Fine-tuning cnn image retrieval with no human annotation. IEEE Trans Pattern Anal Mach Intell 41(7):1655–1668
47. Ren Y, Bugeau A, Benois-Pineau J (2013) Visual object retrieval by graph features
48. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 234–241
49. Rublee E, Rabaud V, Konolige K, Bradski G (2011) Orb : an efficient alternative to sift or surf. In: 2011 IEEE international conference on computer vision (ICCV), IEEE, pp 2564–2571
50. Shen Y, Qin J, Chen J, Yu M, Liu L, Zhu F, Shen F, Shao L (2020) Auto-encoding twin-bottleneck hashing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2818–2827

51. Simonyan K, Zisserman A (2014)
52. Song J, He T, Gao L, Xu X, Hanjalic A, Shen HT (2018) Binary generative adversarial networks for image retrieval. In: Thirty-second AAAI conference on artificial intelligence
53. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
54. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence
55. Tian Y, Fan B, Wu F (2017) 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
56. Wang JZ, Li J, Wiederhold G (2001) Simplicity : semantics-sensitive integrated matching for picture libraries. IEEE Trans Pattern Anal Mach Intell 23(9):947–963
57. Wang G, Hu Q, Cheng J, Hou Z (2018) Semi-supervised generative adversarial hashing for image retrieval. In: Proceedings of the European conference on computer vision (ECCV), pp 469–485
58. Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, Liu D, Mu Y, Tan M, Wang X et al (2020) Deep high-resolution representation learning for visual recognition. IEEE Trans Pattern Anal Mach Intell
59. Wu P, Hoi SC, Hao X, Zhao P, Wang D, Miao C (2013) Online multimodal deep similarity learning with application to image retrieval. In: Proceedings of the 21st ACM international conference on Multimedia, pp 153–162
60. Xiao B, Wu H, Wei Y (2018) Simple baselines for human pose estimation and tracking. In: Proceedings of the European conference on computer vision (ECCV), pp 466–481
61. Yang Z, Yue J, Li Z, Zhu L (2018) Vegetable image retrieval with fine-tuning vgg model and image hash. IFAC-PapersOnLine 51(17):280–285
62. Yang J, Zhang Y, Feng R, Zhang T, Fan W (2020) Deep reinforcement hashing with redundancy elimination for effective image retrieval. Pattern Recogn 100:107116
63. Yuan X, Ren L, Lu J, Zhou J (2018) Relaxation-free deep hashing via policy gradient. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 134–150
64. Zeng S, Huang R, Wang H, Kang Z (2016) Image retrieval using spatiograms of colors quantized by gaussian mixture models. Neurocomputing 171:673–684
65. Zhou B, Zhao H, Puig X, Fidler S, Barriuso A, Torralba A (2017) Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 633–641

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.