A New CBIR Model Using Semantic Segmentation and Fast Spatial Binary Encoding

Achref Ouni, Thierry Chateau, Eric Royer, Marc Chevaldonné, Michel Dhome

A New CBIR Model Using Semantic Segmentation and Fast Spatial Binary Encoding

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

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

Abstract. Content Based Image Retrieval(CBIR) is the task of finding similar images from a query one. Since the term similar means here "with the same semantic content", we propose to explore in this paper, a framework that uses Deep Neural Networks based 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. Several experiments achieved on public datasets show that our binary semantic signature leads to increase the CBIR accuracy and reduce the execution time. We study the performance of the proposed approach on six different public datasets: Wang, Corel 10k, GHIM-10K, MSRC-V1,MSRC-V2, Linnaeus.

Keywords: CBIR · Deep learning · Semantic segmentation · Image Retrieval

1 Introduction

CBIR is the task of retrieving the images similar to the input query from the dataset based from their contents. State of the art mentions two main contributions used for image similarity: BoVW [12] (Bag of visual words) and CNN descriptors [17]. For retrieval, images must be represented as numeric values. Both contributions represents images as vector 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 an input data(image in our case) using multiple layers for a specified task. Furthermore, deep learning used to solve many computer vision problem such as image and video recognition, image classification, medical image analysis, natural language processing... . Particularly Convolutional Neural Network (CNN) have met with great success for image processing. In deep learning(CNN), the image signature is a vector (feature map) of N floats extracted from the feature layer (Example Fc7 layer for AlexNet[17]). Then computing the distance between the input

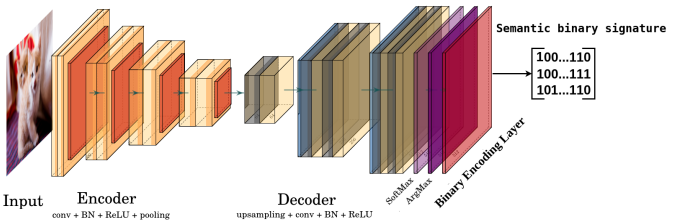

Fig. 1. illustrate a general semantic segmentation architecture with an additional layer (Binary Encoding Layer). This layer transforms the output(2D-map) to a semantic binary signature.

query and dataset using L2 metric or approximate nearest neighbor(ANN) search methods to find the closest images. CNN based features used in existing CBIR works have been trained for classification problems. It is therefore invariant to spatial position of objects. However CBIR applications should take care of spatial position of semantic objects. 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 position of 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 network, originally designed for other application can also be used for CBIR.

Then, by classifying all the pixels of an image, it is then possible to construct abstract representations focusing on objects and their forms. Our approach transforms the semantic 2D-map into binary semantic descriptor. Our descriptor encode the object and forms with their semantic proportion and spatial position in the image. The proposed signature can be localized at the output of the CNN architecture as seen in 1. Our results on six different database highlight the power of our approach.

This article 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 six different datasets and discuss the results of our work in Sect. 4. Section 5 conclusion.

2 State of the art

Many CBIR systems have been proposed in the last years [14] [1] [34] [25] [11] [7]. The content based image retrieval system (figure 2) receives as input a query

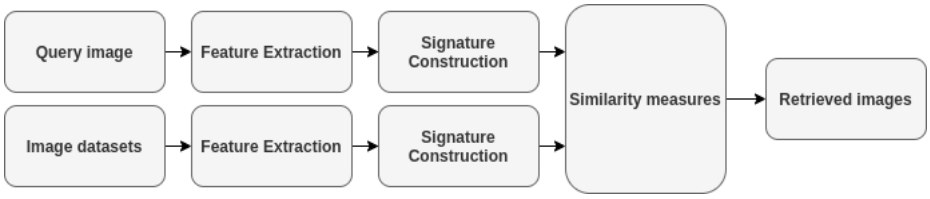

Fig. 2. General CBIR System Architecture

image and returns a list of the most similar images in the database. The framework start with the detection and extraction of the features then the signature construction step. Finally, the closest images to the input query found by the similarity measures between the images signature using dL2 distance. We present a brief overview of approaches based on either visual and learning features.

2.1 Local visual Feature

Bag of Visual Words proposed by [12] is one of the most model used to classify the images by content. This approach is composed of three main steps: (i) Detection and Feature extraction (ii) Codebook construction (iii) Vector quantization. Detection and extraction features in an image can be performed using extractor algorithms. Many descriptors have been proposed to encodes the images into a vector. Scale Invariant Feature Transform (SIFT) [21] and Speeded-up Robust Features (SURF) [6] are the most used descriptors in CBIR. Interesting work from Arandjelović and Zisserman [4] introduces an improvement by upgrading SIFT to RootSift. In other side, binary descriptors has proven useful [27] proposes ORB (Oriented FAST and Rotated BRIEF) to speed up the search. An other work [19] combines two aspects: precision and speed thanks to BRISK (Binary Robust Invariant Scalable Keypoints) descriptor. [15] present a discriminative descriptor for image similarity based on combining contour and color information. Then, creation of clusters from descriptors with K-Means, DBSCAN or another clustering algorithm. The center of each group will be used as the visual word. Finally, creation for each image the histograms of the frequency of vocabularies or visual words, i.e. the image signature. Due the limit of bag of visual words model many improvement have been proposed for more precision. Bag of visual phrases (BoVP) [23] is a high-level description using a more than word for representing an image. formed a phrases using a sequence of n-consecutive words regrouped by L2 metric. [26] Build an initial graph then split the graph into a fixed number of sub-graphs using the N-Cut algorithm. Each histogram of visual words in a sub graph forms a visual phrases. [10] Groups visual words in pairs using the neighbourhood of each point of interest. The pairs words are chosen as visual phrases. Perronnin and Dance [24] applies Fisher Kernels to visual words represented by means of a Gaussian Mixture Model(GMM). Similar approach, introduced a simplification for Fisher kernel. Similar to bag of visual

words, vector of locally aggregated descriptors (VLAD) [16] assign each feature or keypoint to its nearest visual word and accumulate this difference for each visual word.

2.2 Learning-based Feature

Deep learning particularly Convolutional Neural Network have met with great success and in many situation CNN replace local detectors and descriptors. Extracting features using CNN models from images has proven a best result for retrieval. Before, the prediction and extraction features steps the CNN must be trained on large-scale datasets like ImageNet [13]. Neural network training is the process in which the configuration of a neural network determines and calculates the value of each of its weights until the network is able to make correct predictions on images. CNN architecture is composed by a set of layers. The major layers for CNN are :the input layer, hidden layers and the output layer. In CNN the input layer is an image with three dimensional reshaped according to the model. In CNN three applications can be applied on image: Classification, Object detection and Segmentation. Related to CBIR context, in classification category we are interest to extract the vector features from fully connected layer. Many CNN models used for extracting features, including AlexNet [17],VGGNet [29], GoogleNet [32] and ResNet [31]. For example, in AlexNet the size of descriptor from the layer fc7 is 4,096-dim. Similar to Local visual Feature approaches, after extracting all descriptors the retrieval accuracy computed using Euclidean distance between the images. NetVLAD [3] inspired from VLAD is a CNN architecture used for image retrieval. [5] reduce the training time and provides an average improvement in accuracy. Using ACP is frequently in CBIR application thanks to his ability to reduce the descriptor dimension without losing its accuracy. [28] use convolution neural network (CNN) to train the network and support vector machine (SVM) to train the hyerplane then compute the distance between the features image and the trained hyper-plane.

3 Contributions

Encoding is the process of converting the data into a specified format for a specific task. In CBIR, encoding image content have 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 approach based encoding such as BoVW [12], Fisher vector encoding [24], VLAD [16], CNN [17] achieves excellent performance. Consequently, encoding image content is a key element leads to increase the CBIR system performance. Inspired by recent successes of deep learning, we propose a CNN-based model by encoding the output of semantic segmentation architecture for CBIR. So, given a semantic 2D-map, our method (Figure 3) transforms the semantic prediction into a semantic binary signature. The signature construction comprises of two main unsupervised processing units:(i) Encoding of spatial information (ii) Encoding of proportion.

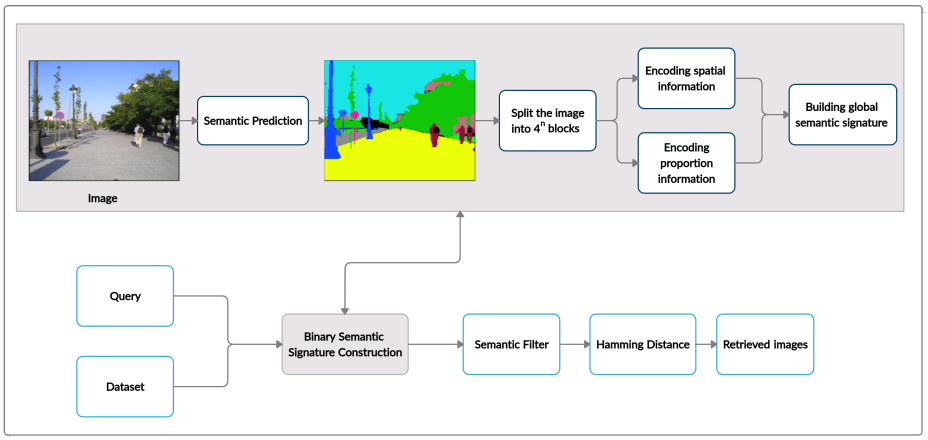

Fig. 3. Global framework

As shown in above of figure 3, given a query image I_q, we obtained the prediction I_{seg} using semantic segmentation algorithm [30] in 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 Encoding of spatial information

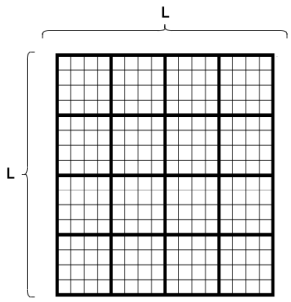

Fig. 4. Illustration of the spatial division. The semantic image divided into 4^n blocks

We propose to encode spatial information using a binary encoding. In a first stage, the image is divided in a recursive way (see fig. 4). 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 L levels, the recursive splitting process generates a set of $n_b = \sum_{n \in \{1,..,L\}} 4^n$ blocks $\mathcal{B} \doteq \{B_{n_b}\}$. 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.

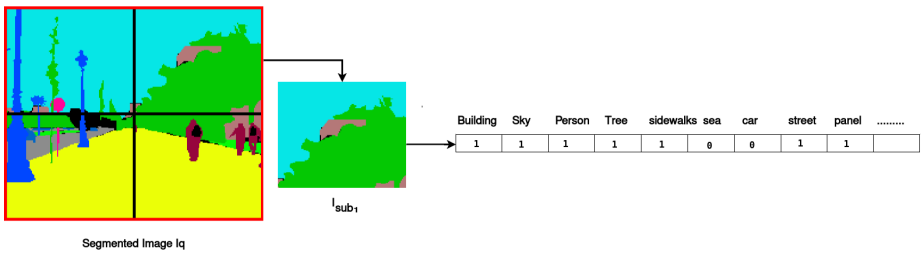

Fig. 5. An example of converting a semantic block to a semantic binary vector.

Figure 5 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.

3.2 Encoding of proportion information

In the second step, we complete the binary spatial presentation with information on the proportion of each semantic class. 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 we calculate its proportion as a percentage P_c in the block. After assigning the percentages of all the classes, a binary conversion process will be applied to each P_c indicated in the equation (1) in order to create a binary signature per block named B_{Pc}.

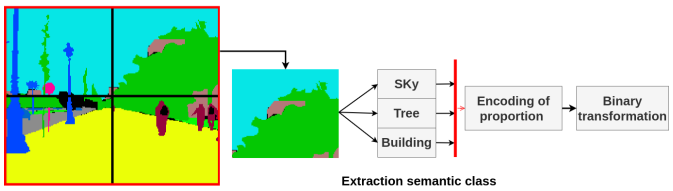

Fig. 6. Example of encoding the proportion information. Given an image divided into 4 blocks, we iteratively select each block to calculate the proportion of the semantic class inside.

$$\begin{cases} \text{if} \quad 0 < P_c <= 0.25 \quad \text{then} \quad B_{Pc} = 0001 \\ \text{if} \quad 0.25 < P_c <= 0.5 \quad \text{then} \quad B_{Pc} = 0011 \\ \text{if} \quad 0.5 < P_c <= 0.75 \quad \text{then} \quad B_{Pc} = 0111 \\ \text{if} \quad P_c > 0.75 \quad \text{then} \quad B_{Pc} = 1111 \end{cases} \tag{1}$$

For cases where the semantic class C_i is not present in the block, an assignment is automatically assigned to it in the block $B_{Pc} = [0000]$. In order to keep all the scores, we collect them together in the B_{Sub-Pc} matrix. This matrix is a binary description of the proportion of a block.

$$B_{Sub-Pci} = \begin{bmatrix} B_{Pc1} \\ B_{Pc2} \\ ..... \\ B_{PcM-1} \\ B_{PcM} \end{bmatrix}$$

where M is the number of classes that the network has learned to detect. Finally, we concatenate all the binary conversions $B_{Sub-Pci}$ to obtain a signature of global proportion S_P corresponding to the segmented input image I_{seg} where $S_P = \{B_{Sub-Pc_1}. B_{Sub-Pc_2} ..... B_{Sub-Pc_M}\}$. We start the tests with large blocks, then we repeat them with smaller to more 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 B_{Sub-Pc}.

Table 1. Database used to evaluate of approach

Name	Size DB / Queries	ground Truth	Query mode
Corel 1K [33] (Wang)	1000 / 1000	100	query-in-ground Truth
Corel 10K [33]	10.000 / 10.000	100	query-in-ground Truth
GHIM-10K [33]	10.000 / 10.000	500	query-in-ground Truth
Linnaeus [9]	6000 / 2000	400	queries/ dataset are disjoint
MSRC v1	241 / 241	-	query-in-ground Truth
MSRC v2	591 / 591	-	query-in-ground Truth

4 Experimental Setup

4.1 Benchmark datasets for retrieval.

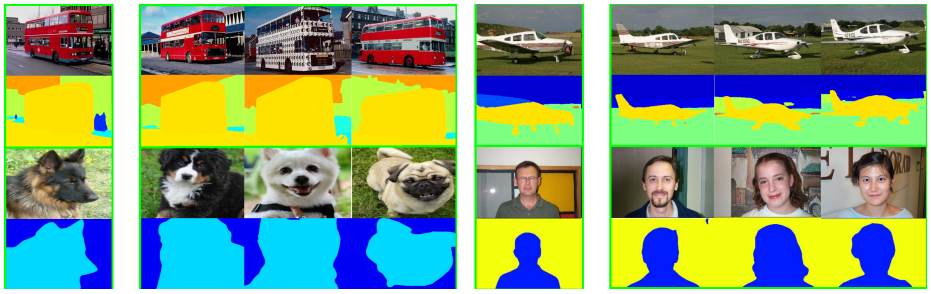

Fig. 7. From different categories selected from different datasets, we show the queries with their corresponding predictions and the three nearest neighbors selected by our method using HRNet-W48 [30] trained on Mseg dataset

Dataset	Images	Merged Classes	All classes	Stuff / Thing classes	Year
Mseg [18]	220K	194	316	102 / 94	2020

Table 2. Details about semantic dataset used to predict the images

Semantic Dataset : Mseg [18]				
Retrieval Dataset \ Size of blocks	$4^0 = 1$	$4^1 = 4$	$4^2 = 16$	$4^3 = 64$
MSRC v1	0.79	0.89	0.83	0.81
MSRC v2	0.64	0.73	0.71	0.67
Linnaeus [9]	0.71	0.77	0.78	0.73
Corel 1K(Wang) [33]	0.77	0.86	0.81	0.80
Corel 10K [20]	0.53	0.55	0.56	0.55
GHIM-10K [20]	0.53	0.52	0.53	0.51

Table 3. MAP evaluations using Coco-stuff datasets

4.2 Results on Benchmark datasets for retrieval

We conducted our experimentation on two different semantic prediction datasets [8] [18] and six retrieval datasets (Table 1). Table 3 presents the mean average precision (MAP) [2] scores for dataset per size of blocks. We start the tests with large blocks to small blocks. When the parameter $n = 1$ then the encoding of semantic spatial information not exist and we encode only the semantic proportion information. Figure 7 clearly indicates that our method capable to select the similar images to input query based on semantic content. The selection is based hamming distance between the query and the images dataset. Experiments with a single thread for each image, the descriptor requires 9 ms on average(Table 4). For [31], [29], [17] we extract from their architectures the features vector from the features layer for evaluating their performance on the datasets using L2 distance.

Retrieval Dataset \ Size of blocks	$4^0 = 1$	$4^1 = 4$	$4^2 = 16$	$4^3 = 64$
MSRC v1	8.8	9.1	12.8	28.1
MSRC v2	8.5	9.8	13.6	30.6
Linnaeus [9]	9.1	11.3	18.6	37.6
Corel 1K(Wang) [33]	10.1	14.3	29.5	41.6
Corel 10K [20]	10.4	14.5	28.9	42.1
GHIM-10K [20]	11.2	15.4	30.1	44.2

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

4.3 Comparison with State-of-the-Art.

Table 5. 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 [12]	0,48	0.30	0,26	0.48	0.30	0.39
n-BoVW [22]	0.58	0.39	0.31	0.60	0.34	0.41
VLAD [16]	0.78	0.41	-	0.74	0.38	0.44
N-Gram [23]	-	-	-	0.37	-	-
AlexNet [17]	0.81	0.58	0,47	0.68	0.40	0.41
VGGNet [29]	0.76	0.63	0,48	0.76	0.45	0.43
ResNet [31]	0.83	0.70	0,69	0.82	**0.59**	0.49
SaCoCo [15]	-	-	-	0.54	0.17	0.15
Ruigang [28]	-	-	0.70	-	-	-
Ayan[7]	-	-	-	0.79	0.52	-
Chu[11]	-	-	-	0.80	0.45	0.51
Ours	**0.89**	**0.73**	**0.78**	**0.86**	0.55	**0.53**

We compare our results against two main categories : (i) Local visual Feature: methods that based on local features like Surf, Sift included the inherited methods such as BoVW, Vlad, Fisher. (ii) Learning based features: methods that based on learning the features using deep learning algorithms. *Hamming* distance is the similarity metric used to compute the similarity between the query and dataset. In Table 5 we compare our results with a large state of the art methods. As indicate the results our proposed present good performance for all datasets. In Table 6 , we compare the precision of the top 20 retrieved image for all categories for Wang dataset. In figure 8 we show the precision performance of top 20 retrieved image for 10 category compared to [14][1] [34][25] methods. The second objective desired in this work is to reduce and minimize the execution time of CBIR system. For any CBIR the execution time depends to time of signature construction. Then, we compare only the time taken by each method to build its signature. We explain here that the extraction, detection and prediction time are not taken into consideration. Table 4 and figure 9 present a comparison time of signature construction for the state of the art methods and our method. Its clearly the interest of our approach in the terms of time against the state of the art methods.

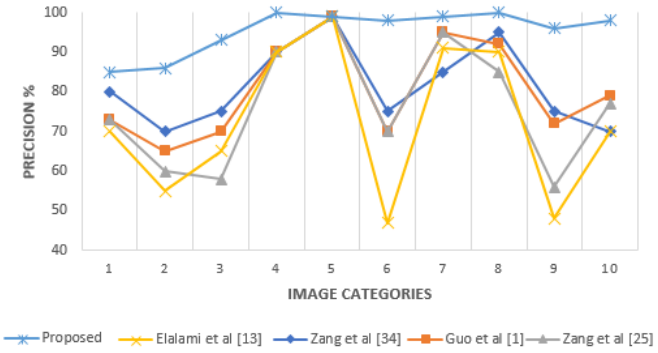

Fig. 8. comparison of precision for top 20 retrieved images for all categories (Corel 1K(Wang) dataset)

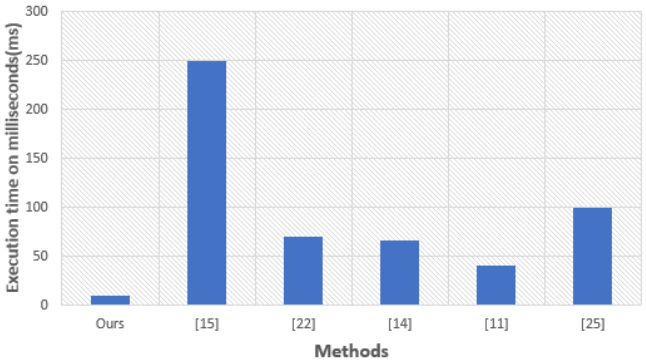

Fig. 9. Comparison of execution time against the state of the art

Table 6. comparison of precision for top 20 retrieved images(Wang dataset)

Methods	Top 20
ElAlami [14]	0.76
Guo and Prasetyo [1]	0.77
Zeng et al. [34]	0.80
Jitesh Pradhan [25]	0.81
Proposed method	0.94

5 Conclusion

We present in this paper a fast and efficient CBIR approach based on semantic segmentation prediction to improve the image similarity. We have shown that by encoding the image information as binary leads to increase the CBIR accuracy. Two gain well shown in our work (i) Time saving (ii) Robust signature based on CNN. Experimental evaluation indicates that our approach achieve a better results in terms of accuracy and time against the state of the art methods.

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. Elli Angelopoulou, Yiannis S Boutalis, Chryssanthi Iakovidou, and Savvas A Chatzichristofis. Mean normalized retrieval order (mnro): a new content-based image retrieval performance measure. 2014.
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. Relja Arandjelovic and Andrew Zisserman. All about vlad. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1578–1585, 2013.
5. 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.
6. Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.
7. Ayan Kumar Bhunia, Avirup Bhattacharyya, Prithaj Banerjee, Partha Pratim Roy, and Subrahmanyam Murala. A novel feature descriptor for image retrieval by combining modified color histogram and diagonally symmetric co-occurrence texture pattern. *Pattern Analysis and Applications*, pages 1–21, 2019.
8. 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.
9. G Chaladze and L Kalatozishvili. Linnaeus 5 dataset for machine learning. Technical report, Tech. Rep, 2017.
10. Tao Chen, Kim-Hui Yap, and Dajiang Zhang. Discriminative soft bag-of-visual phrase for mobile landmark recognition. *IEEE Transactions on Multimedia*, 16(3):612–622, 2014.
11. Kai Chu and Guang-Hai Liu. Image retrieval based on a multi-integration features model. *Mathematical Problems in Engineering*, 2020, 2020.
12. 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.

13. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
14. M Esmel ElAlami. A new matching strategy for content based image retrieval system. *Applied Soft Computing*, 14:407–418, 2014.
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. Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3304–3311. IEEE, 2010.
17. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
18. 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.
19. Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. Brisk: Binary robust invariant scalable keypoints. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2548–2555. IEEE, 2011.
20. 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.
21. Tony Lindeberg. Scale invariant feature transform. 2012.
22. Achref Ouni, Thierry Urruty, and Muriel Visani. A robust cbir framework in between bags of visual words and phrases models for specific image datasets. *Multimedia Tools and Applications*, 77(20):26173–26189, 2018.
23. 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.
24. 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.
25. 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.
26. Yi Ren, Aurélie Bugeau, and Jenny Benois-Pineau. Visual object retrieval by graph features. 2013.
27. Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Computer Vision (ICCV), 2011 IEEE international conference on*, pages 2564–2571. IEEE, 2011.
28. 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.
29. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
30. 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.

31. Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.
32. 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.
33. 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.
34. Shan Zeng, Rui Huang, Haibing Wang, and Zhen Kang. Image retrieval using spatiograms of colors quantized by gaussian mixture models. *Neurocomputing*, 171:673–684, 2016.