A Hybrid Approach for Improved Image Similarity Using Semantic Segmentation

Achref Ouni(✉), Eric Royer, Marc Chevaldonné, and Michel Dhome

Université Clermont Auvergne, CNRS, SIGMA Clermont, Institut Pascal, 63000 Clermont-Ferrand, France
Achref.EL_OUNI@uca.fr

Abstract. Content Based Image Retrieval (CBIR) is the task of finding the images from the datasets that consider 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. To tackle this problem, we present in this paper an efficient CBIR framework based on incorporation between deep learning based semantic segmentation and visual features. We show experimentally that the incorporate leads to the increase of accuracy of our CBIR framework. We study the performance of the proposed approach on four different datasets (Wang, MSRC V1, MSRC V2, Linnaeus).

Keywords: CBIR · Semantic segmentation · Image representation · Features extraction

1 Introduction

Content Based Image Retrieval (CBIR) is a fundamental step in many computer vision applications such as pose estimation, virtual reality, Medical diagnosis, remote sensing, crime detection, video analysis and military surveillance. CBIR is the task of retrieving the images similar to the input query from the dataset based on their contents. CBIR system (see Fig. 1) based on three main steps: (1) Feature extraction (2) Signature construction (3) Retrieval. The performance of any proposed approach depends on the way in an image signature is constructed. Therefore, construction image signature is a key step and the core of CBIR system. State of the art mentions two main contributions used to retrieve the closest image: BoVW (Bag of Visual Words) and CNN (Convolutional Neural Networks) descriptors for image retrieval. Both contributions represent images as vector of valued features. This vector encodes the primitive image such as color, texture and shape.

In this paper, we present a new idea to improve the potential of recovering the relevant images. Our work incorporate the extracted visual features with the

G. Bebis et al. (Eds.): ISVC 2020, LNCS 12510, pp. 647–657, 2020.
https://doi.org/10.1007/978-3-030-64559-5_51

semantic information to build a robust semantic signature. Before computing the distance between the query and the datasets, we have proposed also an efficient test for checking the semantic similarity. This step keeps only the images with the same semantic content with the query and penalize the rest. Our results on 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 four different datasets and discuss the results of our work in Sect. 4. Section 5 conclusion.

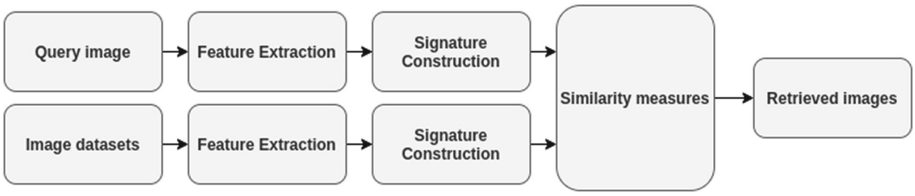

Fig. 1. Cbir system

2 State of the Art

Many CBIR systems have been proposed in last year [1,9,20,28]. In the literature two 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 construction a global signature extracted from the features layer. Let's start by describing the methods based visual features. Bag of visual words (BoVW) or Bag of visual features (BoF) [8] is the popular model used for image classification and image similarity. BoVW treated as following. For each image, the visual features detected then extracted using a visual descriptors such as SIFT [15]. This step will be repeated in a recursive way on all images dataset until collecting all visual descriptors dataset. Then a clustering step using K-MEANS [11] will be applied on the descriptors to build the visual vocabulary (visual words) from the center of each cluster. In order to obtain the visual words, the features query replaced by the index of the visual words that consider the nearest using euclidean distance. Finally, the image described as a histogram of the frequency of the visual phrase exist in the image. Inspired by BoVW, vector of locally aggregated descriptors (VlAD) [10] present an improvement which is assign to each visual feature its nearest visual word and accumulate this difference for each visual word. Fisher Vector encoding [19] uses GMM [21] 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 GMM. Another inspiration from BoVW presented

by Bag of visual phrase (BoVP) [2,17,18]. 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 by different way (sliding windows, k-nearest neighbors, Graph). [2] Local regions are grouped by the method of clustering (single-linkage). [18] Group each key point with its closest spatial neighbors using L2 distance. In other side, deep learning has proven useful in computer vision application. In particular, convolutional neural network (CNN, or ConvNet) is the most commonly applied to analyzing the image by content. CNN algorithms based on architecture for analyzing the images. The architecture is composed by a set of layers. The major layers are: the input layer, hidden layers and the output layer. In CNN for computing the similarity between two images it is necessary to extract the features vector from the feature layer then calculating the distance using L2 metric. Many CNN models have been proposed, including AlexNet [12], VGGNet [23], GoogleNet [26] and ResNet [25]. The fully connected layer (feature layer) usually found towards the end of CNN architectures with vector size of 4096 of float which describe the feature image (color, shape, texture, ...). 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. [4] reduce the training time and provides an average improvement in accuracy. Using ACP is frequently in CBIR application thanks to its ability to reduce the descriptor dimension without losing its accuracy. [22] using convolution neural network (CNN) to train the network and support vector machine (SVM) to train the hyperplane then compute the distance between the features image and the trained hyper-plane.

3 Contributions

In this section, we present a brief explanation of our framework. Our aim is to improve the image representation. The rentability and efficiency of any CBIR system depends on the robustness of the image signature. Figure 2 presents our global framework. Our framework starts with parallel process: extraction visual features and extraction semantic information for both query and datasets. Then, we exploit the extracted information for two main uses: (i) Creation semantic signature (ii) Creation semantic histogram. To build a semantic signature, we incorporate the semantic information with the visual descriptors. Then, we check the resemblance between two images based on their semantic histograms and we compute the distance between the query and the selected candidates using L2 metric.

3.1 Semantic Signature

The most CBIR system describe the image as a vector of N unit. Bag of visual [8] words represent the image as a frequency histogram of vocabulary that are in the image. In deep learning, the image signature is a vector of N float extracted

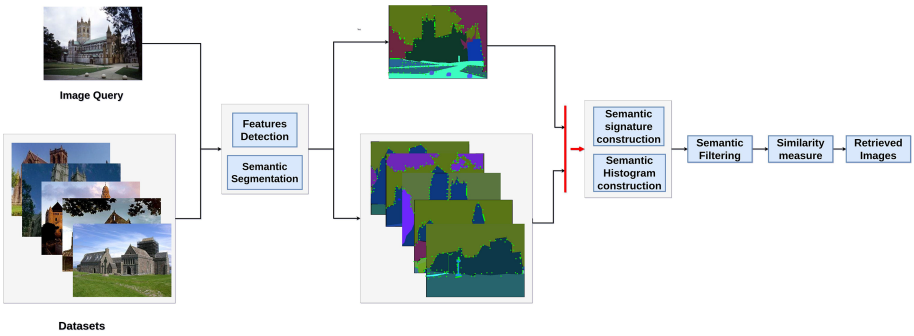

Fig. 2. Global framework

from the feature layer. In this work, we present a new idea to construct an image signature based on incorporation between semantic information and the visual features. We define the signature as a matrix of $N * M$ float where the width N corresponds to the size of descriptor (SIFT 128) and the height M corresponds to the number of classes on which the network was trained. Figure 3 and algorithm 1 describes the different steps of our approach. The process of construction composed of three different steps: (i) Detection and extraction the visual features (ii) Extraction of semantic information (iii) Regrouping the keypoints by class label and computing their center. To compute the center of classes, for each class label on the image we select the set of keypoints that belongs and we apply for them the clustering algorithm (K-MEANS). Consequently, each class label will be presented by a vector of N float. Finally the image signature is composed of N center of clusters that represent the existing classes label in the image. It is not necessarily that the image contains all classes during the prediction. In this case, we assign a null vector for the missing classes.

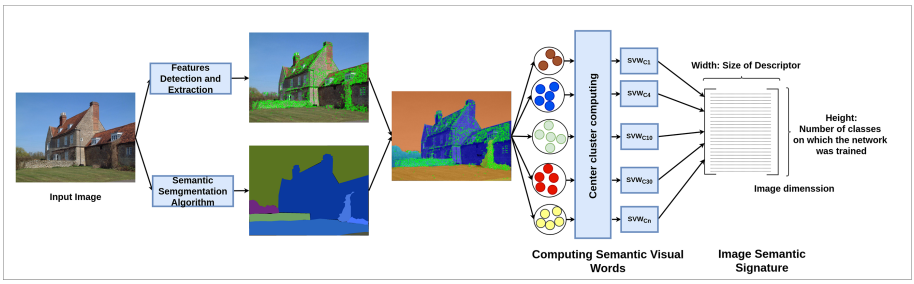

Fig. 3. Semantic signature construction

Algorithm 1 Create Image Signature

Require: Image I, Size N ▷ number of classes on which the network was trained
$Features$=DetectionExtractionFeatures(I)
I_{seg}=SemanticSegmentationPrediction(I)
For i = 1 to N **do**
  IF Exist($Class_i$ in I_{seg})
    SG=∅
    For j = 1 to Size($Features$) **do**
      IF Label($Features_j$)==$Class_i$
        SG= SG . $Features_j$ ▷ Concatenation
      EndIF
    EndFor
    SVW=Kmeans(SG,1) ▷ Semantic visual words
  Else
    SVWs=∅
  EndIF
  Signature(i,:)=SVW
EndFor
Return Signature

3.2 Semantic Histogram

Except that the semantics provides us a class by label, we can also know the objects in the image and their proportion. We exploit this information to check the semantic similarity between the images. Then, we assume that if two images share the same classes label are then semantically similar otherwise the content of the images is different. Consequently, using the semantic information we can select the images which are similar in content with the query. In other side, we can neutral then penalize in the calculation step the dissimilar images with Sim $(I_{query}, I_{dataset}) \leq \epsilon$. To deal with this problem, we proposed to construct for each image a semantic histogram. As shown in Fig. 4, we define the image as a vector of N unit contains the proportion of each class in the image. Then, we measure the semantic similarity between two images using equation (1).

$$Sim(query, candidate) = \sum_{i=0}^{n} |P_{query_i} - P_{dataset_i}| \quad (1)$$

where P are the the proportion of a class in the image.

The main advantage of the checking phase is makes us able to increase the CBIR accuracy by keeping only the images that have the same semantic content with the query and to penalize the rest that consider semantically different (Fig. 5).

4 Experimental Results

4.1 Benchmark Datasets for Retrieval.

In this section, we present the potential of our approach on four 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:

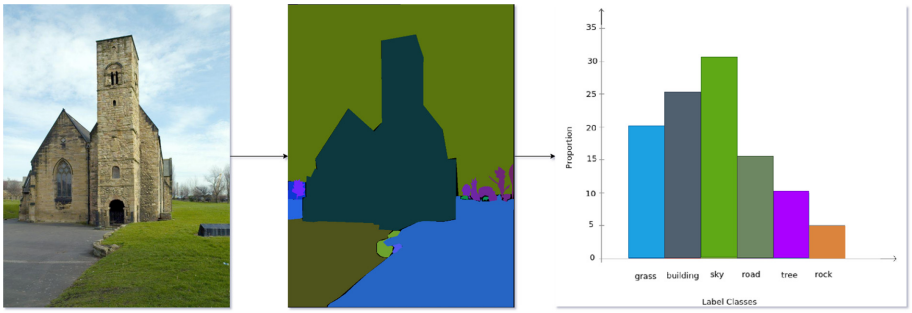

Fig. 4. Semantic histogram

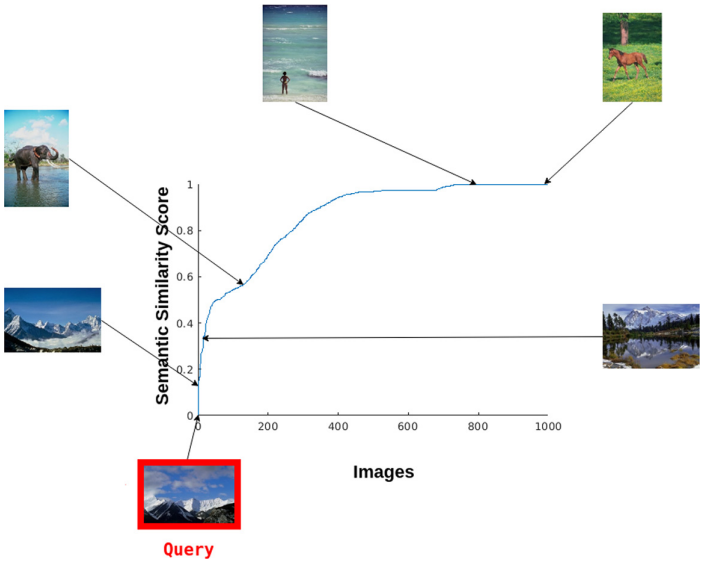

Fig. 5. Semantic similarity between the images

Table 1. Database used to evaluate of approach

Name	Size	Ground	Query mode
	DB/Queries	Truth	
MSRC v1	241/241	–	query-in-ground Truth
MSRC v2	591/591	–	query-in-ground Truth
Wang [27]	1000/1000	100	query-in-ground Truth
Linnaeus [6]	6000/2000	400	queries/dataset are disjoint

- MSRC v1[1] (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[2] (Microsoft Research in Cambridge) contains 591 images included MSRC v1 dataset and divided into 23 categories. The evaluation on MSRC v2 is based on MAP score (mean average precision)
- Corel 1000 [27] or Wang is a dataset of 1000 images divided into 10 categories and each category contains 100 image. The evaluation computed by the average precision of the first 100 nearest neighbors among 1000.
- Linnaeus [6] is a collection of 8000 images of 4 categories (berry, bird, dog, flower). The evaluation on Linnaeus is based on MAP score (mean average precision).

4.2 Performance Metrics

In content based image retrieval (CBIR) the most used evaluate measures is the precision. The precision P is the number of relevant images found compared to the total number of images proposed for a given query.

$$P(I_k) = \sum_{j=1}^{K} \frac{I_j}{K} \tag{2}$$

where k is the number of retrieved images.

In the multi-class case

$$A_v = \frac{1}{M_v} \sum_{k=1}^{M_v} P(I_k) \tag{3}$$

$$mAP = \frac{1}{S} \sum_{k=1}^{S} A_v \tag{4}$$

where M_v is the number of classes and S is the number of queries.

4.3 Benchmark Datasets for Semantic Segmentation

Many semantic segmentation datasets have been proposed in the last years such as Cityscapes [7], Mapillary [16], COCO [14], ADE20K [29], Coco-stuff [5], Mseg [13] and others. The semantic representation is divided into two main categories: Stuff and Things. Things objects have characteristic shapes like vehicle, dog, computer... . Stuff is the description of amorphous objects like sea, sky, tree,... . Therefore, the semantic segmentation datasets are divided into three main categories: (i) Stuff-only (ii) Thing-only (iii) Stuff and Things. To obtain

[1] https://pgram.com/dataset/msrc-v1/.
[2] https://pgram.com/dataset/msrc-v2/.

a robust prediction, we use the recent implementation HRNet-W48 [24] architecture trained on Coco-stuff [5] and Mseg [13] datasets. The main advantage of using Coco-stuff [5] and Mseg [13] datasets is that they are able to predict for both thing and stuff with high number of class predicted for an image.

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

Dataset	Images	Merged classes	All classes	Stuff/Thing classes	Year
Coco-stuff [5]	164K	172	172	92/80	2018
Mseg [13]	220K	194	316	102/94	2020

Table 3. MAP evaluations using Mseg datasets

Retrieval Dataset	Descriptors		
	KAZE	SURF	HOG
MSRC v1	0.79	0.84	0.85
MSRC v2	0.61	0.58	0.60
Linnaeus [6]	0.71	0.73	0.71
Wang [27]	0.73	0.74	0.71
Using semantic filter			
MSRC v1	0.81	0.86	0.87
MSRC v2	0.74	0.73	0.71
Linnaeus [6]	0.73	0.75	0.74
Wang [27]	0.84	0.84	0.83

Table 4. MAP evaluations using Coco-stuff datasets

Retrieval Dataset	Descriptors		
	KAZE	SURF	HOG
MSRC v1	0.77	0.82	0.84
MSRC v2	0.57	0.55	0.61
Linnaeus [6]	0.67	0.68	0.66
Wang [27]	0.71	0.72	0.69
Using semantic filter			
MSRC v1	0.80	0.85	0.86
MSRC v2	0.71	0.72	0.71
Linnaeus [6]	0.72	0.75	0.73
Wang [27]	0.82	0.83	0.80

Table 5. Comparison of precision for top 20 retrieved images (Wang dataset)

Methods	Top 20
ElAlami [9]	0.76
Guo and Prasetyo [1]	0.77
Zeng et al. [28]	0.80
Jitesh Pradhan [20]	0.81
Proposed method	0.91

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

Methods	MSRC v1	MSRC v2	Linnaeus	Wang
BoVW [8]	0,48	0.30	0,26	0.48
n-BoVW [17]	0.58	0.39	0.31	0.60
VLAD [10]	0.78	0.41	–	0.74
N-Gram [18]	–	–	–	0.37
AlexNet [12]	0.81	0.58	0,47	0.68
VGGNet [23]	0.76	0.63	0,48	0.76
ResNet [25]	0.83	0.70	0,69	0.82
Ruigang [22]	–	–	0.70	–
Ours (best)	**0.86**	**0.72**	**0.75**	**0.84**

4.4 Results on Benchmark Datasets for Retrieval

We conducted our experimentation on two different semantic dataset (Table 2) and four retrieval datasets (Table 1). We test our approach using three different descriptors (Kaze, Surf, Kaze). In addition, we compare there with two categories of methods: (i) Local visual Feature: methods that are 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. Tables 3, 4 present the performance of the retrieval on the 4 datasets with three different descriptors. Above, we show the map (mean average precision) scores using only the semantic signature (Fig. 3). Down, we present the results by adding the semantic filter. In experimentation we set epsilon (ϵ) at 0.9 to keep only the images that are considered semantically similar to the input query and we assign a negative score to the rest. It clearly indicates that adding the semantic filter improves the accuracy.

For the methods [1,9,20,28] in Table 5, we compare the precision of the top 20 retrieved images for all categories for Wang dataset. In Table 6 we compare our results with a large state of the art methods. For [12,23,25] we extract from their architecture the features vector from the features layer then we evaluate

the their performance on the datasets using L2 distance. As indicate the results our proposed present good performance for all datasets.

5 Conclusion

In this paper, we have presented an efficient CBIR approach based on incorporation between deep learning based semantic segmentation and visual features. We have presented two main uses of the semantic information: (i) Creation semantic signature (ii) Creation semantic histogram. We have proven that the use of the semantic information increase the CBIR accuracy. With different descriptors (KAZE, SURF, HOG) our approach achieve a better results in terms of accuracy compared to the state of the art methods.

References

1. Admile, N.S., Dhawan, R.R.: Content based image retrieval using feature extracted from dot diffusion block truncation coding. In: 2016 International Conference on Communication and Electronics Systems (ICCES), pp. 1–6. IEEE (2016)
2. Albatal, R., Mulhem, P., Chiaramella, Y.: Visual phrases for automatic images annotation. In: 2010 International Workshop on Content Based Multimedia Indexing (CBMI), pp. 1–6. IEEE (2010)
3. Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
4. Balaiah, T., Jeyadoss, T.J.T., Thirumurugan, S.S., Ravi, R.C.: A deep learning framework for automated transfer learning of neural networks. In: 2019 11th International Conference on Advanced Computing (ICoAC), pp. 428–432. IEEE (2019)
5. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: thing and stuff classes in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1209–1218 (2018)
6. Chaladze, G., Kalatozishvili, L.: Linnaeus 5 dataset for machine learning. Technical report (2017)
7. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
8. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision. ECCV, Prague, vol. 1, pp. 1–2 (2004)
9. ElAlami, M.E.: A new matching strategy for content based image retrieval system. Appl. Soft Comput. **14**, 407–418 (2014)
10. Jégou, H.,Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3304–3311. IEEE (2010)
11. Krishna, K., Narasimha Murty, M.: Genetic k-means algorithm. IEEE Trans. Syst. Man Cybern. Part B (Cybern.) **29**(3), 433–439 (1999)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)

13. Lambert, J., Zhuang, L., Sener, O., Hays, J., Koltun, V.: MSeg: a composite dataset for multi-domain semantic segmentation. In: Computer Vision and Pattern Recognition (CVPR) (2020)
14. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
15. Lindeberg, T.: Scale invariant feature transform (2012)
16. Neuhold, G., Ollmann, T., Bulo, S.R., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4990–4999 (2017)
17. Ouni, A., Urruty, T., Visani, M.: A robust CBIR framework in between bags of visual words and phrases models for specific image datasets. Multimed. Tools Appl. **77**(20), 26173–26189 (2018). https://doi.org/10.1007/s11042-018-5841-8
18. Pedrosa, G.V., Traina, A.J.M.: From bag-of-visual-words to bag-of-visual-phrases using n-grams. In: 2013 XXVI Conference on Graphics, Patterns and Images, pp. 304–311. IEEE (2013)
19. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
20. Pradhan, J., Kumar, S., Pal, A.K., Banka, H.: Texture and color visual features based CBIR using 2D DT-CWT and histograms. In: Ghosh, D., Giri, D., Mohapatra, R.N., Savas, E., Sakurai, K., Singh, L.P. (eds.) ICMC 2018. CCIS, vol. 834, pp. 84–96. Springer, Singapore (2018). https://doi.org/10.1007/978-981-13-0023-3_9
21. Rasmussen, C.E.: The infinite gaussian mixture model. In: Advances in Neural Information Processing Systems, pp. 554–560 (2000)
22. Fu, R., Li, B., Gao, Y., Wang, P.: Content-based image retrieval based on CNN and SVM. In: 2016 2nd IEEE International Conference on Computer and Communications (ICCC), pp. 638–642 (2016)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
24. Sun, K., et al.: High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514 (2019)
25. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
26. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
27. Wang, J.Z., Li, J., Wiederhold, G.: Simplicity: semantics-sensitive integrated matching for picture libraries. IEEE Trans. Pattern Anal. Mach. Intell. **23**(9), 947–963 (2001)
28. Zeng, S., Huang, R., Wang, H., Kang, Z.: Image retrieval using spatiograms of colors quantized by Gaussian mixture models. Neurocomputing **171**, 673–684 (2016)
29. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 633–641 (2017)