Bag of Features vs Vector of Locally Aggregated Descriptors

Farkhunda Younas[1], Junaid Baber[2(✉)], Tahir Mahmood[3], Javeria Farooq[4], and Maheen Bakhtyar[2]

[1] Department of Computer Science, Sardar Bahadur Khan Women's University, Quetta, Pakistan
Farkhundayounas@yahoo.com
[2] Department of Computer Science and Information Technology, University of Balochistan, Quetta, Pakistan
junaidbaber@ieee.org, maheenbakhtyar@um.uob.edu.pk
[3] Department of Computer Science, COMSATS Institute of Infomation Technology, Islamabad, Pakistan
tahirmahmood.cs@gmail.com
[4] Department of Electronic Engineering, Balochistan University of Information Technology, Engineering and Management Sciences, Quetta, Pakistan
javeria.farooq74@gmail.com

Abstract. Image representation by set of local features are common and also state-of-the art for many applications such as image retrieval and image classification. A single image contains on average 2.5 k–3.0 k features. Searching the images based on local features are discriminative compared to global features at the cost of heavy computational overhead. Bag-of-Features (BoF), also known as bag-of-visual words, are used for feature quantization which makes searching local features feasible in very large databases at the cost of distinctiveness. Mostly, the vocabulary size in those applications is kept up-to 1 million. In this research study, we investigated the performance of Vector of Locally Aggregated Descriptors (VLAD) which is recently proposed as an alternative to BoF for different families of descriptor. The VLAD achieves similar or sometimes better performance when compared to BoF despite of limited vocabulary size. The performance of VLAD is mostly compared with BoF on gradient based descriptors in literature. In our experiments, we take gradient based descriptor, intensity based descriptor, and binary descriptor. Scale Invariant Feature Transform (SIFT), Local Intensity Order Pattern (LIOP) and BInarization of Gradient Orientation Histograms (BIGOH) are used to validate the performance of VLAD in parallel to BoF on famous benchmark dataset. VLAD outperforms BoF in gradient based family and intensity based family but non of these are feasible for binary descriptors.

Keywords: Bag-of-Features (BoF) · Local features · Locally aggregated descriptors (VLAD) · SIFT

Y. Bi et al. (eds.), *Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016*, Lecture Notes in Networks and Systems 16, DOI 10.1007/978-3-319-56991-8_10

1 Introduction

Large-scale image retrieval and classification are gotten an important interest in research and industry applications [1,2]. The main challenge of image retrieval is to retrieve images effectively and efficiently from very large datasets. Images are represented either by global features or set of local features. The features computed from the whole image are treated as global feature, such as color histogram. Whereas, multiple features computed from the small patches in the images are called local features such as Scale Invariant Feature Transform (SIFT) [3] and BInarization of Gradient Orientation Histograms (BIG-OH) [4].

Global features suffer from robustness and distinctiveness as small change in image drastically changes the features. There are many transformations where global features performance are not even to the baseline, such as cropping, sub-image insertion, and image fusion. The given two images I and J can be represented by global features $G_I, G_J \in \mathbb{R}^d$. The distance or similarity between these two images can be computed and that score can be used for rank list. In case of Euclidean distance the similarity score between these two images can be computed as follow

$$E(I,J) = \sqrt{\sum (G_I - G_J)^2} = \sqrt{\sum_i^d (G_I^{(i)} - G_J^{(i)})^2} \tag{1}$$

For any dataset that contains N images, there will be N computation for retrieving the rank list based on Eq. 1 which is linear in time and efficiency.

On the other hand, local features, also known as local keypoint descriptors, are the complimentary approach to global features. Initially, keypoint detector is used to find some key locations in the image such as Difference of Gaussian (DoG) detector [3,5], later these keypoints are represented by some robust and distinct descriptors such as SIFT [3]. As mentioned above, there are on average 2.5 k keypoints on single image using DoG detector which enables images to be matched/searched partially in the large corpse. The ability to match/search partially enables local features tackle the challenging transformations effectively, such as cropping, sub-image insertion and image fusion. Local features are widely used for many applications [6–8]. The main limitation of local features is their distance or similarity computation cost. For given images, I and J, SIFT keypoint descriptors are computed and denoted by Q and R respectively. The point pair (q_i, r_j) is assumed to be matched if the following conditions hold

- The Euclidean distance E

$$E(q_i, r_j) = \min_{r_k \in R} E(q_i, r_k) \tag{2}$$

- The following inequality:

$$(q_i, r_j) < \min_{r_l \in R, l \neq j} E(q_i, r_l) \times \beta \tag{3}$$

where $0 < \beta < 1$, the smaller β, the fewer the matched points [9]. We set β to 0.6. Finally, the matching score between two images is computed as follow:

$$\text{Matching score} = \frac{M}{|Q|} \times 100 \tag{4}$$

where M is the number of matched features and $|Q|$ is number of features in I. Rank list is obtained based on descending matching scores.

Image retrieval based on local features is distinctive but not scalable. Therefore, it is important to quantize the feature space. There are various techniques proposed for features quantization such as BoF [10], VLAD [11], and features BInarization [4].

In this paper, we have investigated two famous techniques for quantization on benchmark datasets. The first technique is BoF [10] (explained in Sect. 3.1). This technique is widely used for many applications [1,7,8,10,12]. The second technique is VLAD [11] which is much more compact representation of local features compared to BoF. Despite of compactness, it gives better retrieval compared to BoF.

The remaining paper is organized as follows. Section 2 presents literature of BoF and VLAD. Section 3 describes the quantization techniques. Section 4 describes the experimental setup, dataset, evaluation criteria, features extraction process. Section 5 presents experimental results. Finally, the concluding remarks with future work are given in Sect. 6.

2 Related Work

Patch based descriptors play vital role in large-scale image/object retrieval and image classification. SIFT is one of the gold standard keypoint descriptor which is invariant to many challenging transformations such as rotation, scale change, illumination and view-point change. The detail explanation of SIFT detector and descriptor can be found in Lowe, D. G work [3,5]. Local features have many wide range of applications in computing vision field such as object recognition, image stitching, image retrieval, wide baseline matching, and image tracking.

The LIOP [13] descriptor encodes the local ordinal information of each pixel within the patch and local patch is divided into sub regions based on ordinal information. LIOP does not only perform better in image rotation and monotonic intensity changes but also work for other geometric transformation like image blur, JEPG compression, and image view point. LIOP captures both local patch and intensity ordinal information of selected normalize region which make it more distinctive and robustness.

Baber et al. [4] proposed binary quantization, known as BIGOH, of SIFT descriptor which reduces the memory storage and increases the efficiency of distance computation without affecting the overall performance.

Yu su et al. [14] used the BOF model for the image classification to addressed two problems, one is lack of semantic understanding of visual words, and

other one is polygamous. They proposed two different approaches which contain the semantic attributes which is predicted on whole image and build the intermediate representation using these prediction. They used four challenging datasets, Scene-15[1], MSRCv2[2], PASCAL VOC 2007 [15] and SUN-397 [16] for the experiments.

Baber et al. [17] proposed an efficient framework for video segmentation using BoF. He experimentally showed that better performance can be achieved in video segmentation despite keeping the vocabulary size very small. In video segmentation, every shot is compared with its adjacent shots for scene formation, and in every scene the shots are similar due to the theme of the video. Whereas, in video retrieval the size of vocabulary is kept very high (1 M).

Adnan Hota [18] has used the simulation of image classification to compare the two kernels of Support Vector Machines (SVM) model to analyze the speed, consumption measure of processor power, and accuracy. In the simulation they used VLFEAT software package and PASCAL VOC 2007 benchmark dataset for the experiment. Results show that the non linear Hellinger kernel has better performance than the linear SVM.

Xiaojiang Peng [19] used the VLAD on the two problems one is higher order statistics ignored in VLAD and other problem classification task dictionary is not optimal. They proposed high order VLAD (H-VLAD) to overcome these problems. They used different data set UCF101, PASCAL VOC 2007 and HMDB51 for object classification and video-based action recognition.

3 Features Quantization

In this section, we explain the two famous techniques for features quantization.

3.1 Bag-of-Feature (BoF)

SIFT is 128-D descriptor for each keypoint. There are various implementations of SIFT available. We used VLFEAT[3] API for SIFT and other descriptors. Every cell of SIFT is 1-byte, so there are 256^{128} unique descriptors, which are very big space. To quantize the feature space we train the vocabulary $\mathcal{V}$ of length $\mathcal{K}$ where $\mathcal{V} = \{v_1, v_2, \ldots, v_{\mathcal{K}}\}$. The recommended value of $\mathcal{K} = 1M$ [1,7,8,10,12]. Hierarchical $\mathcal{K}$-mean clustering is widely used for these vocabulary learning [20]. A quantizer $\mathcal{Q}$ is proposed as:

$$\begin{aligned} \mathcal{Q} : \mathbb{R}^d &\rightarrow [1, \mathcal{K}] \\ x &\rightarrow \mathcal{Q}(x) \end{aligned} \tag{5}$$

Any given descriptor $x, x \in \mathbb{R}^d$ is mapped to an integer index between 1 and $\mathcal{K}$ based on the minimum distance x with all the vocabularies $\mathcal{V}$. Finally, histogram

[1] http://qixianbiao.github.io/Scene.html.
[2] http://research.microsoft.com/en-us/projects/objectclassrecognition/.
[3] http://www.vlfeat.org/.

of visual words is computed for the given image. The length of histogram is $\mathcal{K}$ which is very sparse. Now again the rank list based on BoF histogram can be computed in linear time.

3.2 Vector of Locally Aggregated Descriptors (VLAD)

Jegou et al. [11] proposed an efficient, memory smart, and effective framework for large scale image retrieval. BOF framework are used with large vocabularies (hierarchical) improving the descriptors representation but only few million images are tractable by the memory. To overcome this problem, geometric min-hesh are used with better performance. They proposed efficient approach, VLAD, which is obtained by aggregating local descriptors which is very similar to fisher vector (FV) [21]. Vector of Locally Aggregated Descriptors (VLAD) is the improvement and simplification of BoF. Visual vocabulary $\mathcal{V}'$ is learned similarly as BoF, where $\mathcal{V}' = \{v'_1, v'_2, \ldots, v'_{\mathcal{K}'}\}$. The value of $\mathcal{K}' << \mathcal{K}$. Mostly the values of $\mathcal{K}' \in \{16, 32, 64\}$. Once the nearest neighbor visual word $v'_i \in \mathcal{V}'$ for given descriptor x is computed, the accumulated differences all the residuals in cluster v'_i with x is computed as feature vector for descriptor x [11]. The feature extracted using VLAD is of $d \times \mathcal{K}'$.

4 Experimental Setup

4.1 Datasets

Benchmark dataset, VOC 2007, is used for the experiments which are challenging and also widely accepted. The PASCAL VOC 2007 dataset contains 10 K images of 20 object classes. The dataset contains two set of images, training and testing. Average precision is taken as the evaluation criteria for experiments.

4.2 Features Extraction

All the descriptors in literature can be categorized into three broad classes: gradient based descriptors, intensity based descriptors, and binary descriptors. In our experiments we used one descriptor from each category to conform the evaluation. SIFT, LIOP and BIGOH descriptors are used for the feature extraction from the images. Standard keypoint descriptor computation pipeline designed by Mikolajczyk and Schmid is followed [22]. Feature extraction can be divided into three main steps, (1) keypoint detection, (2) keypoint patch normalization, and (3) descriptor computation.

There are number of keypoint detectors available such as SIFT which is also known as Difference of Gaussian (DoG), Hessian affine, and Harris affine detector. The DoG filter images at multiple scales and approximate Laplacian-of-Gausian (LoG) filters. A Gaussian pyramid is constructed by progressively blurring and sub-sampling the image to get the DoG keypoints. Local extrema

are identified and considered as candidate keypoints at each level of pyramid which is obtained by the differences of blurred images [5]. The repeatability of DoG points are weak compared to Harris and Hessian affine keypoints [23–25]. Therefore, Harris affine keypoints are used for the experiments.

For keypoint patch normalization, standard procedure used by many researchers is followed [22,25]. Patch around the keypoint is normalized to 41×41 pixels. Finally, respective descriptor is computed from the normalized patch, as shown in Fig. 1.

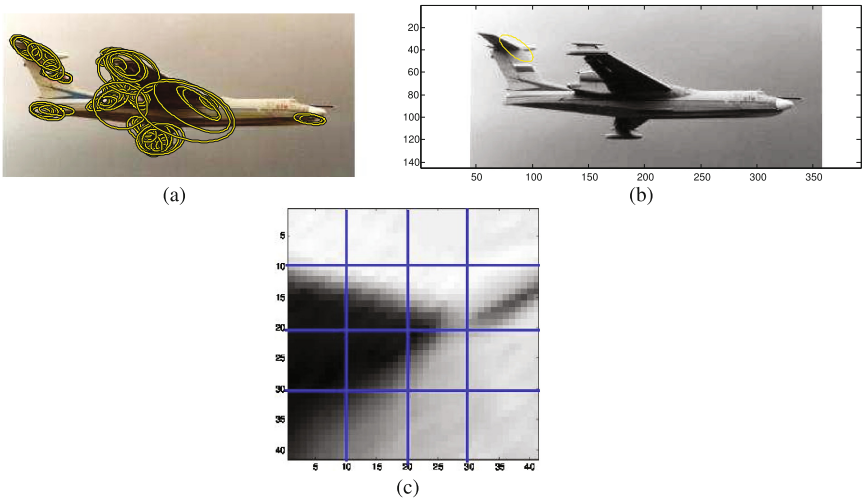

Fig. 1. Example of descriptor computation, (a) show the image with Harris affine keypoints, (b) shows one random elliptical keypoint, and (c) shows the normalized patch of 41×41 pixels divided into 4×4 spatial grid.

4.3 Features Quantization

We quantized these features using BoF and VLAD. To train the VLAD and BoF, the values are $\mathcal{K} = 200000, \mathcal{K}' = 64$. As SIFT is 128D, and there are thousands of keypoints per image. The feature space of SIFT is 256^{128}, in case every cell value in SIFT is 8-bits, which is too large. The quantized images are represented by histogram of visual words of length 200000. So every image has now 200000 D feature when quantized using BoF. Whereas, in case of VLAD it is only $128 \times 64 = 8192$ which is too compact compared to BoF. So every image quantized by VLAD is represented by feature vector of dimensions 8192. BOF and VLAD use the flat k-means clustering for all features, i.e., SIFT, LIOP and BIGOH. To obtain the visual words, k-mean or hierarchical k-mean clustering are widely used. VLFEAT API[4] is used for feature extraction and visual word learning along with executable provided by VGG[5].

[4] http://vlfeat.org/.
[5] http://www.robots.ox.ac.uk/vgg/research/affine/.

In experiment SVM is used as classifier. SVM is tranined on the tranining data provided by VOC [21,26]. For every feature, and both quantizer different SVM is tranined. Every image is represented by single vector of BoF and VLAD after the features are quantized. These quantized features are used for classification. In case of BoF the feature length for every image is 200000, and 8192 D for the VLAD. The feature vector of BoF is very sparse where as VLAD values are dense.

5 Performance Evaluation

There are several ways to compute the performance. It depends on the context we are interested. Since, VLAD and BoF both are quantization techniques. Therefore, we have focused on patch based descriptors size/image, and retrieval

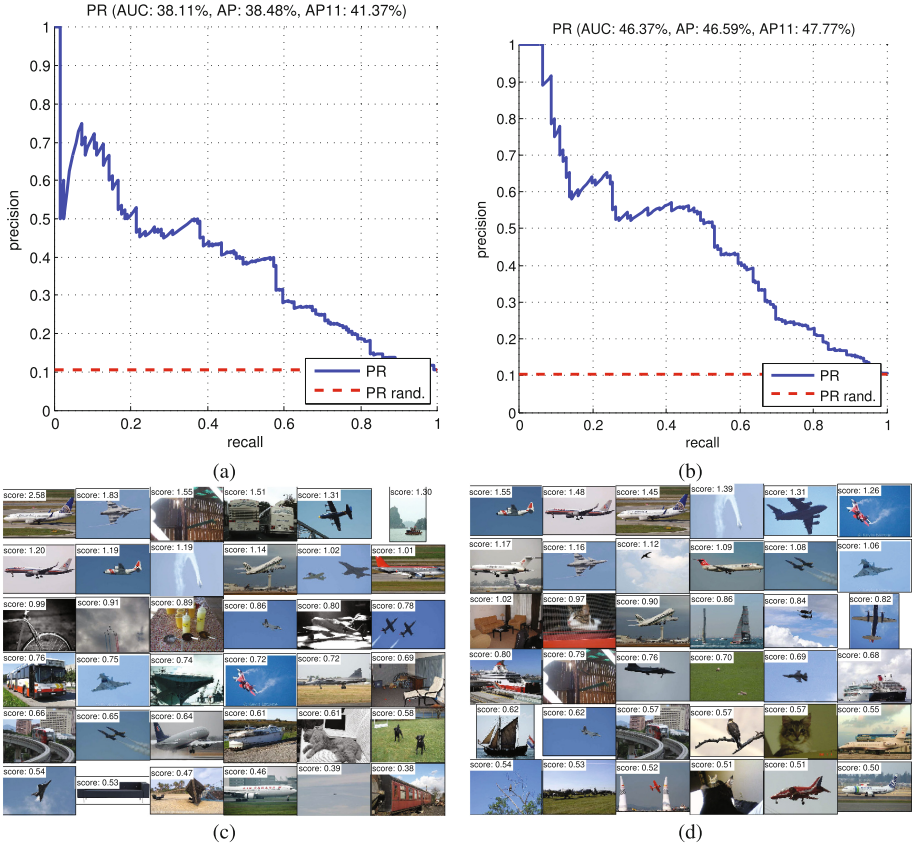

Fig. 2. Precision and recall curves on Aeroplane class using LIOP descriptor. (a) shows for BoF performance (b) shows VLAD performance, (c) shows the visual rank list obtained by BoF and (d) shows VLAD visual rank list.

accuracy. VLFeat is used for experiments. It is portable and open source library for computer vision algorithms. It helps to make easy quick prototyping for computer vision researches. It contains many implementations of many algorithms, such as feature detectors and extractors, k-means clustering, and randomized kd-tree matching. We used VLFEAT library for experiments.

Figure 2 shows the comparative analysis of precision recall curves of Aeroplane class. There are 126 positive images and 2096 negative images. Figure 2 (a) shows the BoF with LIOP descriptor with AP 38.48%, and (*b*) shows the VLAD with LIOP descriptor with AP 46.59%. To obtain the ranklist for evaluation, Euclidean distance is used between pair of images.

Results of AP and mAP are listed in Table 1 for all the classes in VOC 2007 dataset, mean average precision (mAP) is the mean of all average precisions across set of queries. The qualitative view of retrieval is shown in Fig. 2(c) and (d). We have only show the top 36 images in retrieval view.

Table 1. Average precision BOF and VLAD with SIFT, LIOP and BIGOH descriptor

		SIFT descriptor		LIOP descriptor		BIGOH descriptor	
		BoF	VLAD	BoF	VLAD	BoF	VLAD
1	Airplane	39.72%	52.08%	38.48%	46.59%	16.95%	19.95%
2	Bicycle	17.00%	19.50%	18.46%	19.87%	12.27%	11.87%
3	Bird	12.73%	11.21%	15.69%	13.36%	10.92%	12.27%
4	Boat	12.45%	13.04%	16.37%	14.83%	9.25%	8.56%
5	Bottle	16.00%	21.00%	20.90%	17.46%	8.98%	9.97%
6	Bus	15.16%	19.2%	14.85 %	16.82 %	8.42%	8.29%
7	Car	41.05%	51.02%	41.85%	44.91%	28.70%	31.52%
8	Cat	15.70%	13.29%	15.04%	14.11%	12.13%	12.44%
9	Chair	22.30%	21.64%	24.09%	24.88%	16.62%	18.49%
10	Cow	9.21 %	9.39%	10.22%	9.59%	7.60%	7.39%
11	Dining table	12.71%	13.15%	13.43%	13.66%	8.52%	10.40%
12	Dog	18.14 %	16.03%	17.86	16.96%	17.21%	17.72%
13	Horse	23.26%	37.22%	26.98%	38.93%	20.29%	19.72%
14	Motorbike	25.44%	34.37%	19.69%	24.71%	11.43%	11.48%
15	Person	58.98 %	62.48 %	60.37%	63.52 %	50.26%	51.47%
16	Potted Plant	13.77%	12.76%	10.73%	12.45%	9.91%	10.20%
17	Sheep	3.77%	5.38%	7.29%	7.58%	3.98%	3.64%
18	Sofa	13.73 %	13.80%	14.92%	14.74%	11.39%	12.44%
19	Train	16.51%	14.55%	16.46%	15.35%	11.18%	13.83%
20	Tv/monitor	15.34%	15.21%	16.23%	14.95%	11.18%	13.47%
	mAP	20.10%	**22.82%**	21.00%	**22.26%**	14.38%	**15.25%**

VLAD works good both on gradient based descriptors and intensity based descriptors. However, both quantization techniques have limited performance on binary descriptors. Since, BIGOH is obtained by quantizing the SIFT, that is why it gives poor performance when further quantized by visual words. There is still room for binary descriptors as in litereature, best of our knowledge, there is no efficient and affective quantization for binary descriptors using BoF and VLAD. However, significant works are reported in Hash families.

6 Conclusion

VLAD is an effective technique for feature quantization. Despite of compactness, VLAD gives equal or better performance as compared to BoF. During experiments, BoF is 200000 D for given image whereas VLAD is only 8192 D, in case of SIFT. It can be seen that despite of reduced length VLAD outperforms BoF which is considered as one of the golden technique for feature quantization. It can also be seen that both these quantization techniques have limited performance for binary descriptors. Since, binary descriptors have limited feature space and when further quantized to reduce to space, the performance is significantly compromised.

In our future work we want to investigate the performance of VLAD for other machine vision applications such as content based video retrieval, and tracking face in live streaming videos. We are also interested to propose scalable quantization for binary descriptors as both of these quantization, BoF and VLAD, are not appropriate for binary descriptors.

Acknowledgment. This research work is supported by Higher Education Commission (HEC) of Pakistan, SBK women university, and university of Balochistan.

References

1. Yu, F.X., Ji, R., Tsai, M.-H., Ye, G., Chang, S.-F.: Weak attributes for large-scale image retrieval. In: International Conference on Computer Vision and Pattern Recognition, pp. 2949–2956 (2012)
2. Gong, Y., Lazebnik, S., Gordo, A., Perronnin, F.: Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2916–2929 (2013)
3. Lowe, D.G.: Object recognition from local scale-invariant features. In: International Conference on Computer Vision, vol. 2, pp. 1150–1157 (1999)
4. Baber, J., Dailey, M.N., Satoh, S., Afzulpurkar, N., Bakhtyar, M.: BIG-OH: binarization of gradient orientation histograms. Image Vis. Comput. **32**(11), 940–953 (2014)
5. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
6. Wu, Z., Ke, Q., Isard, M., Sun, J.: Bundling features for large scale partial-duplicate web image search. In: Computer Vision and Pattern Recognition, pp. 25–32 (2009)

7. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Computer Vision and Pattern Recognition, pp. 1–8 (2007)
8. Jégou, H., Douze, M., Schmid, C.: Packing Bag-of-Features. In: International Conference on Computer Vision, pp. 2357–2364 (2009)
9. Baber, J., Afzulpurkar, N., Satoh, S.: A framework for video segmentation using global and local features. Int. J. Pattern Recogn. Artif. Intell. **27**(05) (2013)
10. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: International Conference on Computer Vision, pp. 1470–1477 (2003)
11. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: International Conference on Computer Vision and Pattern Recognition, pp. 3304–3311 (2010)
12. Yuan, X., Yu, J., Qin, Z., Wan, T.: A SIFT-LBP image retrieval model based on bag of features. In: IEEE International Conference on Image Processing (2011)
13. Wang, Z., Fan, B., Wu, F.: Local intensity order pattern for feature description. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 603–610. IEEE (2011)
14. Yu, S., Jurie, F.: Improving image classification using semantic attributes. Int. J. Comput. Vis. **100**(1), 59–77 (2012)
15. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
16. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A., et al.: Sun database: large-scale scene recognition from abbey to zoo. In: International Conference on Computer Vision and Pattern Recognition, pp. 3485–3492 (2010)
17. Baber, J., Satoh, S., Afzulpurkar, N., Keatmanee, C.: Bag of visual words model for videos segmentation into scenes. In: Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service, pp. 191–194 (2013)
18. Hota, A.: Comparison of some bag-of-words models for image recognition. In: 2014 X International Symposium on Telecommunications (BIHTEL), pp. 1–5 (2014)
19. Peng, X., Wang, L., Qiao, Y., Peng, Q.: Boosting VLAD with supervised dictionary learning and high-order statistics. In: European Conference on Computer, pp. 660–674 (2014)
20. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: Computer Vision and Pattern Recognition, vol. 2, pp. 2161–2168 (2006)
21. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: European Conference on Computer, pp. 143–156 (2010)
22. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1615–1630 (2005)
23. Adam, B.: Reliable feature matching across widely separated views. In: International Conference on Computer Vision and Pattern Recognition, pp. 774–781 (2000)
24. Lindeberg, T., Gårding, J.: Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure. Image Vis. Comput. **15**, 415–434 (1997)
25. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. Int. J. Comput. Vis. **60**, 63–86 (2004)
26. Malisiewicz, T., Gupta, A., Efros, A., et al.: Ensemble of exemplar-SVMs for object detection and beyond. In: International Conference on Computer Vision, pp. 89–96 (2011)