

Accepted Manuscript

Object-oriented convolutional features for fine-grained image retrieval in large surveillance datasets

Jamil Ahmad, Khan Muhammad, Sambit Bakshi, Sung Wook Baik

PII: S0167-739X(17)31857-5
DOI: <https://doi.org/10.1016/j.future.2017.11.002>
Reference: FUTURE 3794

To appear in: *Future Generation Computer Systems*

Received date: 16 August 2017

Revised date: 18 October 2017

Accepted date: 1 November 2017

Please cite this article as: J. Ahmad, K. Muhammad, S. Bakshi, S.W. Baik, Object-oriented convolutional features for fine-grained image retrieval in large surveillance datasets, *Future Generation Computer Systems* (2017), <https://doi.org/10.1016/j.future.2017.11.002>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



Object-Oriented Convolutional Features for Fine-Grained Image Retrieval in Large Surveillance Datasets

¹Jamil Ahmad, ¹Khan Muhammad, ²Sambit Bakshi, ^{1,*}Sung Wook Baik

¹Digital Contents Research Institute, Sejong University, Seoul, Republic of Korea

²Department of Computer Science & Engineering, National Institute of Technology, Rourkela, India

*Corresponding author: sbaik3797p@gmail.com

Abstract

Large scale visual surveillance generates huge volumes of data at a rapid pace, giving rise to massive image repositories. Efficient and reliable access to relevant data in these ever growing databases is a highly challenging task due to the complex nature of surveillance objects. Furthermore, inter-class visual similarity between vehicles requires extraction of fine-grained and highly discriminative features. In recent years, features from deep convolutional neural networks (CNN) have exhibited state-of-the-art performance in image retrieval. However, these features have been used without regard to their sensitivity to objects of a particular class. In this paper, we propose an object-oriented feature selection mechanism for deep convolutional features from a pre-trained CNN. Convolutional feature maps from a deep layer are selected based on the analysis of their responses to surveillance objects. The selected features serve to represent semantic features of surveillance objects and their parts with minimal influence of the background, effectively eliminating the need for background removal procedure prior to features extraction. Layer-wise mean activations from the selected features maps form the discriminative descriptor for each object. These object-oriented convolutional features (OOCF) are then projected onto low-dimensional hamming space using locality sensitive hashing approaches. The resulting compact binary hash codes allow efficient retrieval within large scale datasets. Results on five challenging datasets reveal that OOCF achieves better precision and recall than the full feature set for objects with varying backgrounds.

Keywords: image retrieval, object-oriented features, convolutional neural network, fine-grained retrieval

1. Introduction

In recent years, we have seen tremendous increase in the production and consumption of multimedia data partly due to advent of the social web and partly because of the progress in surveillance, medical, industrial, mobile and embedded computing technologies [1]. Consequently, multimedia data including images and videos are produced and stored in huge amounts. These multimedia repositories contain wealth of highly useful information for administrators and decision makers, provided that efficient and reliable access to relevant data is ensured [2]. Content-based image retrieval (CBIR) systems attempt to locate images containing objects similar to that of a query image by analyzing their contents. CBIR has several applications in information retrieval, surveillance, medical, e-commerce, industry, and social web. Recently, it has attracted a lot of attention due to the rising interest in making the best use of available multimedia data [3]. The exponential increase in the volume of image data, and the inherent complexity of visual contents (projecting 3D world onto a 2D canvas) has made image retrieval increasingly difficult. This difficulty increases even further with fine-grained image retrieval due to the existence of high degree inter-class visual similarity [4]. One such problem arises when retrieving images from traffic surveillance datasets, where the main object of interest are vehicles [5, 6]. There exists greater visual similarity despite the fact that vehicles may belong to different categories.

Visual surveillance has become an undeniable necessity of the day, producing huge amounts of multimedia data, which is stored for future analysis [7, 8]. Indexing and retrieval of such huge volumes of data requires efficient representation methods [9, 10]. Though there exists numerous ways to represent visual contents in large datasets, complexity in the nature of visual data in surveillance limits the use of traditional image representation schemes. Earlier image retrieval methods used local features like scale-invariant features transform (SIFT) [11] and other feature aggregation schemes like vectors of locally aggregated descriptors (VLAD) [12] and fisher vectors (FV) [13]. In recent years, the success of CNN based features prevailed as the state-of-the-art features for image retrieval and classification. Some of the earlier works by Babenko and Lempitsky [14] and Razavian et al. [15] showed that features from a pre-trained CNN can be used to represent images, yielding state-of-the-art performance in large datasets. However, these approaches directly used activations from various layers without considering the suitability of these features for particular object classes.

In this paper, we investigated convolutional features maps of a pre-trained deep CNN to identify a set of optimal features for representing surveillance objects like vehicles for image retrieval applications. A feature selection procedure is presented for vehicles, allowing us to select appropriate features for fine-grained image search. Main contributions of our work are as follows:

- a. Convolutional activation features have been investigated for vehicles in order to select appropriate features for their effective representation.
- b. An efficient feature selection procedure is presented through which it is shown that the number of feature maps can be considerably reduced without any degradation in performance. The selected features exhibit greater attention to the object of interest than the background.
- c. It has also been shown through experiments that the selected features yield better retrieval performance at higher ranks than the full set of features.

The rest of the paper is organized as: Section 2 introduces relevant literature in the field of image retrieval. Section 3 presents schematics of the proposed approach. Experimental results are discussed in Section 4 and the paper is concluded with future research directions in Section 5.

2. Related Work

Content-based image retrieval has been extensively investigated by the multimedia research community for more than two decades [16, 17]. CBIR systems attempt to retrieve images based on visual content similarity, which require image representation as an essential ingredient [18, 19]. Traditionally, hand-engineered methods including bag-of-words histograms based on SIFT descriptors [20, 21], VLAD [12], GIST [22], and CENTRIST [23], etc. were used to represent images in retrieval systems. In recent years, image descriptors based on activations generated by deep CNNs have substantially improved the state-of-the-art for visual recognition [14, 24, 25]. Several methods were recently proposed for image retrieval which used activations of the fully connected (FC) layers as global descriptors. These methods provided much superior performance than the traditional hand-crafted features. However, directly matching these features in the Euclidean space is inefficient [26]. More recently, researchers found that the features from deep convolutional layers are more useful and naturally interpretable than the features from FC layers [27]. Each activation in the convolution layer is analogous to a local feature corresponding to a local receptive field. Furthermore, spatial layout of these local features is also preserved in convolutional feature maps, which makes them more useful for tasks like object detection and localization.

Deep features from the FC layers are like features from a black box which cannot be naturally interpreted [15]. On the contrary, convolutional layers contain volumes of information which needs to be effectively pooled to construct a global representation. Razavian et al. [15] showed that activations from a pre-trained

CNN can be used as generic features for a variety of tasks including object recognition, localization, and image retrieval. Gong et al. [28] utilized a multi-scale order-less pooling approach to aggregate CNN activations with VLAD based encoding. In [29], the authors aggregated features from the last convolutional layer through global max pooling and achieved better retrieval performance than several deep features based methods. Babenko and Lempitsky [25] showed that aggregating these features using global sum pooling yields better results than max pooling. Comor et al. [30] compared various global pooling approaches for image retrieval and showed that cross-dimensional weighting approach yielded better results with object-heavy datasets, whereas sum pooling performed the best for scene-based datasets.

Traditional image retrieval systems mainly focused on category-level image retrieval where a black car, red car, or a white sports car were considered as the same. However, in fine-grained image retrieval, the CBIR systems need to look deeper and rely on highly discriminative features in order to be able to differentiate between images belonging to the same category [31-33]. Fewer works have been carried out to investigate fine-grained image search. Wang et al. [32] presented a deep ranking method to learn similarity between images. However, their method relied on labelled sets of triplets which required considerable human efforts in annotation. Wei et al. [4] recently proposed selective convolutional descriptor aggregation approach where they utilized selected activations from convolutional layers and pooled those features through average and max pooling. Firstly they eliminated features corresponding to the background and then aggregated the selected descriptors to form a global representation. They showed state-of-the-art performance on several fine-grained image datasets. Our work is based on the findings in [4], however, instead of eliminating some convolutional activations from all the feature maps, we eliminate selective feature maps based on their negligible role in visual representation of the objects within a single category (in this case, vehicles). Based on our observations, we found that convolutional features from a pre-trained CNN model can serve as generic local descriptors for image retrieval which pay attention to particular regions in images. However, for a certain type of objects, all the features may not be equally important. Hence, we studied the role of convolutional feature maps in describing vehicles and found that only a subset of these maps were sufficient to represent vehicle images in surveillance datasets.

In large scale datasets, like surveillance, efficient indexing and retrieval methods are required [34]. In this context, locality sensitive hashing based approaches have been successfully applied to the domain of image retrieval [9]. Two different categories of methods exist: data independent methods like locality sensitive hashing (LSH) [35], spectral hashing (SH) [36], spherical hashing (SpH) [37], density sensitive hashing (DSH) [38], multi-feature hashing (MFH) [39], Kernelized LSH (KLSH) [40, 41], Iterative Quantization (ITQ) [42], Product Quantization (PQ) [43], Compact Quantization (CQ) [44], scalable graph hashing (SGH) [45], and sparse embedding and least variance encoding (SELVE) [46] etc., whereas learning-based methods including Deep Hashing (DH) [47], simultaneous feature learning and hashing [48], and deep semantic ranking based hashing [49], etc. Each of these methods have some strengths and limitations. For instance, LSH uses random projections to generate large number of hash tables in order to achieve considerable precision and recall, thereby requiring a lot of memory and time. DSH avoids purely random projections by utilizing geometric structure of the data, which helps it achieve better performance. KLSH generalizes LSH to the kernel space, however, its training is computationally expensive. SH is fast to train but conversion of features to binary codes requires a lot of time. SpH, on the other hand, requires a lot of time for training. All these methods perform well with image features and have exhibited state-of-the-art performances in large datasets. However, the representative strength of hash codes depends on the discriminative strength of the original features. Hashing methods attempt to project high dimensional feature vectors onto low-dimensional hamming space where feature vectors of

relevant images are placed near to each other such that approximate nearest neighbor search (ANN) approaches can efficiently access relevant data. We have investigated several hashing approaches to show the suitability of proposed features for large scale retrieval.

3. Proposed Method

In this section, we present the object-oriented convolutional features (OOCF) approach for fine-grained image search in large scale datasets. The proposed method consists of a feature selection process which attempts to identify appropriate features for representing objects of a particular class, based on feature attention. Then, the selected features are globally pooled to index and retrieve images. The method can be effectively applied to any type of images. Details of the feature selection, extraction, indexing and retrieval processes are provided in the subsequent sections.

3.1 Analysis of Convolutional Feature Maps

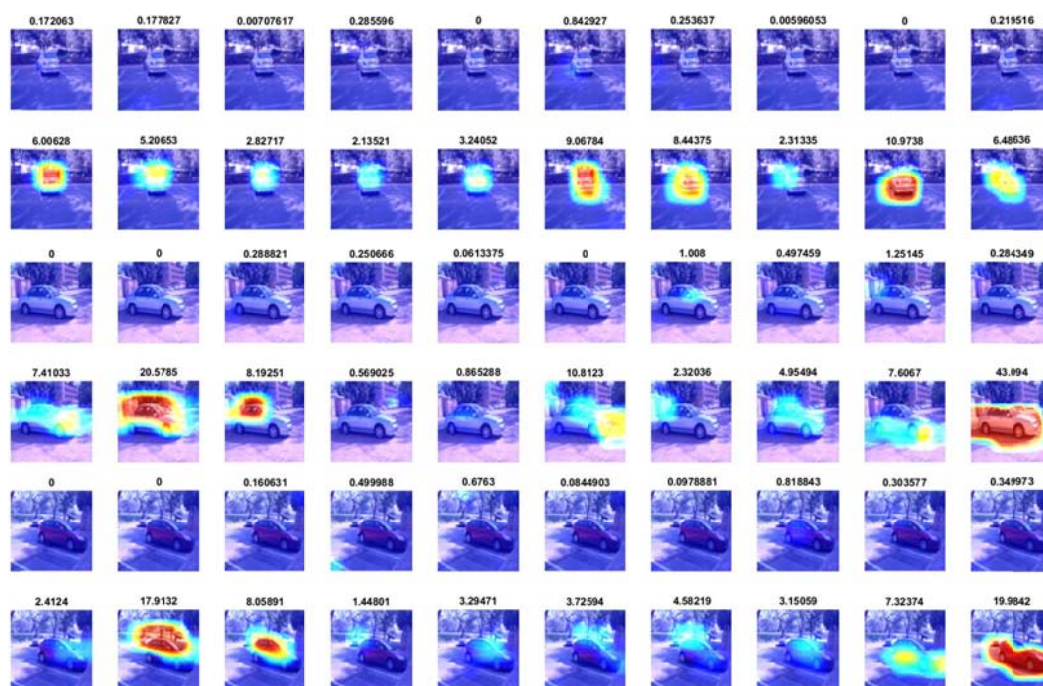
Convolutional layers in a deep CNN constitute an integral component of the feature learning pipeline. These layers learn progressively complex features of images with varying sizes of receptive fields. The shallower layers only look at small neighborhoods which merely contain primitive features, whereas the deeper layers have relatively larger receptive fields and hence are able to model high level semantics. Activations from the deeper convolutional layers and the fully connected (FC) layers have been widely used to retrieve images, yielding state-of-the-art performance in image retrieval. Recently, it has been shown that features from the convolution layers are more powerful and robust than the FC layer features [27]. A number of studies have been conducted to efficiently utilize convolutional features for image retrieval.

At each convolution layer, there exists a number of feature maps formed as a result of convolution operations on the input images. The number of feature maps depends on the number of kernels applied at a certain layer. Given an input image I having size $H \times W$, the output of a convolution layer is a tensor T having $h \times w \times d$ activations, where d corresponds the number of kernels at the layer, h and w represent height and width of the feature maps. The dimensions of the feature maps decrease as we move deeper in the CNN because of the strided convolutions and pooling operations.

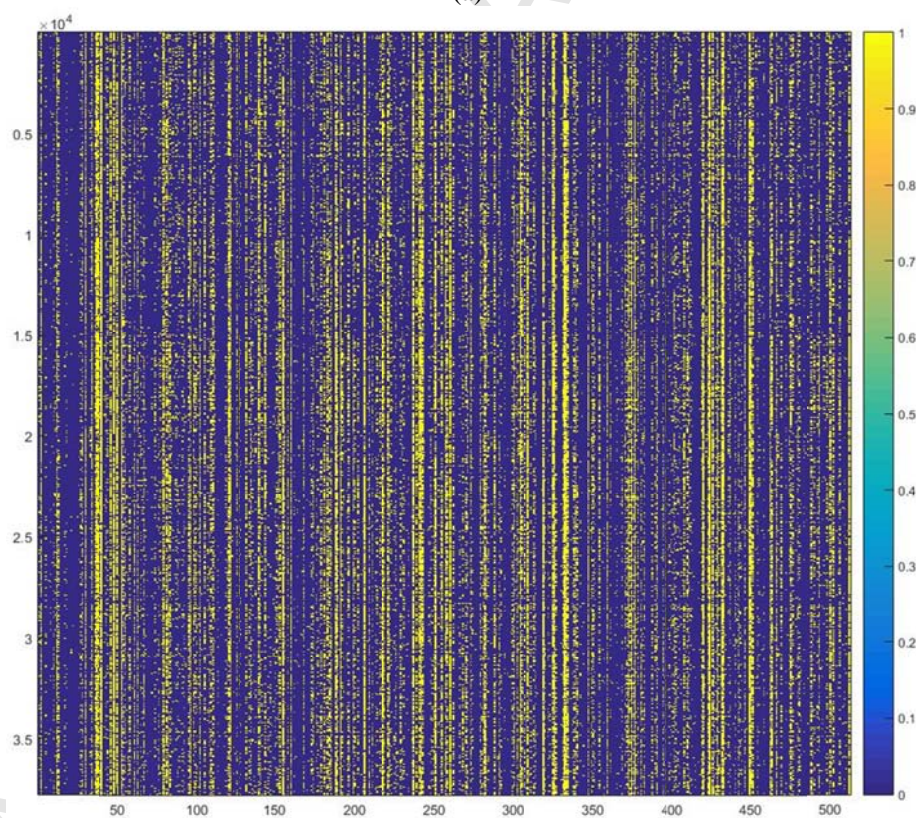
In this study, we investigated output of the “pool6” layer of a VGG-16 model [50] pre-trained on ImageNet dataset [51] having $6 \times 6 \times 512$ dimensions. The deepest convolution layer has the largest receptive field, thereby making each feature map sensitive of particular objects or parts of objects. Recently, Zhou et al. [52] showed that these feature maps or combinations of feature maps can be used to detect objects in images. Based on their results, we formed the basis of our study to determine optimal maps for convolutional features extraction of particular object sets. During this study, we analyzed the outputs of pool6 layer of the pre-trained model in vehicles dataset, and found that a large number of feature maps produce very weak activations or no activations at all. For instance, the feature maps shown in Fig 1 reveal that some feature maps have strong activations at particular parts of the vehicles (even rows in Fig 1), whereas others do not have any activations (odd rows). Though the model was trained on a huge dataset containing diverse images, the feature maps are sensitive to parts of vehicles which can be seen in the even numbered rows in Fig. 1. For different vehicles, the particular feature maps, produce stronger activations at the same object parts. Based on these observations, the proposed algorithm selects appropriate feature maps for a particular set of objects, which we call OOCF. The feature selection algorithm and its schematics are discussed in the following section.

3.2 Object Oriented Features Selection

Selection of appropriate convolutional features serves two important objectives. Firstly, it will allow us to utilize effective and discriminative features for object representation by eliminating irrelevant features. Secondly, the elimination of irrelevant features will result in reduction of the influence of the background in image representation process. As is evident from the feature maps in Fig. 1 and Fig. 2 that the strong activations are produced at parts of the objects rather than the background. Convolutional features with attention to objects of interest or their parts are more suited to represent images than those which may be sensitive to irrelevant objects or the background. The proposed feature selection algorithm is provided in Algorithm 1. We used training images from the vehicle re-identification (VeRI) dataset [53] which contains more than 37K images of vehicles captured by surveillance cameras, to perform the feature selection for vehicles. This dataset is suitable for feature selection because it contains cropped images of vehicles, thereby restricting the influence of background on the extracted features. Convolutional activations ($6 \times 6 \times 512 \times M$) from pool6 were generated for the M training images. Layer-wise mean activations for all the 512 feature maps were computed for M training images and stored. Afterwards, null activation values were identified and marked with 1 to construct a null utilization index (NUI) map as shown in Fig. 1(b). The dark blue columns correspond to the feature maps with non-null activations, whereas the yellow columns indicate those feature maps which generated no activations for the training images. Columns in the NUI maps can be seen which reflect upon our observation that significant number of feature maps can be eliminated without any degradation in performance. We then computed percentage frequencies for all the feature maps in the NUI map. At the end, feature maps with frequencies less than the threshold t were selected. This algorithm effectively eliminates those feature maps which do not react strongly or produce null activations for majority of the images. Remaining feature maps exhibited strong attention to semantic object parts, even in the presence of background objects. Responses of twenty selected feature maps on two images with full background are shown in Fig. 2. The sensitivity of selected feature maps to particular object parts is evident from the stronger activations in Fig. 2. Furthermore, minimal response can be seen for the background which effectively limit its role in the image representation process. Output of a single feature map (473) for eight different images is provided in Fig. 3. This map appears to be sensitive to upper part of the vehicle area where it has produced stronger activations. The mean activation value of the feature map is shown on top of each feature map in these figures.



(a)



(b)

Figure 1. (a) Sample responses of Deep Convolutional Features (b) null utilization index (NUI) map

Algorithm 1: Feature Map Selection

Input:

Training Image Set (TS)

Output

Selected Feature Maps (F_S)

Preparation:

1. Initialize the VGG16-CNN
2. Initialize Null Utilization Indices (NUI) having size $length(TS) \times F_N$ to 0.

Steps:

1. **for each** training image TS_i **in** TS
 - a. Forward propagate TS_i through VGG16-CNN
 - b. Extract $h \times w \times F_N$ feature maps from layer “pool6”
 - c. Compute global mean F_{mi} of each feature map F_i to obtain F_N values
 - d. Locate feature maps whose $F_{mi} = 0$.
 - e. Mark their indices with 1 in the NUI for TS_i
- end for**
2. Compute frequencies of null activations $FREQ_NA_i$ for each F_i
3. Compute percentages of frequencies $P_FREQ_NA_i$ for each F_i
4. Return all F_i as F_S whose $P_FREQ_NA_i < t$

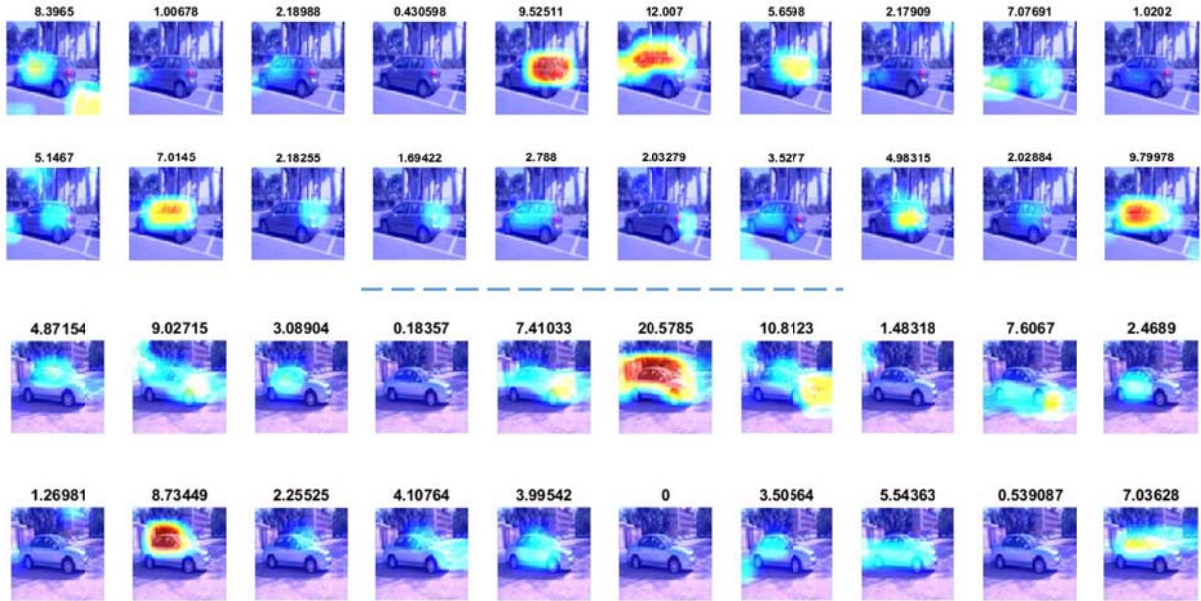


Figure 2. Response of Selective Features



Figure 3. Single feature response on multiple images (feature map # 473)

3.3 Object Oriented Convolutional Features

The proposed feature selection algorithm attempts to isolate feature maps which respond strongly to vehicles or parts of vehicles and produce negligible activations for other objects which may appear in the background. In this way, it minimizes the effects of background features when describing the objects of interest using convolutional features. It is evident from the feature maps shown in Fig. 3 where they responded strongly to parts of vehicles and produced negligible activations at the background. The given maps in Fig. 3 show responses of a single feature map (map # 473) on eight different images. The proposed features extraction framework illustrated in Fig. 4, can serve in fine-grained image classification and retrieval applications.

The input image is forward propagated through the VGG-16 pre-trained CNN. Feature maps from the pool6 layer ($6 \times 6 \times 512$) are extracted. The selected feature maps, identified through Algorithm 1, are isolated and their layer-wise global mean is computed. Each global mean represents the strength of response of a particular feature map for the input image. The combination of global layer-wise mean values for the selected feature maps result in the OOCF vector. For large scale retrieval, this feature vector can be projected into low-dimensional hamming space with locality sensitive hashing. Details of this transformation are provided in the subsequent section.

3.4 Transformation of OOCF to compact hash code

Hashing is a widely used approach for ANN search, which aims to transform high-dimensional feature vectors to low-dimensional hamming space. The resultant representation consists of a short sequence of binary digits, known as hash code. There exists two major categories of procedures for this transformation: locality sensitive hashing (LSH) or learning based hashing. The LSH methods are data independent and can be effectively applied to transform any feature vector to hash code, with locality sensitivity property. These approaches aim to map the query item to the target items in hamming space, allowing relevant items to be accessed efficiently and accurately using ANN search schemes. These characteristics of LSH techniques allow faster searching in big data by directly accessing areas of the feature space where potential relevant items could be found, thereby eliminating the need to exhaustively search the entire database. In our work, we evaluated six different schemes including DSH, SpH, MFH, SELVE, ITQ, and SGH. All of these data-independent methods aim to derive short binary representations for high dimensional features. The discriminative strength of hash codes is directly related to the discriminative capability of original features. The OOCF features will yield better retrieval performance if it is more discriminative than the full feature set. Hash-based image retrieval can be formulated as:

Given the query OOCF vector $q \in \mathbb{R}^d$ and the set of N d-dimensional vectors in the database $D \in \mathbb{R}^{d \times N}$ $D = \{f_1, f_2, \dots, f_n\}$, a set of hash functions H can be employed $H = \{h_1, h_2, \dots, h_k\}$ to compute a K -bit code $H_q = \{y_1, y_2, \dots, y_k\}$ for q such that

$$H_q = \{h_1(q), h_2(q), \dots, h_k(q)\} \quad (1)$$

where k^{th} bit is computed as $y_k = h_k(q)$. Each hash function performs the mapping as $h_k : \mathbb{R}^d \longrightarrow B$. This kind of encoding corresponds to mapping the original data point to a binary valued hamming space.

$$H : q \longmapsto \{h_1(q), h_2(q), \dots, h_k(q)\} \quad (2)$$

Given the set of hash functions, all data points in the database $D = \{f_1, f_2, \dots, f_n\}$ can be transformed to binary codes as:

$$H_f = H(D) = \{h_1(f), h_2(f), \dots, h_k(f)\} \quad (3)$$

where the hash codes for D will be $Y \in \mathbb{R}^{k \times N}$, the value of k corresponds to the number of hash functions applied on the data points or the length of the hash code. Typical hash code lengths range from 16-bits to 512-bits, depending on the type of image data being represented with these codes. Once the hash codes are computed, ANN search can be performed by computing hamming distance d_H between codes H_q and H_f as:

$$d_H(H_q, H_f) = |H_q - H_f| = \sum_{k=1}^K |h_k(H_q) - h_k(H_f)| \quad (4)$$

The objective of locality sensitive hash codes is to compute codes H for both q and f , such that the hamming distance between q and f strongly correlates with the hamming distance between H_q and H_f . If the Euclidean distance between q and f is large, then their hamming distance must also be large and vice versa. This characteristic will allow us to search relevant items in the hamming space without exhaustively searching the whole dataset. The hash code for q will lead us to a location in the hamming space where the probability of relevant items will be the highest. We will then retrieve the nearest neighbors and rank them according to their hamming distances from the query. Items with smaller distances will appear at higher ranks and those with larger distances will be placed at lower ranks. In big data, such a scheme can dramatically reduce search space and improve efficiency considerably [9].

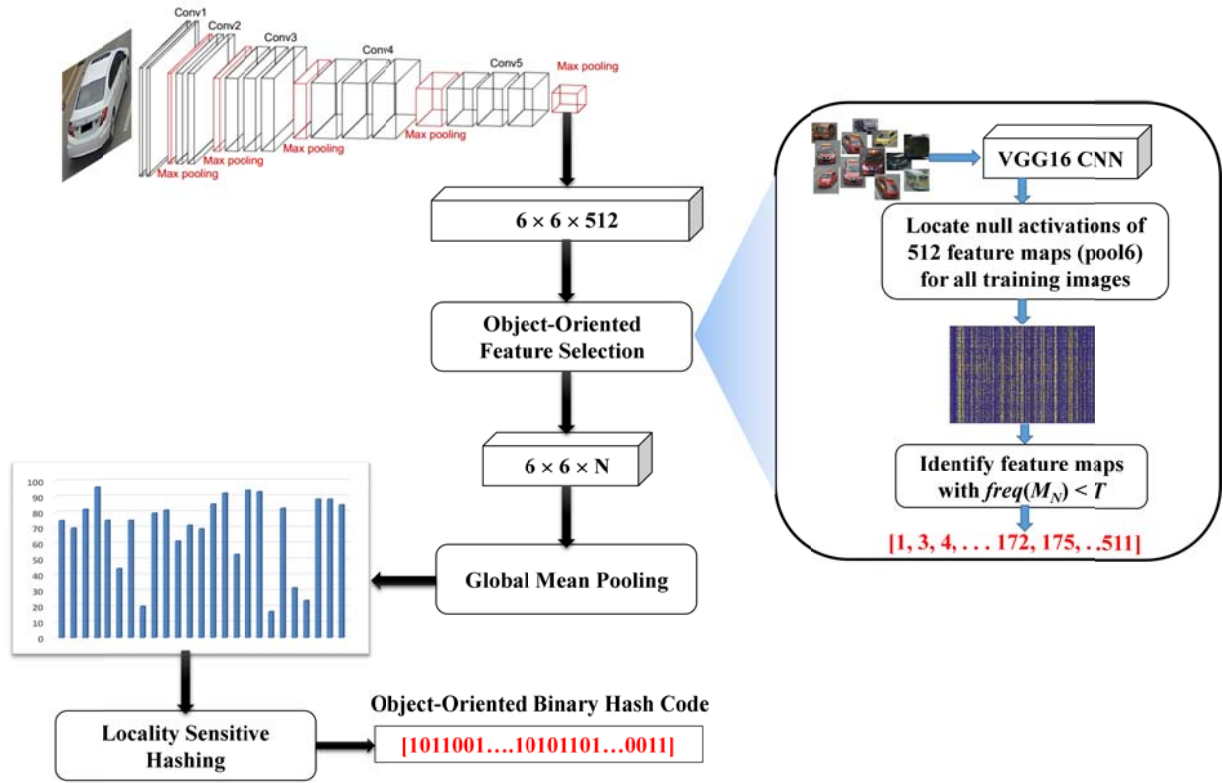


Figure 4. Proposed features extraction framework

4. Experiments and Results

The aim of this study was to develop a procedure to select appropriate features for representing objects of interest for fine-grained image search. We chose vehicle images captured by surveillance cameras to evaluate the proposed scheme. We also experimented with recent hashing methods to determine appropriateness of the proposed features for transforming them into compact binary codes. Results of various experiments and their outcomes are thoroughly discussed in this section.

4.1 Datasets and Evaluation Metrics

We evaluated the proposed method on three vehicles datasets namely Vehicle Re-Identification (VeRI) [53], Weizmann Cars ViewPoint (WCVP) [54], and Stanford Cars dataset [55]. The VeRI dataset consists of 51035 images of cropped vehicles captured from surveillance cameras. The entire dataset is split into training and test sets such that 37,778 images are used for training and 11,579 images are used for testing. The remaining 1678 images are used as query images. We used the training images of this dataset to select appropriate features for vehicles, and used the query images to measure retrieval performance in the test set. The second dataset, WCVP consists of 1464 images captured from 22 different cars. Cars in this dataset are not cropped and the dataset was used to evaluate retrieval performance of the selected features. This dataset was used because we wanted to analyze whether the selected OOCF limits the role of background in the overall image representation process. The third dataset Stanford Cars contains 16185 images, captured from 196 different cars. We used the entire dataset to test retrieval performance of the OOCF features which were selected using images from VeRI dataset.

Besides these datasets, we also evaluated retrieval performance using two other fine-grained datasets namely Aircraft [56] and Flowers [57]. The Aircrafts dataset consists of 10,200 images corresponding to 102 different aircraft models. Results of experiments are provided in the subsequent sections. The Flowers dataset consists of 1360 images corresponding to 17 different flower species. Both of these datasets contain background in significant amounts and the proposed method can be effectively used to extract features from the object of interest with little influence of the background.

Retrieval performance of the proposed method was measured in terms of precision (P), and recall (R) scores which are computed as:

$$P = \frac{\text{Number of Relevant images retrieved}}{\text{Total number of images retrieved}} \quad (5)$$

$$R = \frac{\text{Number of Relevant images retrieved}}{\text{Total Number of relevant images in the dataset}} \quad (6)$$

4.2 Results on VeRI Dataset

VeRI dataset consists of a test set having 11,559 images and a query set of 1678 images. To obtain performance scores on this dataset, we ran queries for all the images in the query set and retrieved top-k images using the proposed OOCF. We found that OOCF, despite having low dimensionality outperforms the full feature set, especially at higher ranks. This performance boost can be attributed to the selection of appropriate features for image representation, which eliminate the role of image background. Hence, the proposed features are able to effectively represent images even with different backgrounds as the chosen features focus only on the objects of interest. Results of randomly selected queries are shown in Fig. 5, where the first image is the query image and the remaining are top retrieved images using the proposed features. In the first query, our method retrieved accurately at top ranks, however some irrelevant, yet visually similar images were also retrieved at ranks 7-10, and 13-15. Likewise, some incorrect images were retrieved for 7th and 8th query due to visual similarity in colors and shapes. For the second and last two queries, the system was able to retrieve images with high precision, despite the fact that visually similar images of other vehicles were also present in the test set. Similarly, better performance was noticed for the rest of the queries where relevant images were retrieved accurately at top ranks, despite slight view variations, occlusions and different backgrounds. Quantitative results presented in Fig. 6 also exhibit superiority of the OOCF compared to the full convolutional features of the same layer. It is interesting to note that the proposed features yield better precision at lower recall rates but performs slightly poorly at very high recall rates. Since, typical CBIR systems seldom require all relevant images to be retrieved in response to a query. High precision at top ranks is usually favored over high precision at lower ranks.

In Fig. 6, real values in the legend correspond to the value of t used to select features in Algorithm 1, and the values inside brackets show the number of selected feature maps for the particular value of t . Choosing $t = 0.01$, only 123 highly reactive convolutional features maps are selected. However, effective representation of the vehicles is not achieved which leads to poor retrieval performance. Increasing value of t to 0.05, 230 feature maps were selected. With these features, significant improvement was noticed in the retrieval performance at all recall rates. Further increase in the number of features had little effect on performance. Best results were obtained for $t = 0.1$ where 267 feature maps were used to represent images. At this setting, slightly better precision rates were achieved for recall up to 0.4, as compared to other subsets of features. However, it performed relatively poor at high recall. If sufficient features are

selected for representing a particular type of object, the irrelevant features can be safely removed. At the same time, retrieval performance can be improved for low recall because of highly focused features used for object representation.

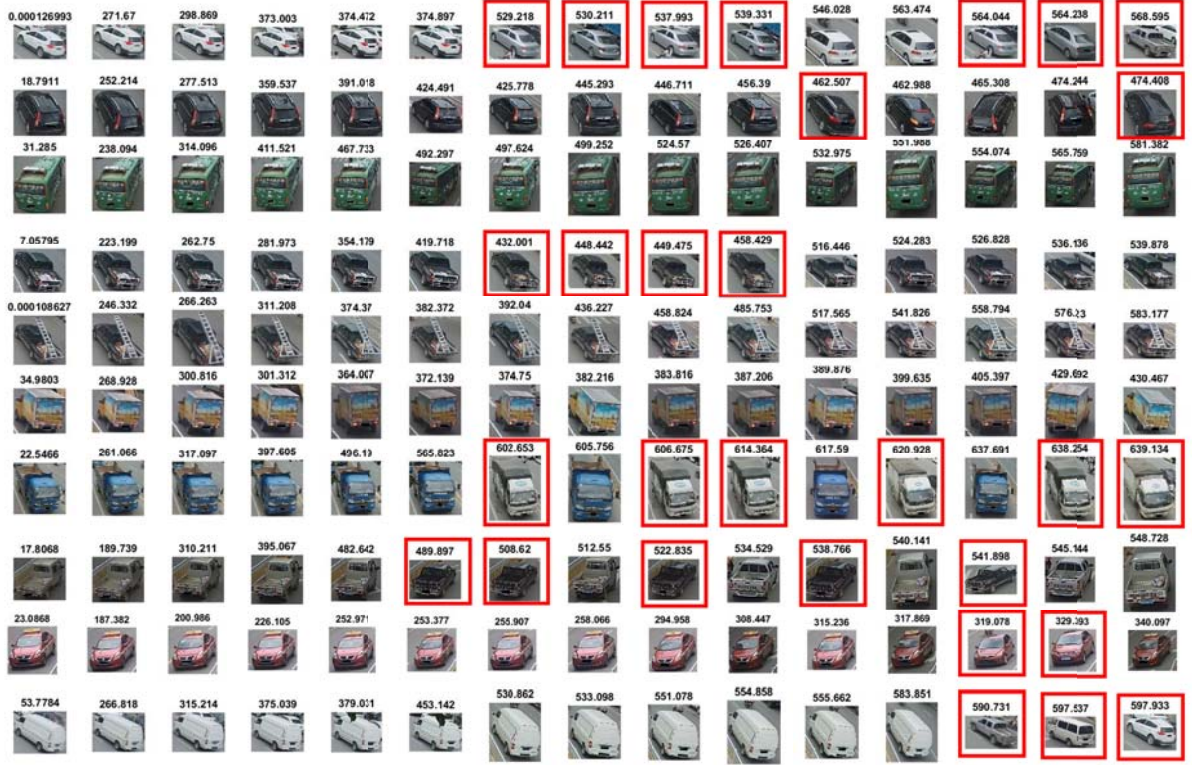


Figure 5. Retrieval results in VeRI dataset

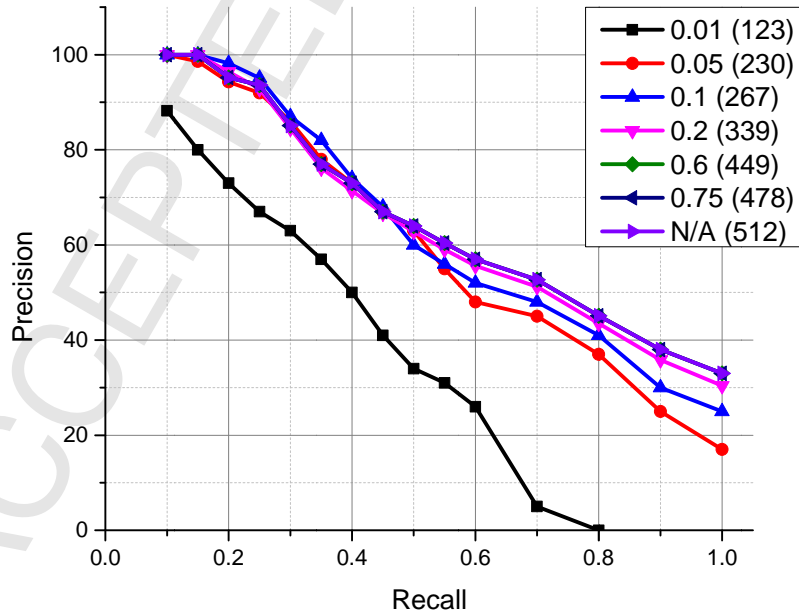


Figure 6. Precision recall rates for varying subsets of selected features (VeRI dataset)

4.3 Results on WCVF Dataset

This dataset is relatively smaller but more challenging than the VeRI dataset due to the presence of a variety of backgrounds. Images are captured from different viewpoints and the vehicles are not cropped due to which different background objects appear in images. Despite the differences in backgrounds, the proposed features successfully retrieved relevant images at top ranks. For instance as we move down the image ranks at first query, we notice the change in the background, yet the proposed method correctly retrieved them. Similar is the case with other queries, particularly, 3rd, 6th, and 9th query. Fig. 7 shows results of 10 random queries with top ranked images. Though some images were incorrectly retrieved at lower ranks, visual similarity of those images to the query image can be noticed which reflects upon the discriminative ability of the proposed features.

Similar to the qualitative results in VeRI dataset, the proposed features exhibit better precision for lower ranks as compared to the full set of convolutional features, shown in Fig. 8. It is important to highlight, that the improvement achieved with the optimal feature set is significantly higher than the previous dataset. This is because the presence of background in this dataset acts as a distractor in the object representation process. In the VeRI dataset, test images are also segmented images of vehicles with very little background. On the contrary, images in WCVF dataset contain significant background, providing much better set of images for proving effectiveness of the proposed features. When the full set of features is used to represent these images, features corresponding to the background result in retrieval of irrelevant images at top ranks. Instead, when the optimal subset of features are used, the influence of the background is significantly reduced, and precision at higher ranks improves. Results for both datasets show that once sufficient number of features are selected to represent objects, the rest of the features can be safely discarded without any drastic change in performance, particularly at higher recall rates.

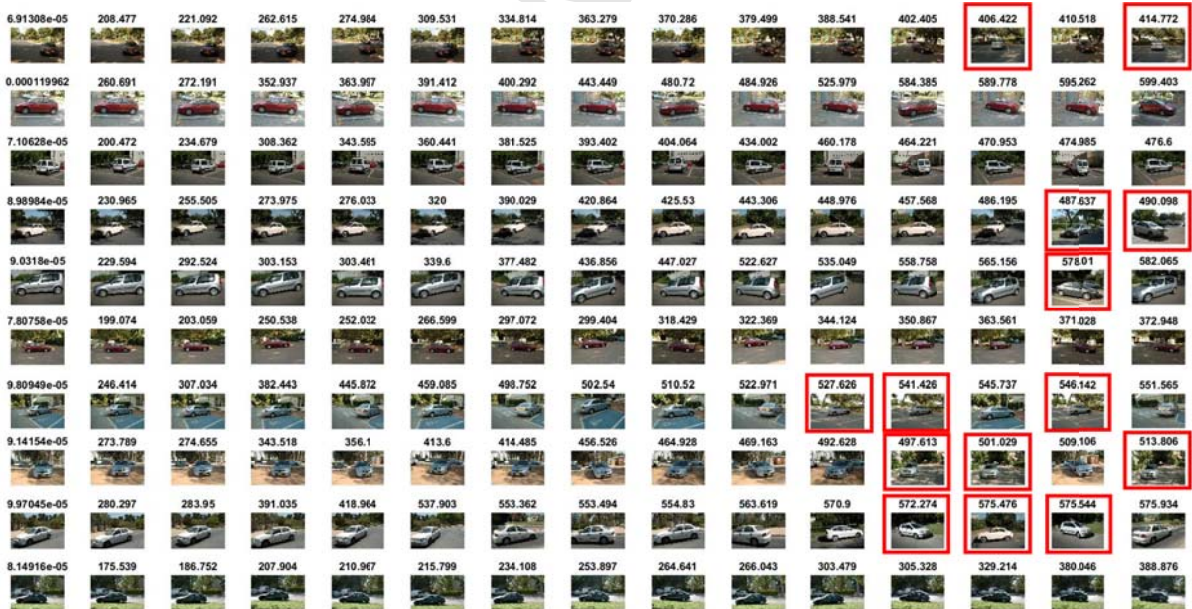


Figure 7. Retrieval results in WCVF dataset

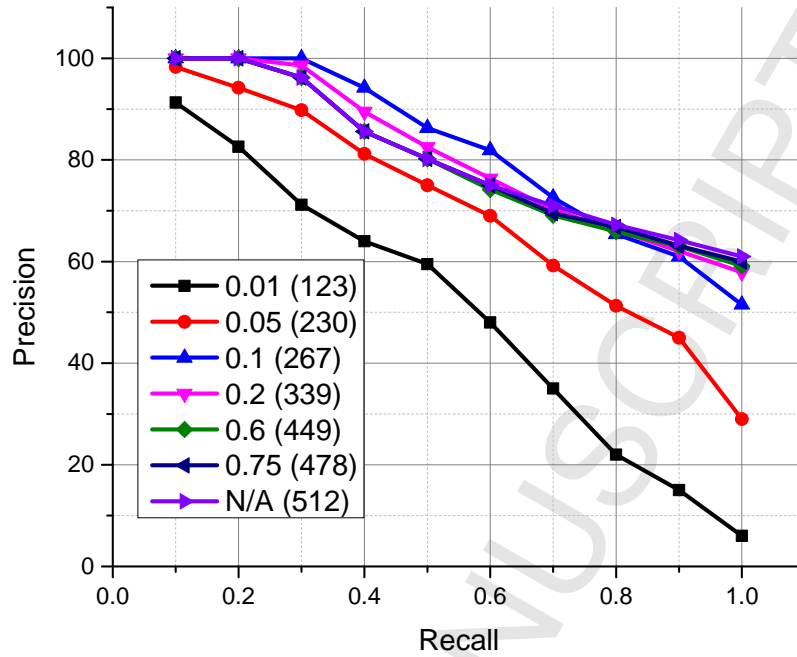


Figure 8. Precision recall rates for varying subsets of selected features (WCVP dataset)

4.4 Results on Stanford Cars Dataset

This dataset is the most challenging one, partly due to the huge volume of images and partly due to the high degree of viewpoint variations with which the images were taken. This is not primarily a surveillance dataset, rather the images were collected to evaluate fine-grained recognition tasks. Given a query image, the objective is to retrieve images of the same model car irrespective of the viewpoint variations, and changes in colors or textures etc. The images in this dataset were taken with varying backgrounds which make it a fine candidate for evaluating the suitability of our method. We extracted OOCF features from these images and used them to retrieve top ranked images. The objective was to retrieve as many relevant images as possible. Fig. 9 contains results of 10 randomly chosen query images, where the left most image is the query and the remaining are top ranked images based on L2 distance between the query and the dataset images. The images enclosed within red boxes indicate incorrect retrieval. In the first query, the proposed system retrieved accurately at ranks 3 and 7, irrespective of the fact that there is a high degree of variation in viewpoints and colors. Similarly, in query 2, top 5 images have been retrieved correctly. In the third query, relevant images were retrieved at ranks 2, 4, 5, 6, and 8, despite the fact that there exists huge disparity in their backgrounds. In the rest of the queries, relevant images were retrieved at top ranks which exhibit the capabilities of proposed features. Though the results on this dataset are not very strong, we believe that if a more powerful CNN model is used, these results can be significantly improved. Fig. 10 presents the precision recall scores for Stanford Cars dataset using varying subsets of features. Like the previous two datasets, the best results were achieved in this dataset with 267 features.



Figure 9. Retrieval results in Stanford Cars dataset

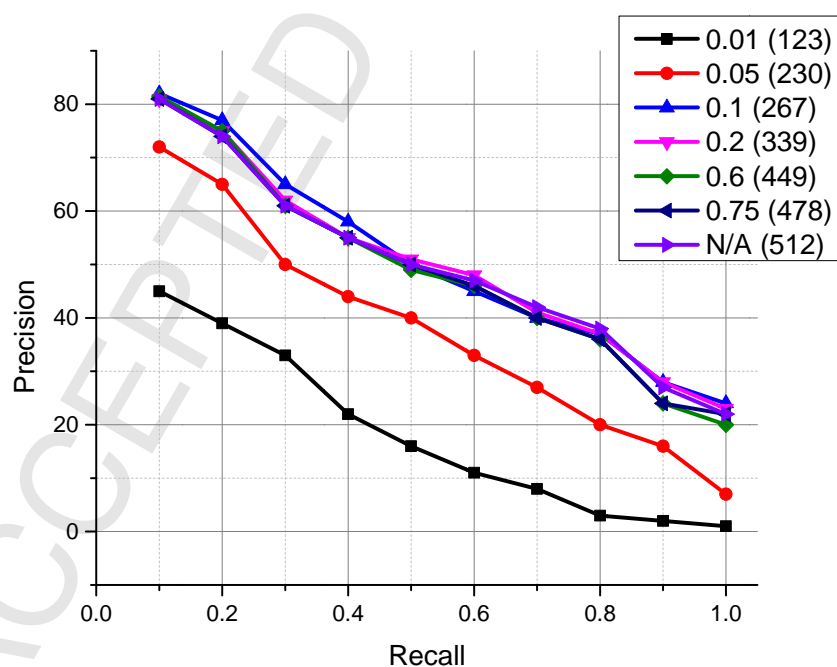


Figure 10. Precision recall rates for varying subsets of selected features (WCVP dataset)

4.5 Large scale image retrieval with hash codes based on OOCF

In these experiments, we evaluated the suitability of proposed features to be transformed to compact binary codes for large scale image retrieval using ANN search techniques. We used six different methods to derive hash codes of varying lengths for the proposed features. Precision recall scores were computed for each code and compared with the hash codes computed for the various subsets of features. Based on the parameter t in Algorithm 1, we selected seven distinct sets of features and computed hash codes of various lengths so that their performance could be compared with the hash codes generated by all the features. As witnessed in the previous experiments, eliminating 10 to 15% percent feature maps had absolutely no effect on retrieval performance. However, decreasing the number of features even further results in better performance at top ranks. It is interesting to note that this improvement appears only when a certain percentage of feature maps are eliminated. It is because the feature maps corresponding to the background have been eliminated and that the chosen features effectively model only the object features. Keeping in view these outcomes, we recommend at least 50 % of features to be selected so that effective representation for objects of interest could be achieved. We evaluated retrieval performance for code lengths 16, 32, 64, 128, and 256. However, 16, 32, and 64 bit codes could not achieve reasonable performance. Best retrieval rates were obtained for the optimal subset of features (at $t = 0.1$) with 128 bit hash codes for both datasets. A further increase in code length resulted in negligible improvements, therefore, we used 128-bits in these experiments. Retrieval performance for OOCF and full feature set with 128-bit hash codes generated using the different methods is provided in Fig. 11, 12, and 13 for VeRI, WCVF, and Stanford Cars datasets, respectively.

Fig. 11 shows precision recall scores for VeRI dataset with 128-bit hash codes generated using six different methods. In the previous experiments with these datasets, where OOCF features were used to retrieve images in this dataset, we see little improvements with OOCF as compared to the full feature set. Similar results have been achieved in these experiments as well. In DSH and SELVE methods, OOCF performs slightly better than the full feature set, whereas in other approaches, the performance is either similar or slightly poor. Negligible performance improvements in these results reflect the fact that the images in VeRI dataset were segmented and contained very little background. Hence, there exists no significant room for improvement for the OOCF. In Fig. 12, however, much better precision-recall scores have been achieved with OOCF in the WCVF dataset. This is due to the presence of background in this dataset, which allows OOCF to perform much better than the full feature set. OOCF yields better performance with SpH, DSH, SGH, SELVE and ITQ. In MFH, its performance remains the same for low recalls, however, its performance drops slightly at higher recalls. Similar kinds of results were achieved with Stanford Cars dataset (Fig. 13), where OOCF outperforms full feature set with SpH, DSH, MFH, SGH, and ITQ. With SELVE, it performs either similar scores or slightly lower scores. However, with majority of these methods, OOCF achieves better precision and recall scores, when there exists significant amounts of background in the images.

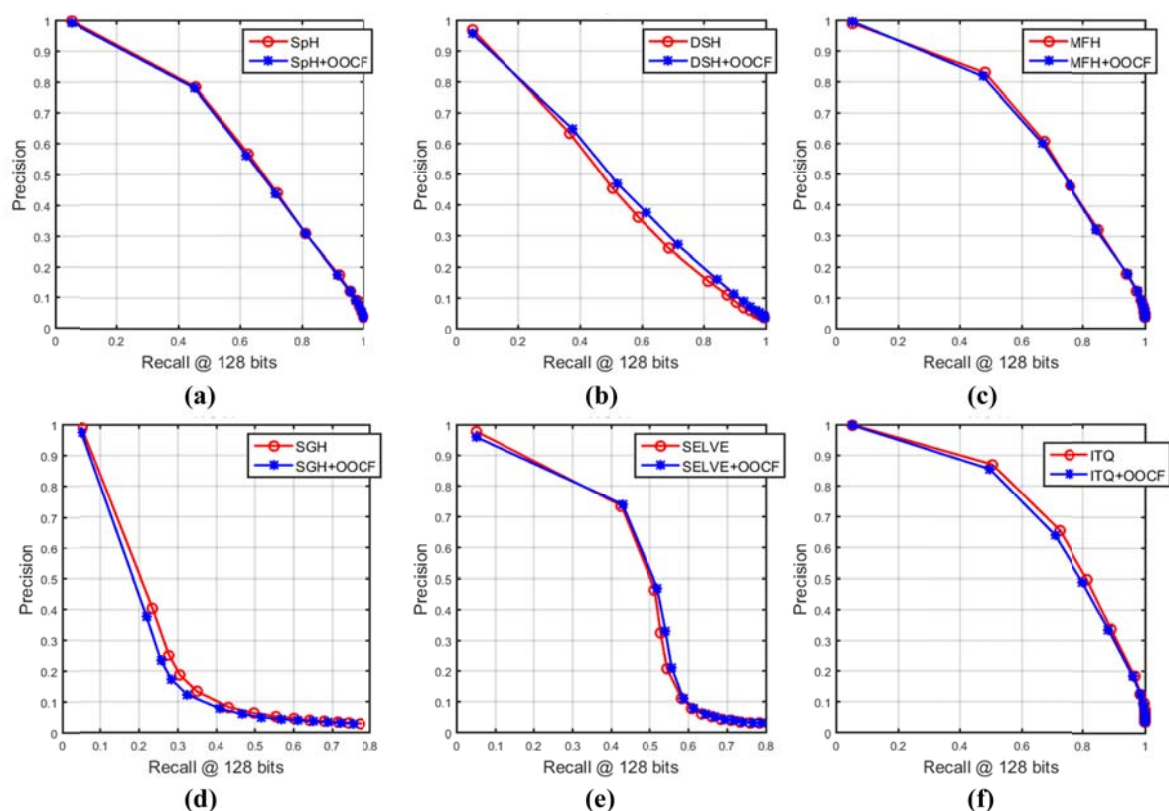


Figure 11. Retrieval performance for hash codes generated using full feature set and OOCF for VeRI dataset

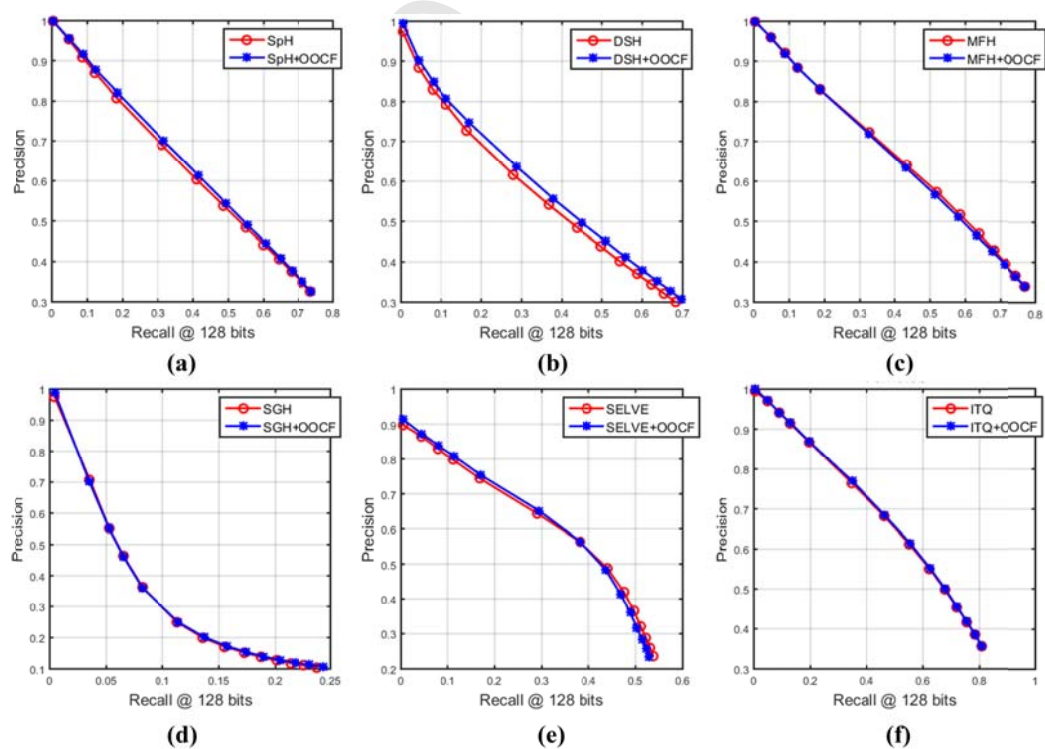


Figure 12. Retrieval performance for hash codes generated using full feature set and OOCF for WCVF dataset

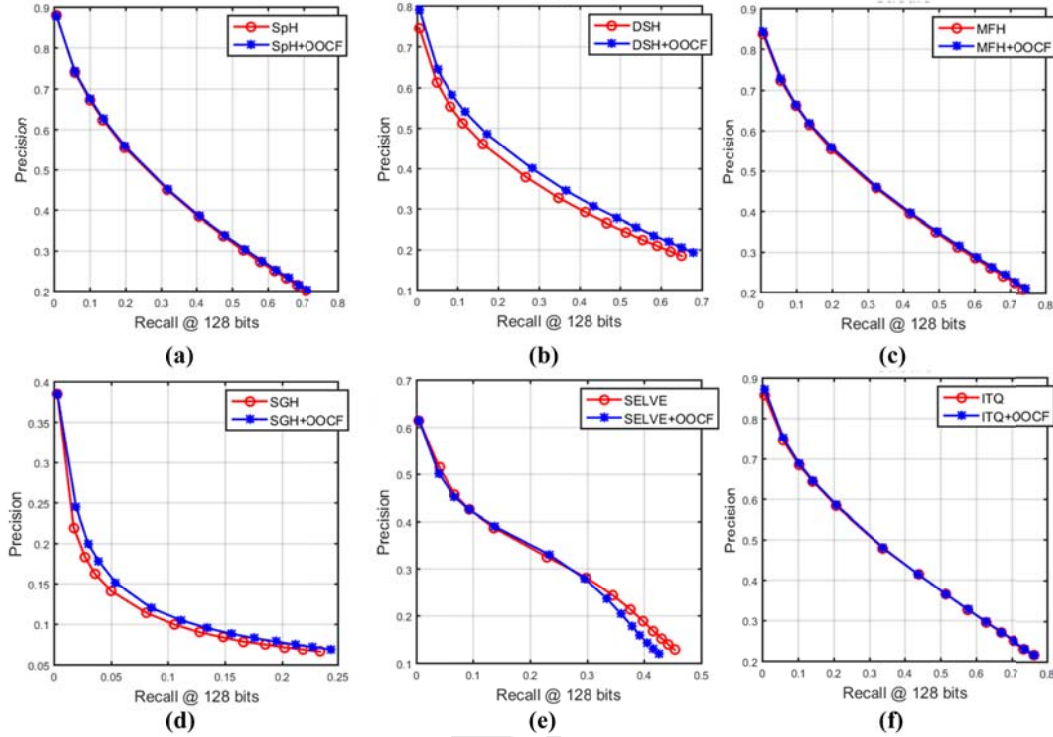


Figure 13. Retrieval performance for hash codes generated using full feature set and OOCF for Stanford Cars dataset

4.6 Performance of OOCF on other fine-grained retrieval datasets

In these experiments, we used two famous datasets for fine-grained image retrieval namely Aircraft and Flowers. The objective of these experiments, is to assess the capability of OOCF for object representation in general. Here, two different sets of objects were represented using OOCF and their retrieval performance was assessed using several hash-based retrieval algorithms. The Aircraft dataset is highly challenging due to the similar structure, and colors of the aircrafts. Even if the background is eliminated, it is very difficult to capture fine-grained discriminative features. On the contrary, the Flowers dataset is relatively easier because the difference among species of flowers can be relatively easily identified. For each of these datasets, we used a small portion of the dataset to select optimal features for representation using the proposed algorithm. The remaining images were used to retrieve similar images. For the Aircraft dataset, we merely used 10% of the dataset for feature selection, which amounts to 1020 images. The bounding boxes of the selected images were used to detect object oriented feature maps using Algorithm 1. We set $t = 0.15$ to obtain 284 features. In the remaining images, we retrieved top ranked images as shown in Fig. 14 (a), where the left most query image was used to retrieve the remaining images. Despite the highly challenging nature of this dataset, we were able to retrieve significant number of relevant images at top ranks. In the Flowers dataset, we segmented the flowers by eliminating the backgrounds containing leaves, and ground. The segmented flower images were used in the feature selection phase. For flowers, we obtained optimal set of 228 features by setting the value of t in Algorithm 1 to 0.2. Fig. 14 (b) contains results of top ranked images in Flowers dataset, where the proposed features accurately retrieved relevant images at top ranks. Only one image in the last query was

incorrectly retrieved at the last rank. The visible differences in the different flower species make retrieval in this dataset relatively easy. However, the similarity in background can affect retrieval performance if it is not effectively avoided.

Quantitative results depicted in Fig. 15 show that the proposed OOCF yield better retrieval scores than the full feature set on the Aircraft dataset. ITQ and MFH achieved the highest scores in terms of both precision and recall, whereas, SGH and SELVE achieved lowest scores on this dataset. With SpH, DSH, MFH, SGH, and ITQ, OOCF achieve relatively better scores than the full feature set. Similar results can be seen in Fig. 16 where OOCF achieves better performance with DSH, MFH, and ITQ, particularly at top ranks. These results vindicate the superiority of OOCF features for object based image representation in fine-grained image search applications. Identifying reactive convolutional features to objects of interest improves image representation which eventually leads to superior overall retrieval performance.

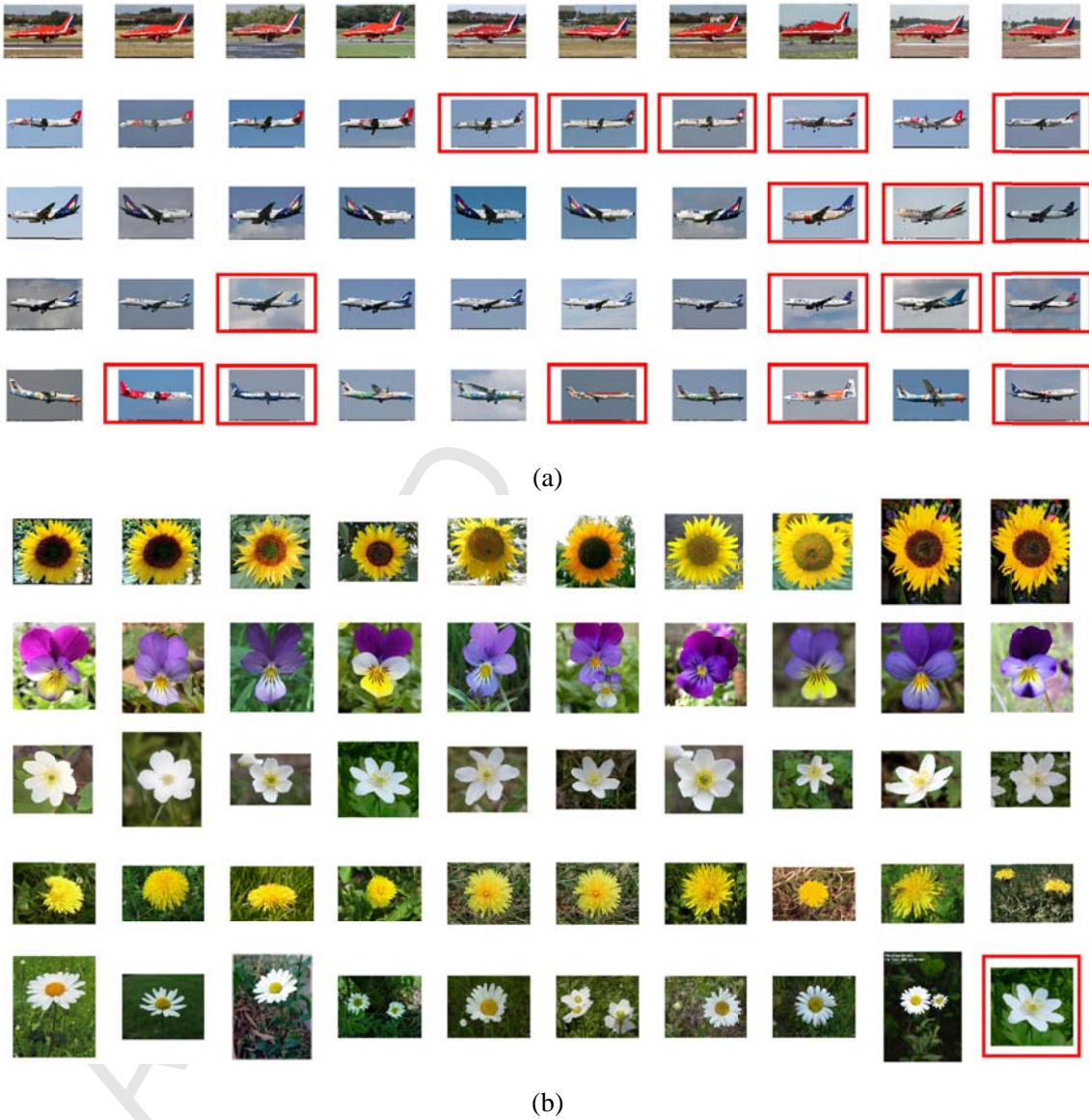


Figure 14. Retrieval results in (a) Aircraft and (b) Flowers dataset

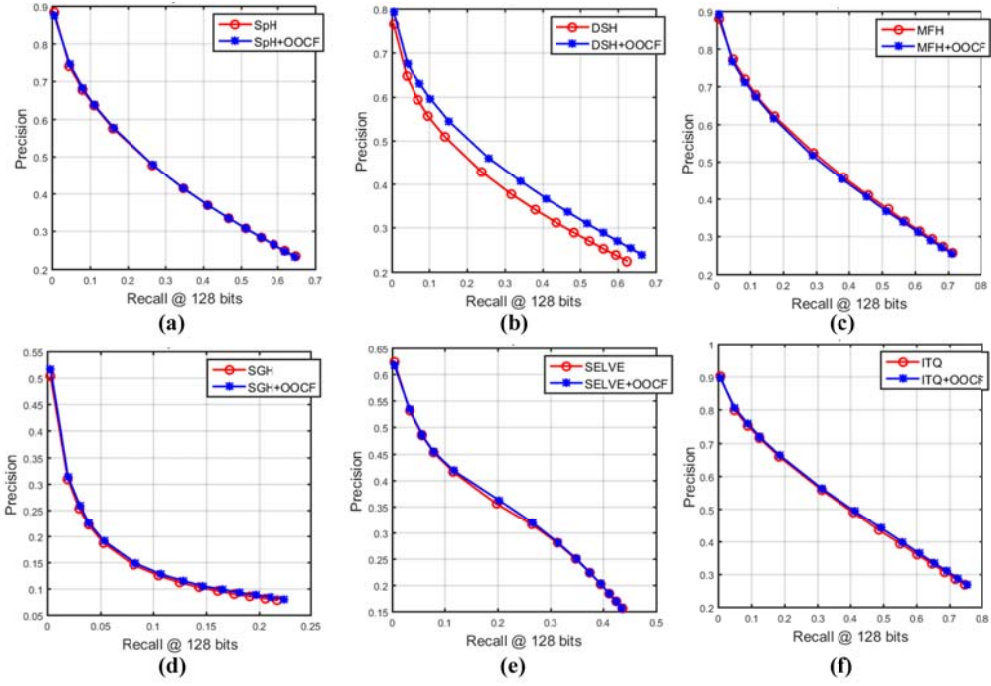


Figure 15. Retrieval performance for hash codes generated using full feature set and OOCF for Aircraft dataset

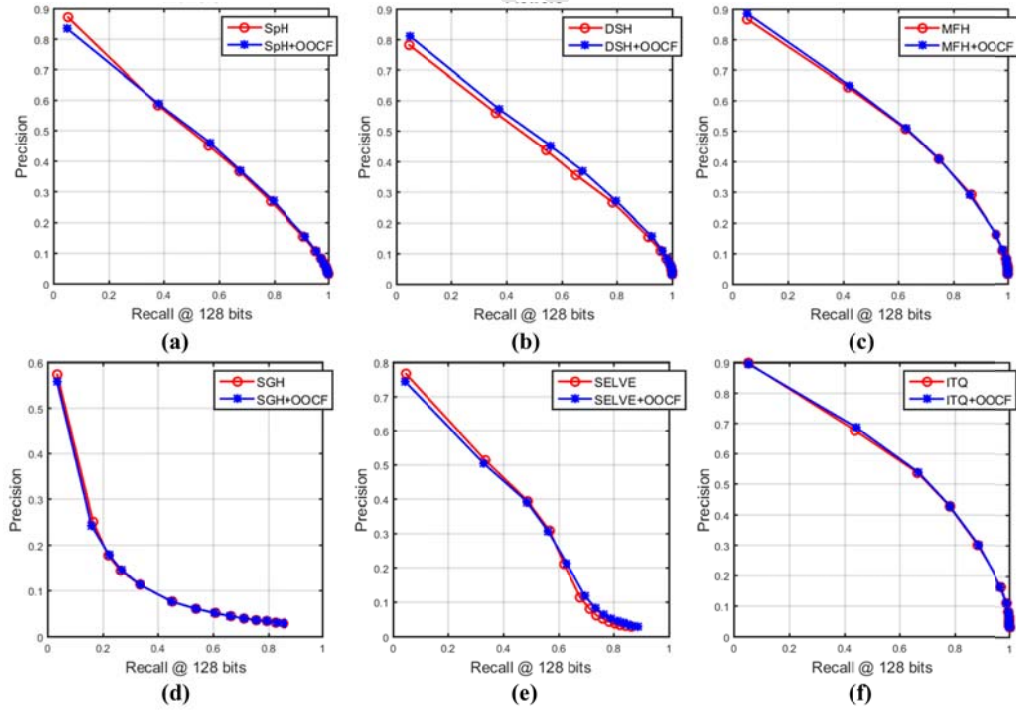


Figure 16. Retrieval performance for hash codes generated using full feature set and OOCF for Flowers dataset

4.7 OOCF sensitivity to target objects

In this section, we evaluated the proposed OOCF for detecting target objects (vehicles and aircrafts) in the presence of other objects or background. Since, convolutional feature maps preserve the spatial locations of detected features in images, the selected features (OOCF) can be used to detect and localize the objects of interest in images. In these experiments, we tested the OOCF for localization of vehicles in images other than the ones we used in the previous experiments. We also show that the selected feature maps can be utilized to detect any object of interest. Fig. 17 shows sample test images from the VOC2007 dataset [58] and their corresponding activations of the selected feature maps on those images. The activation maps for each image has been obtained by taking the mean of the selected activation values at each pixel location. Further, these maps have been resized to fit the size of the input image, so that the position of object can be identified. The activation maps overlayed on each image show that the OOCF features can effectively represent objects of interest, keeping them under focus. Such attention based features extraction has helped us achieve better results in the previous experiments. Fig. 17 (a) shows activations on vehicle images in the presence of various backgrounds as well as other objects. We used the VeRI dataset to select the feature maps for vehicles using the proposed algorithm. The selected feature maps were successful in detecting their locations in the images despite the variations in their scales. Vehicles with smaller sizes, partially occluded, and with viewpoint variations have been successfully detected. For aircrafts, we used some images from the Aircrafts dataset to select feature maps. Their responses on sample images are shown in Fig. 17 (b). Like the vehicles, the aircrafts have been successfully detected even in the presence of complex backgrounds. These results indicate that the proposed feature selection method effectively selects feature maps by analyzing their attention to our objects of interest.

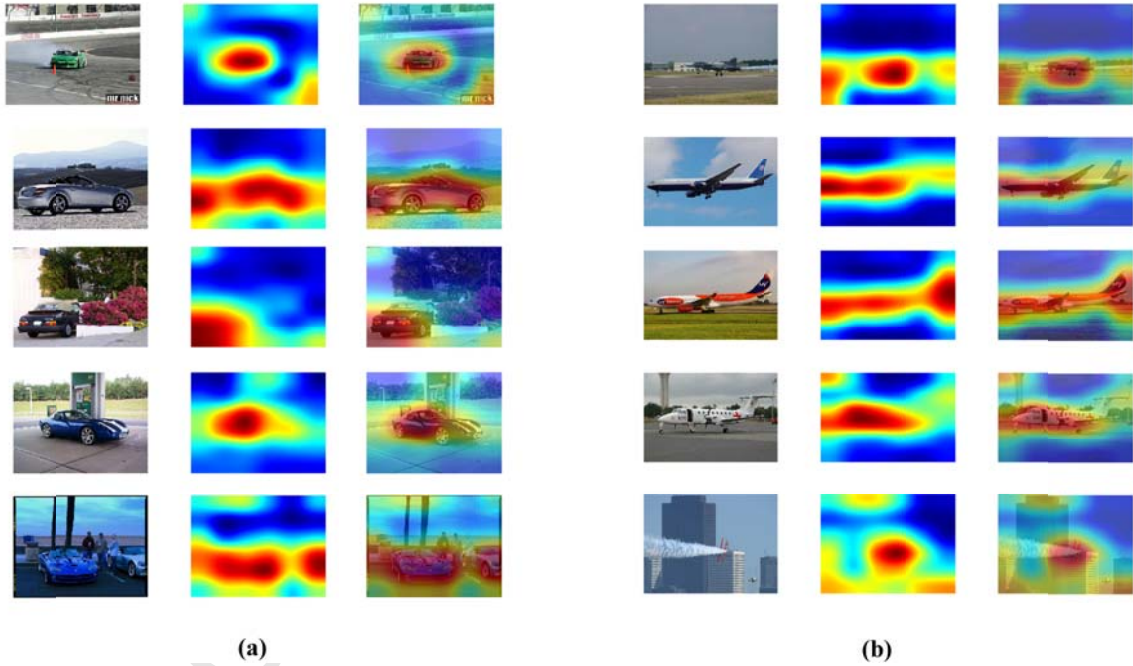


Figure 17. OOCF attention to features in the target objects (a) vehicles (b) aircrafts

5. Conclusions and Future Work

In this paper, we presented an efficient feature selection method of convolutional feature maps for a particular object category. The selected features focus on the objects of interest in the presence of background, eliminating the need to remove background prior to features extraction. We experimented on large surveillance datasets containing vehicle images captured by surveillance cameras, and two other

datasets. Analysis of the convolutional feature maps on segmented vehicles images revealed that a small subset of feature maps can adequately represent the vehicles. Some of the feature maps contained no activations whereas others produced negligible activations, which the proposed method attempted to eliminate. As a result, only those features are selected which discriminatively represent the objects of interest. Furthermore, the effect of background on the extracted features is also significantly reduced. We conducted several experiments to evaluate performance on vehicles and other datasets and showed that the selected features improve retrieval performance at higher ranks, particularly when optimal subset of features is selected. Furthermore, we also showed that the selected features can be effectively transformed into compact binary hash codes to allow efficient retrieval of images in large scale datasets.

In this study, the proposed method has been applied to vehicles, flowers, and aircrafts datasets, however, we strongly believe that the method can be easily applied to any fine-grained image recognition task. The only weakness of the current method is the requirement of segmented objects for the training set, which may not be always available for all datasets. Further research needs to be carried out to automatically select appropriate feature maps without requiring segmented objects of interest. One possibility is to employ visual saliency methods to identify object of interest and then use that information to isolate the foreground from background during the training process.

In future, we aim to improve the feature selection mechanism by considering more parameters through deeper analysis of the convolutional feature maps. Furthermore, we also intend to extract optimal features for object detection and localization, tracking, and representation of images having multiple objects. Currently, we focused on images with a single object. In future, we will try to extend the framework for multiple objects per image.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIP) (No.2016R1A2B4011712).

References

- [1] Y. Wu, G. Min, D. Zhu, and L. T. Yang, "An analytical model for on-chip interconnects in multimedia embedded systems," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 13, p. 29, 2013.
- [2] A. Gutub and N. Alharthi, "Improving Hajj and Umrah Services Utilizing Exploratory Data Visualization Techniques," presented at the Hajj Forum, Umm Al-Qura University – King Abdulaziz Historical Hall, Makkah, Saudi Arabia, 2016.
- [3] A. Gutub, "Exploratory Data Visualization for Smart Systems," in *Smart Cities 2015-3rd Annual Digital Grids and Smart Cities Workshop*, 2015.
- [4] X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE Transactions on Image Processing*, vol. 26, pp. 2868-2881, 2017.
- [5] J. Ahmad, I. Mehmood, S. Rho, N. Chilamkurti, and S. W. Baik, "Embedded deep vision in smart cameras for multi-view objects representation and retrieval," *Computers & Electrical Engineering*, vol. 61C, pp. 297-311, 2017 2017.
- [6] J. Ahmad, I. Mehmood, and S. W. Baik, "Efficient object-based surveillance image search using spatial pooling of convolutional features," *Journal of Visual Communication and Image Representation*, vol. 45, pp. 62-76, 2017.
- [7] Y. Wu, G. Min, K. Li, and B. Javadi, "Modeling and analysis of communication networks in multicluster systems under spatio-temporal bursty traffic," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, pp. 902-912, 2012.

- [8] N. A. Al-Otaibi and A. A. Gutub, "2-layer security system for hiding sensitive text data on personal computers," *Lect. Notes Inform. Theory*, vol. 2, 2014.
- [9] J. Wang, W. Liu, S. Kumar, and S.-F. Chang, "Learning to hash for indexing big data—a survey," *Proceedings of the IEEE*, vol. 104, pp. 34-57, 2016.
- [10] N. Alharthi and A. Gutub, "Data visualization to explore improving decision-making within Hajj services," *Sci Modell Res*, vol. 2, pp. 9-18, 2017.
- [11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91-110, 2004.
- [12] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, pp. 1704-1716, 2012.
- [13] M. Douze, A. Ramisa, and C. Schmid, "Combining attributes and fisher vectors for efficient image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 745-752.
- [14] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Computer Vision—European Conference on Computer Vision (ECCV)*, ed: Springer, 2014, pp. 584-599.
- [15] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 512-519.
- [16] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, pp. 1349-1380, 2000.
- [17] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, pp. 262-282, 2007.
- [18] J. Ahmad, M. Sajjad, I. Mehmood, S. Rho, and S. W. Baik, "Saliency-weighted graphs for efficient visual content description and their applications in real-time image retrieval systems," *Journal of Real-Time Image Processing*, pp. 1-17, 2016.
- [19] J. Ahmad, M. Sajjad, S. Rho, and S. W. Baik, "Multi-scale local structure patterns histogram for describing visual contents in social image retrieval systems," *Multimedia Tools and Applications*, vol. 75, pp. 12669-12692, 2016.
- [20] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*, 2007, pp. 197-206.
- [21] T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua, "Contextual bag-of-words for visual categorization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, pp. 381-392, 2011.
- [22] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, pp. 145-175, 2001.
- [23] J. Wu and J. M. Rehg, "CENTRIST: A visual descriptor for scene categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 1489-1501, 2011.
- [24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.
- [25] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1269-1277.
- [26] K. Lin, H.-F. Yang, J.-H. Hsiao, and C.-S. Chen, "Deep learning of binary hash codes for fast image retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 27-35.
- [27] L. Liu, C. Shen, and A. van den Hengel, "The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4749-4757.

- [28] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *European conference on computer vision*, 2014, pp. 392-407.
- [29] H. Azizpour, A. Sharif Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 36-45.
- [30] I. Comor, Y. Zhao, Z. Gao, L. Zhou, and L. Wang, "Image Descriptors from ConvNets: Comparing Global Pooling Methods for Image Retrieval," in *Digital Image Computing: Techniques and Applications (DICTA), 2016 International Conference on*, 2016, pp. 1-8.
- [31] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 49-58.
- [32] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1386-1393.
- [33] R. Zhang, J. Shen, F. Wei, X. Li, and A. K. Sangaiah, "Medical image classification based on multi-scale non-negative sparse coding," *Artificial Intelligence in Medicine*, 2017.
- [34] V. Sugumaran, A. K. Sangaiah, and A. Thangavelu, "Computational Intelligence Applications in Business Intelligence and Big Data Analytics," ed: CRC Press, 2017.
- [35] M. S. Charikar, "Similarity estimation techniques from rounding algorithms," in *Proceedings of the thirty-fourth annual ACM symposium on Theory of computing*, 2002, pp. 380-388.
- [36] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Advances in neural information processing systems*, 2009, pp. 1753-1760.
- [37] J.-P. Heo, Y. Lee, J. He, S.-F. Chang, and S.-E. Yoon, "Spherical hashing," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 2957-2964.
- [38] Z. Jin, C. Li, Y. Lin, and D. Cai, "Density sensitive hashing," *IEEE transactions on cybernetics*, vol. 44, pp. 1362-1371, 2014.
- [39] J. Song, Y. Yang, Z. Huang, H. T. Shen, and R. Hong, "Multiple feature hashing for real-time large scale near-duplicate video retrieval," in *Proceedings of the 19th ACM international conference on Multimedia*, 2011, pp. 423-432.
- [40] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, pp. 1092-1104, 2012.
- [41] J. Ahmad, M. Sajjad, I. Mehmood, and S. W. Baik, "SiNC: Saliency-injected neural codes for representation and efficient retrieval of medical radiographs," *PloS one*, vol. 12, p. e0181707, 2017.
- [42] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 817-824.
- [43] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, pp. 117-128, 2011.
- [44] T. Zhang, C. Du, and J. Wang, "Composite Quantization for Approximate Nearest Neighbor Search," in *ICML*, 2014, pp. 838-846.
- [45] Q.-Y. Jiang and W.-J. Li, "Scalable Graph Hashing with Feature Transformation," in *IJCAI*, 2015, pp. 2248-2254.
- [46] X. Zhu, L. Zhang, and Z. Huang, "A sparse embedding and least variance encoding approach to hashing," *IEEE transactions on image processing*, vol. 23, pp. 3737-3750, 2014.
- [47] V. Erin Liong, J. Lu, G. Wang, P. Moulin, and J. Zhou, "Deep hashing for compact binary codes learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2475-2483.

- [48] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3270-3278.
- [49] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1556-1564.
- [50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 248-255.
- [52] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," *arXiv preprint arXiv:1412.6856*, 2014.
- [53] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Multimedia and Expo (ICME), 2016 IEEE International Conference on*, 2016, pp. 1-6.
- [54] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich, "Viewpoint-aware object detection and continuous pose estimation," *Image and Vision Computing*, vol. 30, pp. 923-933, 2012.
- [55] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 554-561.
- [56] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.
- [57] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on*, 2008, pp. 722-729.
- [58] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303-338, 2010.

Authors' Biographies

Jamil Ahmad received his BCS degree in Computer Science from the University of Peshawar, Pakistan in 2008 with distinction. He received his Master's degree in 2014 with specialization in Image Processing from Islamia College, Peshawar, Pakistan. He is also a regular faculty member in the Department of Computer Science, Islamia College Peshawar. Currently, he is pursuing PhD degree in Sejong University, Seoul, Korea. His research interests include deep learning, medical image analysis, content-based multimedia retrieval, and computer vision. He has published several journal articles in these areas in reputed journals including Journal of Real-Time Image Processing, Multimedia Tools and Applications, Journal of Visual Communication and Image Representation, PLOS One, Journal of Medical Systems, Computers and Electrical Engineering, SpringerPlus, Journal of Sensors, and KSII Transactions on Internet and Information Systems. He is also an active reviewer for IET Image Processing, Engineering Applications of Artificial Intelligence, KSII Transactions on Internet and Information Systems, Multimedia Tools and Applications, IEEE Transactions on Image Processing, and IEEE Transactions on Cybernetics. He is a student member of the IEEE.

Khan Muhammad (S'16) received the bachelor's degree in computer science from the Islamia College Peshawar, Pakistan, in 2014, with a focus on information security. He is currently pursuing the M.S. leading to Ph.D. degree in digital contents from Sejong University, Seoul, South Korea. He has been a Research Associate with the Intelligent Media Laboratory since 2015. He has authored over 24 papers in peer-reviewed international journals and conferences, such as Future Generation Computer Systems, the IEEE ACCESS, the Journal of Medical Systems, Biomedical Signal Processing and Control, Multimedia Tools and Applications, Pervasive and Mobile Computing, SpringerPlus, the KSII Transactions on Internet and Information Systems, the Journal of Korean Institute of Next Generation Computing, the NED University Journal of Research, the Technical Journal, the Sindh University Research Journal, the Middle-East Journal of Scientific Research, MITA 2015, PlatCon 2016, and FIT 2016. His research interests include image and video processing, information security, image and video steganography, video summarization, diagnostic hysteroscopy, wireless capsule endoscopy, computer vision, deep learning, and video surveillance.





Sambit Bakshi is currently with Centre for Computer Vision and Pattern Recognition of National Institute of Technology Rourkela, India. He also serves as Assistant Professor in Department of Computer Science

& Engineering of the institute. He earned his PhD degree in Computer Science & Engineering in 2015. He serves as associate editor of International Journal of Biometrics, IEEE Access, and Plos One. He is technical committee member of IEEE Computer Society Technical Committee on Pattern Analysis and Machine

Intelligence. He received the prestigious Innovative Student Projects Award - 2011 from Indian National Academy of Engineering (INAE) for his master's thesis. He has more than 30 publications in journals, reports, conferences.

Sung Wook Baik received the B.S degree in computer science from Seoul National University, Seoul, Korea, in 1987, the M.S. degree in computer science from Northern Illinois University, Dekalb, in 1992, and the Ph.D. degree in information technology engineering from George Mason University, Fairfax, VA, in 1999. He worked at Datamat Systems Research Inc. as a senior scientist of the Intelligent Systems Group from 1997 to 2002. In 2002, he joined the faculty of the College of Electronics and Information Engineering, Sejong University, Seoul, Korea, where he is currently a Full Professor and Dean of Digital Contents. He is also the head of Intelligent Media Laboratory (IM Lab) at Sejong University. He served as professional reviewer for several well-reputed journals such as IEEE Communication Magazine, Sensors, Information Fusion, Information Sciences, IEEE TIP, MBEC, MTAP, SIVP and JVCI. His research interests include computer vision, multimedia, pattern recognition, machine learning, data mining, virtual reality, and computer games. He is a professional member of the IEEE.

Authors Photographs

			
Jamil Ahmad	Khan Muhammad	Sambit Bakshi	Sung Wook Baik

Highlights

1. Convolutional activation features have been investigated for vehicles in order to select appropriate features for their effective representation.
2. An efficient feature selection procedure is presented through which it is shown that the number of feature maps can be considerably reduced without any degradation in performance.
3. It has also been shown through experiments that the selected features yield better retrieval performance at higher ranks than the full set of features.