A Salient Based Bag of Visual Word Model (SBBoVW): Improvements toward Difficult Object Recognition and Object Location in Image Retrieval

Leila Mansourian[1], Muhamad Taufik Abdullah[1],
Lilli Nurliyana Abdullah[1], Azreen Azman[1], Mas Rina Mustaffa[1]
[1] University Putra Malaysia, Faculty of Computer Science and Information Technology, Department of Multimedia,
UPM Serdang, Selangor Darul Ehsan, Malaysia, 43400
[e-mail: leila.m@student.upm.edu.my, {mta, liyana, azreenazman, masrina} @upm.edu.my]
*Corresponding author: Muhamad Taufik Abdullah

Received March 6, 2015; revised June 22, 2015; revised August 27, 2015; revised September 29, 2015; accepted October 22, 2015; published February 29, 2016

Abstract

Object recognition and object location have always drawn much interest. Also, recently various computational models have been designed. One of the big issues in this domain is the lack of an appropriate model for extracting important part of the picture and estimating the object place in the same environments that caused low accuracy. To solve this problem, a new Salient Based Bag of Visual Word (SBBoVW) model for object recognition and object location estimation is presented. Contributions lied in the present study are two-fold. One is to introduce a new approach, which is a Salient Based Bag of Visual Word model (SBBoVW) to recognize difficult objects that have had low accuracy in previous methods. This method integrates SIFT features of the original and salient parts of pictures and fuses them together to generate better codebooks using bag of visual word method. The second contribution is to introduce a new algorithm for finding object place based on the salient map automatically. The performance evaluation on several data sets proves that the new approach outperforms other state-of-the-arts.

Keywords: saliency map, SIFT feature, Bag of Visual Words model (BoVW), image retrieval, object recognition, and object location.

http://dx.doi.org/10.3837/tiis.2016.02.018 **ISSN : 1976-7277**

1. Introduction

Finding an appropriate method to retrieve objects inside images is a controversial issue among the image retrieval and image annotation researchers. Most of these foundational algorithms used Bag of Visual Words method (BoVW) Csurka et al. [1] for image classification, treating image features as words. BoVW is a number of occurrences of a vocabulary of image features, and an image can be conducted as a document. Furthermore, the definition of "words" is obligatory, which in images is it also needs to be defined. To apply the BoVW method, the following steps are taken into consideration: Feature detection, feature description and codebook generation. Most of the recent works used BoVW with Scale Invariant Feature Transform (SIFT) by Lowe [2]. It is not the subject to affine, transformations, occlusion, lighting as well as intra-class variations. Other features are usually not flexible on size and their performances are usually unsatisfactory. This is because global features cannot express the basic objects in the image individually. Indeed, among difficult objects like frog or camel that are hiding on their backgrounds by Vedaldi and Fulkerson [3] and Wang et al. [4] got less accurate results.

In the previous experiment, the researchers of the current study used BoVW of SIFT features with different SVM methods (using LIBSVM Chang and Lin [5]) for animal recognition. However, most of the animals are the same as their environment. Doing so, nature protects them against enemies. The problem is, in the traditional BoVW model in which cannot collect visual words based on their locations in the picture. Therefore, all the visual words collected and treated the same as each other even if they are from the important part or background of the picture. Based on Oquab et al. [6], this means that the classifier often relies on visual words that fall in the background and merely describe the context of the object. Therefore, the recognition of difficult objects requires more precision and care.

Thus, based on this problem, a Salient Based Bag of Visual Word (SBBoVW) model for difficult object recognition and object location is proposed. This model can collect visual words of the whole and salient part of the picture on the basis of spatial histograms to overcome the problem mentioned above. We used Caltech-101 object categories dataset of Fei-Fei et al. [7], which has around 40 to 800 images per category. It contains a total of 9146 images, split between 101 distinct objects (including faces, watches, ants, pianos, etc.) and a background category (for a total of 102 categories). Additionally, an algorithm was run on other datasets like Caltech 256 by Griffin et al. [8].

A. Vedaldi, B. Fulkerson 2010’s PHOW method and the previous experiment (BoVW) were selected. Our dataset was Caltech 101 as well as a subset of Caltech 256. For comparison, 13 states of the arts were chosen. The comparison of all the approaches was done based on the same datasets. It has to be mentioned that, the comparison of Veldaldi 2010’s research and BoVW are done on the same train and test images.

It is worth mentioning that our train approach consists of six main steps. The first one is saliency map computation based on Jiang et al. [9], which discriminates the background from the main object. The second step is a saliency of the rectangular parts of the dataset pictures which is extracted to speed up computation dense SIFT feature selected for feature extraction. The third step is the SIFT feature of rectangular saliency parts and normal pictures that are also extracted. The codebook is created in the fourth stage by K-Means classification. In the fifth stage, spatial histogram descriptors are quantized based on the KD-trees to get the visual words. Then, in order to test our model, spatial histogram of visual words of testing pictures are compared with spatial histograms of visual words based on SVM Chi-square. Afterward, the appropriate concept names are extracted from test images. Finally, a new algorithm is conducted to find object place.

The rest of the current paper is structured as follows. In Section 2, some related works are reviewed. Section 3 introduces SBBoVW model toward difficult object recognition. Section 4 presents experiments and the experimental setup. A discussion of the proposed model, research results and usefulness of SBBoVW model are going to be explained in section 5. The paper is concluded by making some comments on future ideas in section 6.

2. Related Works

Despite significant progress shown by statistical approaches for images annotation, finding an appropriate method for object recognition and object location is a recent controversial discussion in Bannour and Hudelot [10], Kim and Yoon [11], Zhong et al. [12], Long et al. [13], Kulkarni et al. [14], Murphy et al. [15], Lampert et al. [16]. In the traditional BoVW model, all the visual words are collected and treated the same as each other even if they are from the important part or background of the picture. Oquab et al. [6] mentioned this means that the classifier often relies on visual words that fall in the background and merely describe the context of the object. Based on this problem, a Salient Based Bag of Visual Word (SBBoVW) model for difficult object recognition and object location is proposed. This model can collect visual words of the whole and salient part of the picture. In what follows, we first briefly review common stages in object recognition and object location techniques. After that, in the last subsection, the proposed method will be compared and contrasted with others'.

The first but not the compulsory stage is image segmentation. The segmentation algorithm divides images into different parts based on feature similarity. Different segmentation approaches proposed in the literature, are background removing based, clustering based, grid based, model based, contour based, graph-based, region growing based and salient based method. For a comprehensive segmentation review, readers are referred to Dey et al. [17]. In this study, the focus is on salient based methods. Because of the object location, removing the background parts is one of the important stages. Recently, many types of research have been done to design various models to compute the saliency maps. Based on the survey research conducted by Borji et al. [18], there are two attributes for detecting salient or interesting objects in images: *Block-based vs. Region-based analysis* and *intrinsic cues vs. Extrinsic cues*.

Block-based vs. Region-based analysis. Block (i.e. pixels and patches) based is an early method of finding a salient object, while regions are widespread generation with the development of superpixel algorithms.

Intrinsic cues vs. Extrinsic cues. The key difference is for using attributes from one image (i.e. Intrinsic cues) or cooperation similar images (e.g. user annotations, depth map, or statistical information) to facilitate detecting salient objects in the image (i.e. Extrinsic cues).

Based on this survey and mentioned attributes most of the existing salient object detection approaches can be divided into three major categories, *block-based models with intrinsic cues*, *region-based models with intrinsic cues*, and *models with extrinsic cues*.

Block-based Models with Intrinsic Cues. These models detect salient objects based on blocks (i.e. pixels or patches) with only utilizing intrinsic cues. Their drawbacks are: they detect high contrast edges as a salient object instead of the real salient object, and if the size of blocks is large, the boundary of the salient object is not protected very well. To control these problems successfully, region based maps are considered as the new generation of researchers. Because the number of regions is much less than the number of blocks better features can be extracted from regions.

Region-based Models with Intrinsic Cues. In these models, the first input image is segmented into regions aligned with intensity edges and then regional saliency map is computed. Three types of region extraction methods are used for saliency computation (Graph-based segmentation algorithm, mean-shift algorithm, or clustering quantization). The first advantage of this method in comparison with block-based is that for improving these models, there are several choices like backgroundness, objectness, focusness and boundary connectivity. Besides, regions give more advanced cues (e.g. color histogram). Another advantage of using regions instead of blocks (i.e. pixels or patches) is for computational cost because each image has far fewer regions than pixels, computation of regional saliency would be less than producing full-resolution saliency maps. Despite these advantages, the new generation will be used extrinsic cues. Jiang et al. [19] proposed an approach based on multi-scale local region contrast, which calculates saliency values across multiple segmentations and combines these regional saliency values to get a pixel-wise saliency map.

Models with Extrinsic Cues. These models help salient object extraction in images and videos. These cues can be derived from the ground truth annotations of the training images, similar images, the video sequence, a set of input images containing the common salient objects, depth maps, or light field images. Borji et al. [18] concluded that the DRFI, which is presented by Jiang et al. [9], is an extrinsic cue model. This model had been only trained on a small subset of MSRA5K, and it still consistently outperforms other methods on all datasets.

Borji et al. [20] qualitatively and quantitatively compared 40 state-of-the-art models (28 salient object detections, 10 fixation predictions, 1 objectness, and 1 baseline) over 6 datasets for the purpose of benchmarking salient object detection and segmentation methods. In their comprehensive comparisons, DRFI, which is presented by Jiang et al. [9], discriminatively trains a regression model to predict region saliency according to a 93-dimensional feature vector. Also, it is a region-based approach, and in comparison with another block –based models, it always preserves the object boundary well. It uses a combination of Saliency Cut method with a sophisticated salient object detection model that has got the best segmentation results. Therefore, this salient method would be the best selection for salient region extraction.

Previous categorizations (*Block-based vs. Region-based analysis* and *intrinsic cues vs. Extrinsic cues*) were based on salient object detection. Borji et al. [18] mentioned that there exist some other research whose main research effort is not on the saliency map computation nor can it segment or localize salient objects directly with bounding boxes. They classified them as *Localization models, Segmentation models, Aggregation and Optimization models* and *Active models.*

Localization models. The output of these models is rectangles around the salient objects by converting the binary segmentations to bounding boxes. The most common approach is using a sliding window and classifying each of them as either a target or a background. For example, Lampert et al. [16] proposed an object localization method based on maximization of sub-images with branch and bound scheme, but their research cannot find two or more important objects in one picture. Another problem of using sliding windows occurred when the local image information is insufficient e.g. when the target is very small or highly occluded. In these cases, other parts of the picture will help us to classify the picture Murphy et al. [15].

Therefore, K. Murphy et al. presented a combination model of local and global (gist) features of the scene. This would be useful for solving the previous problem. They found that local features alone would cause a lot of false positives. Sometimes the scale estimation is incorrect as well. Also, they concluded that using global features can correct the estimation and decrease the ambiguity caused by only using local object detection methods. However, the basic idea of previous approaches that at least one salient object exists in the input image may not always behold as some background images that contain no salient objects at all. Wang et al. [21], investigated the problem of detecting the existence and the place of salient objects on thumbnail images using random forest learning approach.

Segmentation models. In these models separating the salient object from the background is the main approach. Kim and Grauman [22] proposed a region detection approach. Their method used dense local region detector to extract suitable features for object recognition and image matching. Having applied boundary-preserving local regions (BPLRs), they asserted that their method can find the connectivity of pixels, and it can save the object boundaries for foreground discovery and object classification. Wang et al. [23] presented a framework to segment the salient object by contextual cues usage automatically. Their method incorporates texture, luminance and color cues. Also, it measures the similarity between neighboring pixels and computes the edge probability map to label them as background/foreground. Recently, Jiang et al. [9] offered a saliency estimation as a regression problem and still their method consistently outperforms other methods on all datasets.

Aggregation and optimization models. These models try to combine M saliency maps and in order to form an accurate map to help the detection of salient objects. Borji et al. [20] proposed a standard saliency aggregation. Recently, Yan et al. [24] combined saliency maps based on the hierarchical segmentation to get a tree-structure graphical model from three layers of different scales. In this model, each node is related to a region. They concluded that hierarchical algorithms could select optimal weights for each region instead of global weighting superpixels.

Active models. These models combine two stages in one (the most salient object detection and segmentation). Recently, Borji [25] presented an active model, which can locate the salient object by finding the peak pixels of the fixation map. Then it segments the picture by superpixels. Their method can connect fixation prediction and salient object segmentation.

The next stage for object recognition is feature extraction. The first category of approaches used SIFT descriptor for object recognition. Later, Bay et al. [26] proposed SURF, which is a quicker SIFT. Liu et al. [27] suggested a fast algorithm for the computation of a dense set of SIFT descriptors. Dalal et al. [28] used the Histogram of Oriented Gradient (HOG) descriptor for pedestrian detection. One of the popular structures for adding object location is the hierarchical structures e.g. Oquab et al. [6] recommended a weakly supervised convolutional neural network (CNN) (a hierarchical convolutional structure) for object classification that only depends on image-level labels. Their method predicts the estimated place (x, y) of objects in the scene, not their bounding box. Also, spatial pyramids are considered as other popular hierarchical structures for adding location information to Bag of Visual Words image representation model e.g. Elfiky et al. [29] presented a framework to compact pyramid representation. Since all the features are not important, nowadays reducing the size of feature vectors or feature selection is one main technique for dimensionality reduction that involves identifying a subset of the most useful features. Therefore, identifying a subset of the most useful features is still another way for researchers. For this reason, recently Li et al. [30] suggested Clustering-Guided Sparse Structural Learning (CGSSL) algorithm, an unsupervised feature selection method, which mixes cluster analysis and structural analysis into one framework. While, other researchers try to uncover the hidden subspaces for the purpose of image understanding more recently, Li et al. [31] proposed Robust Structured Subspace Learning (RSSL) algorithm by mixing image understanding and feature learning into a joint learning framework. Their model used the visual geometric structure as well as the local and global structural consistencies over labels at the same time to reveal the important subspace robust to the outliers and noise.

It deserves mentioning that feature representation and codebook generation are considered as the last and the most important stage. This means that a proper feature representation considerably improves the performance of the semantic learning techniques. In the existing image retrieval methods both regional and global image representations are applied. Regional based feature extraction is the future development in this area.

It has to be mentioned that our proposed approach differs from existing algorithms on two points. In term of the object detection, a new saliency map method based on Jiang et al. [9] was used to distinguish the main rectangular parts of the picture. It is concluded that extracting SIFT features of these main rectangular parts improved the performance of existing methods. Because of removing the background from the main object, it got better results for difficult objects. However, in many state-of-the-art methods they can only decide whether an object is present in an image or not. Murphy [15] mentioned that they cannot distinguish the exact place of the object. While some previous approaches used pre-segmented or semi-supervised techniques to estimate the exact place of main objects like Russell et al. [32] (LabelMe) or PASCAL, the algorithm can automatically estimate the exact place of an object based on saliency map. Also, it does not matter how many times the object is repeated. SBBoVW model can find important objects properly.

3. Overview of the SBBoVW model toward difficult object recognition

This paper proposes a Salient Based Bag of Visual Word (SBBoVW) model for difficult object recognition. For the sake of simplicity this model is divided into two parts, **Fig. 1** and **Fig. 2**.

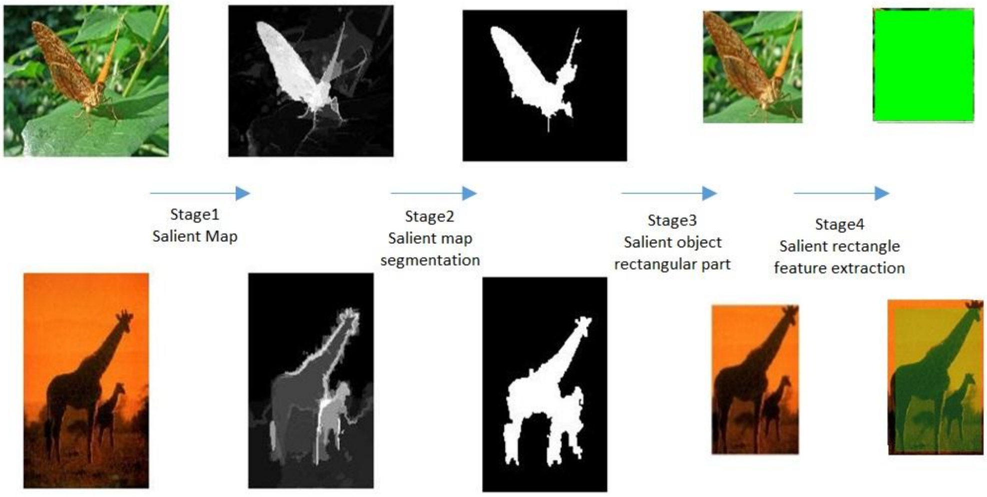
Stage1
Salient Map
Stage2
Salient map segmentation
Stage3
Salient object rectangular part
Stage4
Salient rectangle feature extraction

Fig. 1. Salient rectangle features extraction steps

Steps for object recognition by SBBoVW are:

First part depicted in **Fig. 1**.

- Saliency map computation
- Saliency rectangular parts computation
- SIFT feature of saliency rectangular parts and normal pictures

Second part depicted in **Fig. 2**.

- Codebook creation based on K-Means classification
- Visual words spatial histograms quantization based on KD-trees
- Testing of the model for unseen pictures
- Object place finding

In the first step, a saliency map is computed by fusing the saliency maps across multiple levels of segmentations based on Jiang et al. [9]. In detail, saliency map creation runs in three steps. Multi-level segmentation is the first step. Region saliency computation is the next and the key step of saliency map creation. Finally, a saliency map is computed by merging the saliency maps across multiple levels of segmentations. The second step is called salient map segmentation. In fact, applying this step generates a salient binary mask. With this mask, we can find salient objects rectangle parts. Algorithm 1 describes this phase of the method in details. The third step of **Fig. 1** (i.e. Salient object rectangular part) shows the results of the algorithm. The last step is salient rectangle feature extraction, which is depicted in the last step of this figure. Although, the resulting picture completely covered by a green or yellow mask, those are dense SIFT features (DSIFT).

The key step of SBBoVW model lies in the second step, SBBoVW training model. This step is depicted as a sequential diagram in **Fig. 2**. As illustrated in this figure, the model combines several feature vectors (SIFT feature) in order to achieve an appropriate codebook. As it can be seen, the SBBoVW training process is performed through the following steps:

The first one is dense SIFT feature extraction of the test original pictures and saliency rectangular parts of them (which are extracted from the previous model **Fig. 1**).The codebook is created in the next step by K-Means classification. In the third step, descriptors are quantized based on KD-trees to get the visual words. Then spatial histograms of visual words that are considered as a joint distribution of appearance and location of the visual words in an image are extracted.

```
Algorithm 1  SalObjRect(image)
saliency Map=salmapcreation(image)
segmentation Result = Segmentation(image, saliency Map);
s=region props(segmentation Result,'Bounding Box','area');
image3=ones(size(segmentation Result))
mask1 = false(size(image,1),size(image,2));
mask=mask1;
SalObjRect= image;
for k=1:length(s)
    starty = s(k).xleftTopcorner
    stopy = starty+s(k).xrightDowncorner-1;
    startx = s(k). yleftTopcorner ;
    stopx = startx+s(k). yrightDowncorner -1;
    mask1(startx:stopx,starty:stopy) = true;
    mask(mask1)=true;
end
SalObjRect (~mask) = 0;
```

Testing the model is the third step. In this step, spatial histograms of visual words of testing pictures are compared with spatial histograms of visual words of test pictures using SVM Chi-square. Afterward, the appropriate concept names are extracted from test images. This model is concluded in

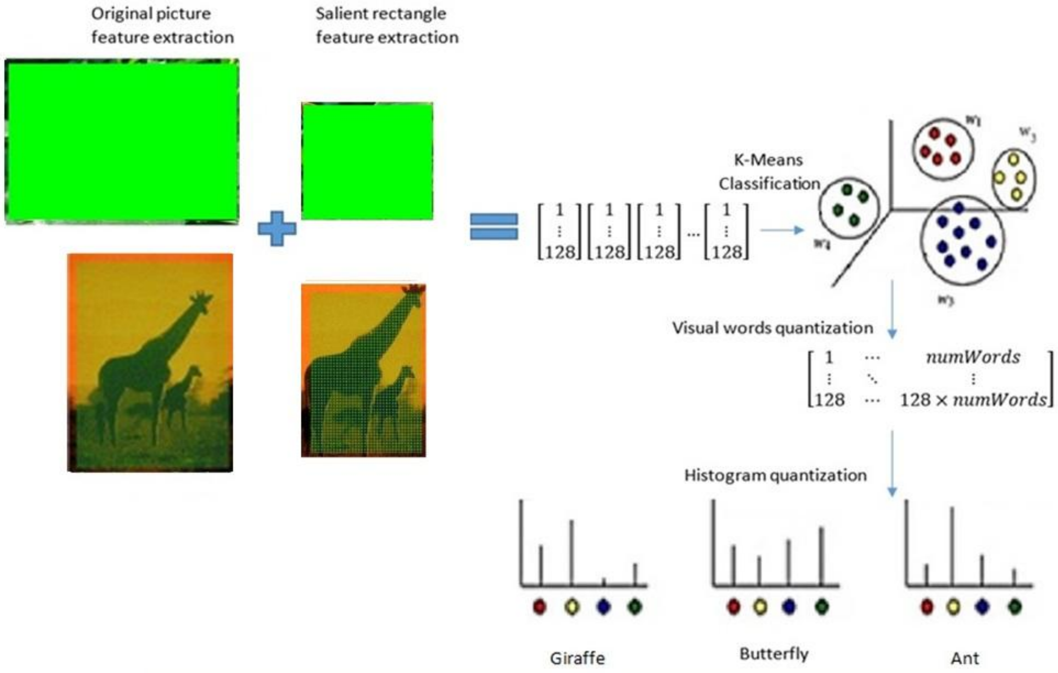

Fig. 2. SBBoVW training model.

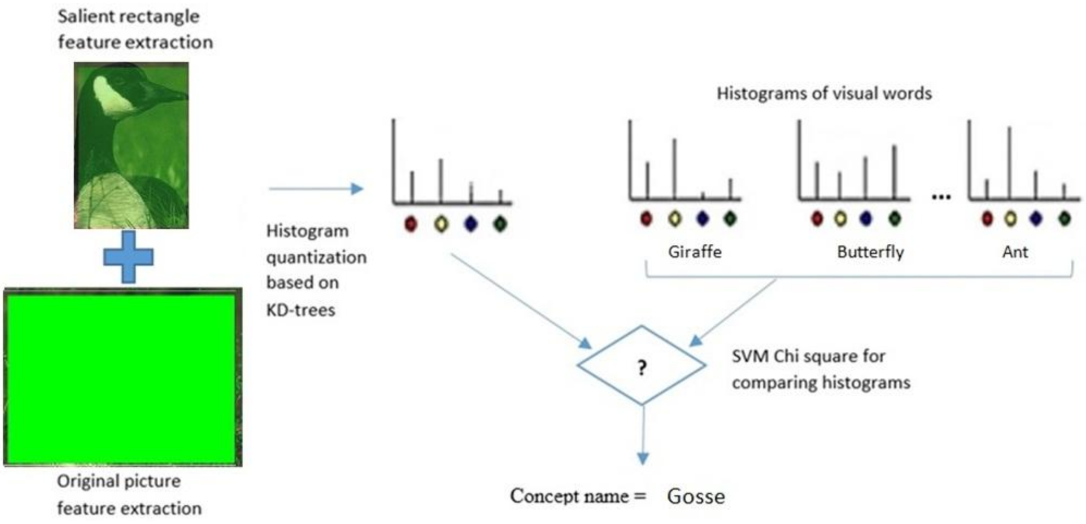
Salient rectangle feature extraction
Histograms of visual words
Histogram quantization based on KD-trees
Giraffe
Butterfly
Ant
?
SVM Chi square for comparing histograms
Original picture feature extraction
Concept name = Gosse

Fig. 3. SBBoVW testing model.

At last, a new algorithm is conducted to find object place automatically. Fig. 4 shows four examples of the algorithm results. Algorithm 2 describes the new algorithm in detail.

```
Algorithm 2  object_place (image)
saliency Map=salmapcreation(image)
 segmentation Result = Segmentation(image, saliency Map);
 se = strel('disk',10);
 bw2 = imdilate(segmentation Result, se);          % dilates the binary mask
 s=region props(bw2,'BoundingBox','area');
 imshow(image);
 hold on
 for k=1:length(s)
       rectangle('position',s(k).BoundingBox,'edgecolor','y','linewidth',2);
 end
 hold off;
```

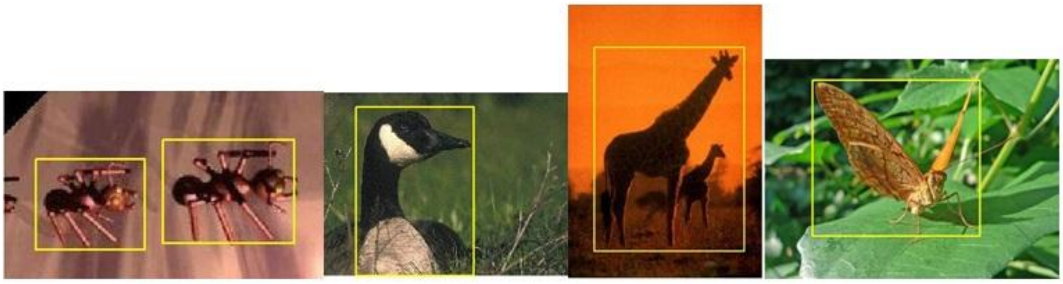

Fig. 4. The output of the algorithm for finding object place.

After that, the building of our object recognition and place is fully automatically performed, i.e. without any human intervention.

4. Experiments

As it was mentioned earlier, this paper aims at investigating the potentiality and accuracy of a Salient Based Bag of Visual Word model (SBBoVW) toward difficult object recognition in image retrieval. Evaluations are performed on Fei-Fei et al. (Caltech-101) [7], in addition to the animal subset of Griffin et al.'s (Caltech-256) [8] dataset.

4.1 Caltech-101 dataset

This dataset has around 40 to 800 images per category. It contains a total of 9146 images, split between 101 distinct objects (including faces, watches, ants, pianos, etc.) and a background category (for a total of 102 categories). As suggested by Wang et al. [4] and also by many other researchers Griffin et al. [8] and Zhang et al. [33], the dataset is partitioned into 5, 10,…, 30 training images per class and no more than 50 testing images per class. To make a comparison between this method and Vedaldi and Fulkerson's [3] the same train and test images opted. Well-known formulas measured the accuracy: Precision, Recall and Accuracy Tousch et al. [34], Chiang [35], Fakhari and Moghadam [36], Lee et al. [37] and classification rate (mean($\frac{No.of\ accurate results}{No.oftotal\ test\ pictures}$)), used in all of the compared states of arts e.g. Berg [33], Vedaldi and Fulkerson [3].

4.2 Caltech-256 dataset

From Caltech-256 dataset, 20 different animals (bear, butterfly, camel, house-fly, frog, giraffe, goose, gorilla, horse, hummingbird, ibis, iguana, octopus, ostrich, owl, penguin, starfish, swan, dog, zebra) were selected from different environments (lake, desert, sea, sand, jungle, bushy, etc.). The common training setup (15, 30, 45 and 60 training images for each class) was followed Wang et al. [4]. Also, this amount is not more than 50 testing images per class. To compare this method with the basic BoVW model, the same train, and test images were chosen. The number of extracted codewords is 1500. The accuracy is measured or computed by well-known formulas: Precision, Recall and Accuracy and classification rate

This method is expandable, all we need to do is to separate the folder of the new concept and change its name. Then all the steps can be automatically done by the algorithm.

Essential software for running the program is Matlab 2013a/ 2014a.

5. Discussion

As illustrated in **Table 1** and **Table 2**, the method for object recognition provides better results than the other 12 state of arts results on Caltech-101 and Caltech-256 datasets, with average accuracies of 60.78%, 64.38%, 68.14%, 70.39%, 74.35%, and 76.72% on Caltech-101. The comparison under the same train and test images shows the method gains of +2.94%, -1.21%, +2.45% , +1.57%, +1.47%, and +0.98% according to Vedaldi and gains of +15% , +10.63%, +12.5%, +16.37%, and 13.67% according to the previous experiment (BoVW method). In **Table 1** the improvements of SBBoVW method in each running (5, 10, …, 25, 30 for Caltech 101 and 15, 30, 40, 45, 60 for a subset of Caltech256) as well as other states of arts can be clearly observed.

These results approve the effectiveness of the proposed approach, and the importance of salient region information to improve object recognition.

Table 1. Classification rate comparison based on percentage of $\frac{No.of\ accurate\ results}{No.of\ total\ test\ pictures}$ in Caltech-101 dataset and different methods

No. of Training images	**5**	**10**	**15**	**20**	**25**	**30**
Lazebnik 2006 [38]			56.4			64.6
Zhang 2006 [33]	46.6	55.8	59.1	62		66.2
Griffin 2007 [8]	44.2	54.5	59	63.3	65.8	67.6
Boiman 2008 [39]			65			70.4
jain 2008 [40]			61			69.10
Gemert 2008 [41]						64.16
Yang 2009 [42]			67			73.20
Vedaldi 2010 [3]	57.84	**65.59**	65.69	68.82	72.88	75.74
Wang 2010 [4]	51.15	59.77	65.43	67.74	70.16	73.44
Bilen 2014 [43]						75.31
Biagio 2014 [44]			66.57			
Maeda 2014 [45]			67.6			73.9
SBBoVW (Ours)	**60.78**	64.38	**68.14**	**70.39**	**74.35**	**76.72**

Table 2. The comparison of classification rate between BoVW (the previous experiment) and SBBoVW (the new model) in animal subset of Caltech-256

No. of Training images	**15**	**30**	**40**	**45**	**60**
BoVW	27.5	36.25	38.5	38.63	36.66
SBBoVW (Ours)	**42.5**	**46.88**	**51**	**55**	**50.33**

In **Fig. 5,** the present researchers compare their framework accuracy for object recognition with their previous experiment (BoVW). As it can be seen in this figure, the SBBoVW framework for object recognition outperforms on 16 classes in comparison with the other ones.

In order to get accurate results, similar experimental setup was taken into consideration. The same training/testing sets from the Caltech-101 and Caltech-256 dataset and the same features of images were selected. Therefore, it is clear that the proposed framework for object recognition makes a significant improvement in the object recognition accuracy.

In **Fig. 5** and **Fig. 7**, the framework accuracy for object recognition is compared with the following methods: BoVW and Vedaldi and Fulkerson [3] (VLFeat). After 5 times of program running on the same train and test images of an animal subset of Caltech-256 dataset, it is observed that, the proposed method has improved 16 concepts (bear, butterfly, camel, frog, giraffe, goose, gorilla, horse, house-fly, ibis, octopus, ostrich, penguin, starfish, swan, zebra) and have not improved 4 concepts (dog, hummingbird, iguana and owl) out of. For Caltech-101 again we compared our method 6 times under the same train and test images, it is found that it has improved 79 concepts out of 102 objects. This means, 23 concepts (bonsai, brontosaurus, cellphone, chair, crab, crocodile, crocodile_head, cup20, electric_guitar, ewer, gerenuk, headphone, helicopter, mayfly, menorah, nautilus, panda, pigeon, platypus, scorpion, stegosaurus, strawberry, wrench) have not been improved. **Fig. 6** and **Fig. 8** show their precision results respectively.

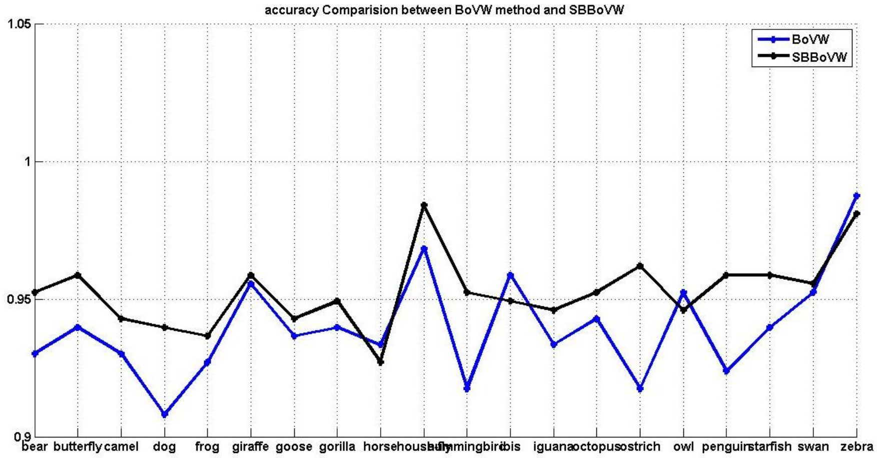

Fig. 5. Comparison of the accuracy of the previous and proposed model on the subset of Caltech-256.

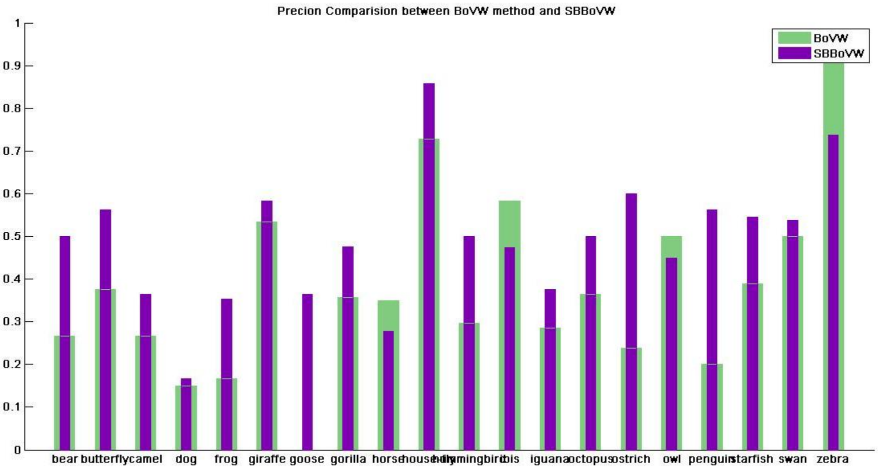

Fig. 6. precision comparison of BoVW and proposed model on a subset of Caltech-256.

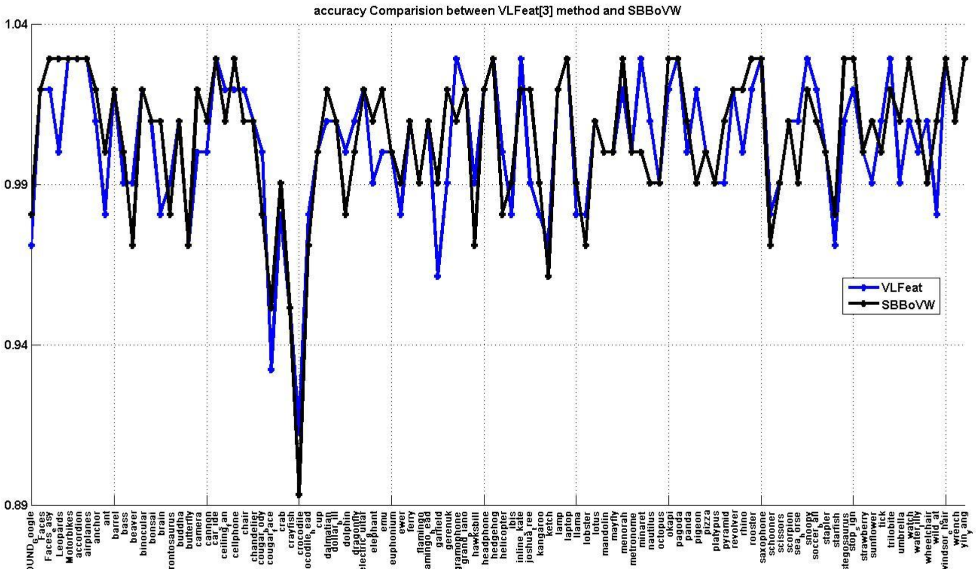

Fig. 7. Accuracy comparison of VLFeat and the proposed model on Caltech101**.**

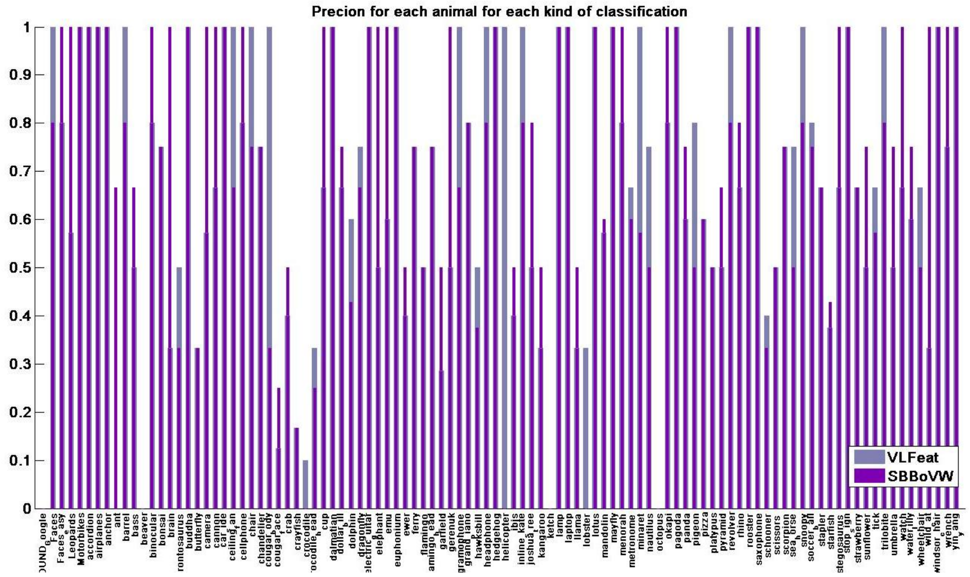

Fig. 8. precision comparison of the VLFeat and the proposed model on Caltech101.

In summary, SBBoVW got better results, especially for difficult concepts. In Caltech-101, 79 concepts were improved (i.e. BACKGROUND_Google, Faces, Faces_easy, Leopards, Motorbikes, accordion, airplanes, anchor, ant, barrel, bass, beaver, binocular, brain, Buddha, butterfly, camera, cannon, car_side, ceiling_fan, chandelier, cougar_body, cougar_face, crayfish, Dalmatian, dollar_bill, dolphin, dragonfly, elephant, emu, euphonium, ferry, flamingo, flamingo_head, Garfield, gramophone, grand_piano, hawksbill, hedgehog, ibis, inline_skate, joshua_tree, kangaroo, ketch, lamp, laptop, llama, lobster, lotus, mandolin, metronome, minaret, octopus, okapi, pagoda, pizza, pyramid, revolver, rhino, rooster, saxophone, schooner, scissors, sea_horse, snoopy, soccer_ball, stapler, starfish, stop_sign, sunflower, tick, trilobite ,umbrella, watch, water_lilly, wheelchair, wild_cat, windsor_chair, yin_yang).

Also, in the subset of Caltech-256, 16 animals were improved (i.e. bear, butterfly, camel, frog, giraffe, goose, gorilla, horse, house-fly, ibis, octopus, ostrich, penguin, starfish, swan, zebra). This clearly proves that SBBoVW outperforms other state-of-the-arts, especially to uncover and recognize hidden concepts (i.e. most of the animals are as the same as their environment because nature protects them against enemies).

Because of this problem of the traditional BoVW model, we cannot collect visual words on the basis of their locations in the picture. Moreover, all the visual words are collected and treated the same as each other even if they are from the important part or background of the picture. A Salient Based Bag of Visual Word (SBBoVW) model is presented for difficult object recognition and object location. In fact, SBBoVW model collected visual words of the whole and salient part of the picture to overcome the mentioned problem.

In General, SBBoVW is better than the other state of arts because of the following reasons:

Features of the pictures are duplicated based on their locations, i.e. if the feature is located in the salient part of the picture, SBBoVW repeats it. For more explanation, our model decreases the number of background features and increases the number of the foreground or important part of the picture. This location persisting improves the performance of our model.

As our experiment, feature extraction of the salient part definitely will improve the accuracy. If we look closer to **Table 1**, we can easily understand that in comparison with Maeda's method [45] which is a two-stage classification framework for combining parametric approaches (BoF+SVM) and non-parametric approaches (NN), our model improved +2.82 and 0.54 for 15 and 30 training images respectively. In order to have a more detailed comparison, Vedaldi's [3] PHOW method selected which has image categorization with a dense set of multi-scale SIFT descriptors, spatial histograms of visual words, Chi2 SVM and KD-trees to add the location of the visual words in an image, and homogeneous kernel map.Because more features of salient part of the picture, are added to our algorithm, after conducting the experiment, SBBoVW outperforms +2.94, -1.21, +2.45, +1.57, +1.47, +0.98 for 5, 15, 20, 25 and 30 numbers of training images. Only it did not work well for 10 training images because the classification rate decreases to -1.21. Also, for the rest of the Compared Literature SBBoVW classification rate got better results.

Another interesting achievement about this technique is that researchers of the present study trained and tested their model on the salient rectangular part of the picture and found that the accuracy was decreased dramatically. This means that removing the whole information of background is not satisfactory for total accuracy, and some information of environment is required in order to distinguish the objects.

6. Conclusion

In this paper, the potential usage of Salient map in difficult object recognition and estimation of important object place in pictures was investigated. It is found that feature extraction of the salient rectangular part of the picture as well as the whole picture will affect the results precisely. Besides, the only usage of features of the salient rectangular part, cannot improve the accuracy. After implementation, the model (SBBoVW) shows more accurate results than the other 13 state-of-art models. However, this model still needs improvements for really hidden object in darks such as owl or object with the same patterns of background like an iguana. In future, with the help of some segmentation methods and improvements in salient rectangular parts of pictures difficult objects can be recognized more accurately. Also, multi-object pictures such as VOC-7 dataset can be used in order to improve the proposed model.

Acknowledgments

This article was kindly supported by Ministry of Higher Education (Malaysia) under Fundamental Research Grant Scheme (FRGS).

References

[1] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray, and D. Maupertuis, "Visual Categorization with Bags of Keypoints," *Work. Stat. Learn. Comput. vision, ECCV*, vol. 1, pp. 1–2, 2004. Article (Ref Link)

[2] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. of Seventh IEEE Int. Conf. Comput. Vis.*, vol. 2, pp. 1150–1157, 1999. Article (CrossRef Link)

[3] A. Vedaldi and B. Fulkerson, “VLFeat - An open and portable library of computer vision algorithms,” *Design*, vol. 3, no. 1, pp. 1–4, 2010. Article (CrossRef Link)

[4] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, “Locality-constrained linear coding for image classification,” in *Proc. of IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3360–3367, 2010. Article (CrossRef Link)

[5] C.-C. Chang and C.-J. Lin, “Libsvm,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011. Article (CrossRef Link)

[6] M. Oquab, I. P. France, B. Leon ; L. Ivan ,S. Josef, “Is object localization for free ? – Weakly-supervised learning with convolutional neural networks,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 685-694. 2015. Article (CrossRef Link)

[7] L. Fei-Fei, R. Fergus, and P. Perona, “Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories,” *Comput. Vis. Image Underst.*, vol. 106, no. 1, pp. 59–70, Apr. 2007. Article (CrossRef Link)

[8] P. Griffin, G. Holub, AD. Perona, “Caltech-256 object category dataset,” *California Institute of Technology*, 2007. Article (Ref Link)

[9] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, “Salient object detection: A discriminative regional feature integration approach,” in *Proc. of IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 2083–2090, Jun. 2013. Article (CrossRef Link)

[10] H. Bannour and C. Hudelot, “Building and using fuzzy multimedia ontologies for semantic image annotation,” *Multimed. Tools Appl.*, pp. 2107–2141, May 2013. Article (CrossRef Link)

[11] M.-U. Kim and K. Yoon, “Performance evaluation of large-scale object recognition system using bag-of-visual words model,” *Multimed. Tools Appl.*, Jun. 2014. Article (CrossRef Link)

[12] S. Zhong, Y. Liu, Y. Liu, and F. Chung, “Region level annotation by fuzzy based contextual cueing label propagation,” *Multimed. Tools Appl.*, vol. 70, no. 2, pp. 625–645, Jan. 2012. Article (CrossRef Link)

[13] X. Long, H. Lu, and W. Li, “Image classification based on nearest neighbor basis vectors,” *Multimed. Tools Appl.*, vol. 71, no. 3, pp. 1559–1576, Nov. 2012. Article (CrossRef Link)

[14] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, “Baby talk: Understanding and generating simple image descriptions,” *Cvpr 2011*, pp. 1601–1608, Jun. 2011. Article (CrossRef Link)

[15] K. Murphy, A. Torralba, D. Eaton, and W. Freeman, “Object Detection and Localization Using Local and Global Features,” pp. 382–400, 2006. Article (CrossRef Link)

[16] C. H. Lampert, M. B. Blaschko, and T. Hofmann, “Beyond sliding windows: Object localization by efficient subwindow search,” in *Proc. of 2008 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1–8, Jun. 2008.Article (CrossRef Link)

[17] V. Dey, Y. Zhang, M. Zhong, and G. Engineering, “A REVIEW ON IMAGE SEGMENTATION TECHNIQUES WITH REMOTE SENSING PERSPECTIVE,” *na*, vol. XXXVIII, pp. 31–42, 2010. Article (Ref Link)

[18] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, “Salient Object Detection: A Survey,” *arXiv preprint arXiv:1411.5878*, pp. 1–26, Nov. 2014. Article (Ref Link)

[19] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, “Automatic Salient Object Segmentation Based on Context and Shape Prior,” BMVC, Vol. 6. No. 7, pp. 110.1--110.12,. 2011. Article (CrossRef Link)

[20] A. Borji, D. N. Sihite, and L. Itti, “Salient Object Detection : A Benchmark,” in *Proc. of the 12th European Conference on Computer Vision - Volume Part II, Springer-Verlag*, pp. 414–429, 2012. Article (CrossRef Link)

[21] P. Wang, J. Wang, G. Zeng, J. Feng, H. Zha, and S. Li, “Salient object detection for searched web images via global saliency,” in *Proc. of 2012 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3194–3201, Jun. 2012. Article (CrossRef Link)

[22] J. Kim and K. Grauman, “Boundary preserving dense local regions,” *Cvpr 2011*, pp. 1553–1560, Jun. 2011. Article (CrossRef Link)

[23] G. H. Le Wang, Jianru Xue, Nanning Zheng, "Automatic salient object extraction with contextual cue," in *Proc. of 2011 Int. Conf. Comput. Vis.*, pp. 105–112, Nov. 2011. Article (CrossRef Link)

[24] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical Saliency Detection," in *Proc. of 2013 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1155–1162, Jun. 2013. Article (CrossRef Link)

[25] A. Borji, "What is a salient object? A dataset and a baseline model for salient object detection," *Image Processing, IEEE Transactions on*, Vol. 24(2), pp.742-756, 2015. Article (CrossRef Link)

[26] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF : Speeded Up Robust Features," in *Computer vision–ECCV 2006, Springer Berlin Heidelberg*, pp. 404–417, 2006. Article (CrossRef Link)

[27] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, "SIFT Flow : Dense Correspondence across Different Scenes," in *Computer Vision–ECCV 2008, Springer Berlin Heidelberg*, Vol. 1, no. 1, pp. 28–42, 2008. Article (CrossRef Link)

[28] N. Dalal, B. Triggs, and D. Europe, "Histograms of Oriented Gradients for Human Detection," in *Proc. of Computer Vision and Pattern Recognition, 2005. CVPR 2005, IEEE Computer Society Conference on*, vol. 1, pp. 886-893, 2005. Article (CrossRef Link)

[29] N. M. Elfiky, F. Shahbaz Khan, J. van de Weijer, and J. Gonzàlez, "Discriminative compact pyramids for object and scene recognition," *Pattern Recognit.*, vol. 45, no. 4, pp. 1627–1636, Apr. 2012. Article (CrossRef Link)

[30] Z. Li, J. Liu, Y. Yang, X. Zhou, and S. Member, "Clustering-Guided Sparse Structural Learning for Unsupervised Feature Selection," *Knowledge and Data Engineering, IEEE Transactions on* Vol. 26, no. 9, pp. 2138–2150, 2014. Article (CrossRef Link)

[31] Z. Li, J. Liu, J. Tang, and H. Lu, "Robust Structured Subspace Learning for Data Representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. X, no. X, pp. 1–1, 2015. Article (CrossRef Link)

[32] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A Database and Web-Based Tool for Image Annotation," *Int. J. Comput. Vis.*, Vol. 77, no. 1–3, pp. 157–173, Oct. 2007. Article (CrossRef Link)

[33] A. C. Berg, "SVM-KNN : Discriminative Nearest Neighbor Classification for Visual Category," in *Proc. of Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference* on. Vol. 2, pp. 2126-2136, 2006. Article (CrossRef Link)

[34] A. M. Tousch, S. Herbin, and J. Y. Audibert, "Semantic hierarchies for image annotation: A survey," *Pattern Recognit.*, vol. 45, no. 1, pp. 333–345, Jan. 2012. Article (CrossRef Link)

[35] C.-C. Chiang, "Interactive tool for image annotation using a semi-supervised and hierarchical approach," *Comput. Stand. Interfaces*, vol. 35, no. 1, pp. 50–58, Jan. 2013. Article (CrossRef Link)

[36] A. Fakhari and A. M. E. Moghadam, "Combination of classification and regression in decision tree for multi-labeling image annotation and retrieval," *Appl. Soft Comput.*, vol. 13, no. 2, pp. 1292–1302, Feb. 2013. Article (CrossRef Link)

[37] C.-H. Lee, H.-C. Yang, and S.-H. Wang, "An image annotation approach using location references to enhance geographic knowledge discovery," *Expert Syst. Appl.*, vol. 38, no. 11, pp. 13792–13802, May 2011. Article (CrossRef Link)

[38] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *Proc. of 2006 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Vol. 2*, vol. 2, pp. 2169–2178, 2006. Article (CrossRef Link)

[39] O. Boiman, E. Shechtman, and M. Irani, "In Defense of Nearest-Neighbor Based Image Classification," in *Proc. of Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1-8. IEEE, 2008 Article (CrossRef Link)

[40] P. Jain, B. Kulis, and K. Grauman, "Fast Image Search for Learned Metrics," in *Proc. of Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1-8. IEEE, 2008. Article (CrossRef Link)

[41] and A. W. S. Jan C. Gemert, Jan-Mark Geusebroek, Cor J. Veenman, "Kernel Codebooks for Scene Categorization," in *Proc. of the 10th European Conference on Computer Vision: Part III (ECCV '08)*, 2008, p. 7_52. Article (CrossRef Link)

[42] and T. H. J. Yang, K. Yu, Y. Gong, "Linear spatial pyramid matching using sparse coding for image classification," in *CVPR '09*, 2009. Article (CrossRef Link)

[43] H. Bilen, V. P. Namboodiri, and L. J. Van Gool, "Object and Action Classification with Latent Window Parameters," *Int. J. Comput. Vis.*, vol. 106, no. 3, pp. 237–251, Aug. 2013. Article (CrossRef Link)

[44] M. S. Biagio, L. Bazzani, M. Cristani, and V. Murino, "Weighted bag of visual words for object recognition," in *Proc. of Image Processing (ICIP), 2014 IEEE International Conference on*, pp. 2734–2738, 2014.Article (CrossRef Link)

[45] C. Engineering, "MULTI-STAGE OBJECT CLASSIFICATION FEATURING CONFIDENCE ANALYSIS OF CLASSIFIER AND INCLINED LOCAL NAIVE BAYES NEAREST NEIGHBOR," in *Proc. of Image Processing (ICIP), 2014 IEEE International Conference on*, pp. 5177–5181, 2014. Article (CrossRef Link)

Leila Mansourian is currently a Ph.D. student in the Department of Multimedia, Faculty of Computer Science and Information Technology at Universiti Putra Malaysia. She received her M.S. degree in Artificial Intelligence from Islamic Azad University, Science and Research Branch, Tehran, in 2007, and Bac. in Computer Science from Islamic Azad University of Mashhad, in 2002. Her research interests include Multimedia Systems and Applications, Multimedia Information Retrieval, and Pattern Recognition.

Muhamad Taufik Abdullah is an Associate Professor at the University Putra Malaysia. He received his BS degree in Computer Science from the Universiti Pertanian Malaysia in 1990, his MS degree in Computer Science from the Universiti Teknologi Malaysia in 1992, and his Ph.D. degree in Information Retrieval from the Universiti Putra Malaysia in 2006. His current research interests include information retrieval, natural language processing, and multimedia computing.

Lili Nurliyana Abdullah is currently an Assistant Professor in the Department of Multimedia, Faculty of Computer Science and Information Technology at Universiti Putra Malaysia. She received her Ph.D. degree in Information Science from Universiti Kebangsaan Malaysia, in 2007, M.S. degree in Engineering (Telematics) from University of Sheffield, United Kingdom, in 1996, and Bac. in Computer Science from UPM in 1990. Her research interests include multimedia system, video processing and retrieval, computer modeling and animation, image processing and computer games.

Azreen Azman is a Senior Lecturer at the Universiti Putra Malaysia. He received a Diploma in Software Engineering from the Institute of Telecommunication and Information Technology in 1997. Immediately, he was accepted directly for the second year in Multimedia University, Malaysia to study Bachelor of Information Technology majoring in Information Systems Engineering. He completed his bachelor degree in 1999. After serving in the industry for a few years, he enrolled for a Ph.D. in January 2003, studying Computing Science specializing in Information Retrieval at the University of Glasgow, Scotland and completed his study in September 2007. His current research interests include information retrieval, text mining, opinion mining and semantic technology.

Mas Rina Mustaffa is currently a Senior Lecturer at the Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. She received her BCompSc degree (Multimedia) and MSc degree (Multimedia Systems) in 2003 and 2006 respectively. She received her Ph.D. for studies in Content-based Image Retrieval (CBIR) in 2012. All of the three degrees are from Universiti Putra Malaysia (UPM), Serdang, Selangor, Malaysia. She has authored several publications in various journals and proceedings and presented at many conferences. She also has been actively involved in several international conferences as technical program committee. She is a member of IEEE, ACM, Malaysian Society of Information Retrieval and Knowledge Management (PECAMP), and International Association of Computer Science and Information Technology (IACSIT). Her primary research interests are multimedia systems and applications, Content-Based Image Retrieval (CBIR), image processing, pattern recognition, and interactive multimedia.