J Inf Process Syst, Vol.14, No.5, pp.1176~1194, October 2018
https://doi.org/10.3745/JIPS.02.0095

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

# A Hybrid Proposed Framework for Object Detection and Classification

Muhammad Aamir*, Yi-Fei Pu*, Ziaur Rahman*, Waheed Ahmed Abro**,
Hamad Naeem*, Farhan Ullah***, and Aymen Mudheher Badr*

**Abstract**

The object classification using the images' contents is a big challenge in computer vision. The superpixels' information can be used to detect and classify objects in an image based on locations. In this paper, we proposed a methodology to detect and classify the image's pixels' locations using enhanced bag of words (BOW). It calculates the initial positions of each segment of an image using superpixels and then ranks it according to the region score. Further, this information is used to extract local and global features using a hybrid approach of Scale Invariant Feature Transform (SIFT) and GIST, respectively. To enhance the classification accuracy, the feature fusion technique is applied to combine local and global features vectors through weight parameter. The support vector machine classifier is a supervised algorithm is used for classification in order to analyze the proposed methodology. The Pascal Visual Object Classes Challenge 2007 (VOC2007) dataset is used in the experiment to test the results. The proposed approach gave the results in high-quality class for independent objects' locations with a mean average best overlap (MABO) of 0.833 at 1,500 locations resulting in a better detection rate. The results are compared with previous approaches and it is proved that it gave the better classification results for the non-rigid classes.

**Keywords**

Image Proposals, Feature Extraction, Object Classification, Object Detection, Segmentation

## 1. Introduction

Humans use their eyes and brains to visually sense the world. The computer systems cannot sense due to different degrees of viewpoint variations, illumination, scale, deformation and high intraclass variations. It is the first step to permit machines to recognize object, which is the foundation of the visual world. The computer vision is a field of computer science that works on enabling computers to sense and react to real-world visual media. It includes the growth of a theoretical and algorithmic basis to attain automatic visual understanding. It is mainly used to obtain dimensional real-world data and then, to process it into computer understandable decisions. The recent years have witnessed a rapid evolution in computer vision in a wide variety of real-world applications, i.e., automatic face recognition, optical character recognition, automated medical imagining, motion recognition, augmented reality, autonomous cars, domestic/service robots, image restoration, object tracking, and object detection [1,2].


Manuscript received November 17, 2017; first revision February 12, 2018; second revision July 20, 2018; accepted August 25, 2018.
**Corresponding Author:** Muhammad Aamir (aamirshaikh86@hotmail.com)
* College of Computer Science, Sichuan University, Chengdu, China (aamirshaikh86@hotmail.com, puyifei_007@163.com, {ziaurrahman167, hamadnaeemh, aemn_1978}@yahoo.com)
** School of Computer Science and Engineering, Southeast University, Nanjing, China (engr.waheedabro@gmail.com)
***COMSATS University Islamabad - Sahiwal Campus, Punjab, Pakistan (farhankhan.cs@yahoo.com)

Currently, the object detection and classification in an image are the most significant problems in computer vision and image processing. Many users and hackers can access a huge amount of visual information using the internet. It became more challenging because of the various viewpoint variations, i.e., size, angle, perspective, occlusion, and illumination. Current research is progressing in both directions, with numerous different techniques being proposed to achieve state-of-the-art detection and classification performance. The fast detector for a key point detection and binary local features descriptor for key point description being used. The Fast Library for Approximated Nearest Neighbors (FLANN) is applied to match the tested an actual image in the dataset. Further, a homograph matrix is estimated from corresponding pairs by using the optimized random sample consensus (ORSA) algorithm. Next, the significant color feature is used to calculate the global color histogram since it reflects the main content of the primitive image and also ignores noises [3]. Object classification is an important task in computer vision. It is the process of tagging objects into predefined and semantically significant classes using trained datasets. They used tensor features with Scale Invariant Feature Transform (SIFT) to improve the accuracy of the problem [4].

For covering a diverse set of regions, different kinds of grouping strategies and color spaces are used which produced good recall results. Selective search [5] grouping technique is used for object recognition and detection based on the hierarchical segmentation using super-pixels. But it has no scoring mechanism on the locations of the pixels. The edge-box [6] is the window-based approach used to track the location of an object in an image. It produced object location directly from the edges of an image. The initial edges are computed using detectors and then combined into a group of eight edges. It used a sliding window search over a scale to generate a candidate box and then scores each box using a huge number of locations. In order to achieve better detection and classification performance, there should be a joint improvement to speed in both detection and classification task. In the last decade, classification accuracy has been increased by using bag of words (BOW) algorithm [7], which was initially introduced in text analysis. Currently, the BOW method can also be used in image processing through SIFT feature extraction algorithm. The SIFT uses the critical points of an image by converting into the features descriptor and then performs the K-means clustering algorithm to get a cluster of features. Finally, the images are classified by an support vector machine (SVM) classifier [8]. The SVM is a supervised learning algorithm SVM, first proposed by Cortes and Vapnik [9]. The first version was developed for two-class classification problem. Then, it has been extended for multi-class classification problem and regression. The SVM is used to find the linear splitting hyperplane with the best margin using kernel functions. The ensemble support vector machine (ESVM) classifier is used to extract static length feature vector from self-organizing map (SOM) in given input [10,11].

The SIFT algorithm is mostly used in an extensive variety of applications such as object detection, object tracking, 3D modeling and successfully applied in the field of medicine. The author estimated the descriptor on numerous nodule morphology classification problems. Then, compare it with state-of-the-art approaches for 3D shape in medical imaging and computer vision [12]. Furthermore, a classification method based on the multi-level brain partitions, extracted SIFT features from the brain as the basic visual words to increase the classification performance [13]. Moreover, some extended versions of SIFT have been developed such as PCA-SIFT, GLOH, and SURF [14,15]. These methods provided dimension reduction and other functionalities such as scale and rotation change, blur change, illumination change, and affine change. However, since conventional SIFT algorithms are not efficient

at extracting the features from the noisy image, Tahir et al. [16] proposed a simple and useful approach that uses GIST feature vector. GIST provides a global feature representation of an image. However, features cannot distinguish the foreground from the background of an image; thus, big data images cannot be classified.

In this paper, we propose a new hybrid object detection and classification technique. Firstly, selective search [5] proposal generation scheme and edge-box [6] score criteria combine to increase the detection performance with significantly higher rates. Secondly, a feature fusion technique which combines GIST with SIFT Feature Vector Through Weight is formed to improve accuracy and reduce the computational complexity of the traditional BOW paradigm. Lastly, the VLFeat linear SVM classifier is used for performing classification. Results compared with existing approaches show our new proposed combined technique is effective for detecting and classifying an object accurately and timely. The main contributions of the paper are as:

1. High-Quality object locations with less number of candidate box
2. Rank the proposal according to region score—which is defined as a number of contours wholly enclosed in the proposed region, passing only the top object proposal for the post-classification.
3. Feature extraction on generated proposals/locations by combining both local and global features

This paper is organized as follows: Section 2 discusses the existing approaches to object detection and classification. In Section 3, we propose a hybrid proposed framework for object detection and classification In Section 4, we analyze and compare the existing and proposed schemes in terms of detection and classification accuracy. Section 5 concludes this paper. Section 6 discusses future work.

## 2. Previous Work

In this section, we briefly review prior approaches to object detection, feature extraction, and object classification. The object detection is drawing the ever-increasing efforts and it is still a challenging task. In natural images, it is very hard to find the object of consideration due to small objects and complex background. Furthermore, the gap between machine learning and human perception makes the task even harder [17]. The object detection methods are generally divided into two categories, i.e., grouping and window scoring. The grouping method is used to generate multiple segments of an image which contain objects. The hierarchical image segmentation is a grouping method approach to merge segments according to the similarities. In the author used algorithm of Felzenszwalb and Huttenlocher [18] to generate a set of small initial regions. It defined segmentation as graph problems where each vertex is an element to be segmented, and edges are between two neighboring regions. Then, it produced region comparisons, with each segment corresponding to a connected component in the graph. CMPC [19] and methods of Endres and Hoiem [20] are used to solve multiple graph-cuts with different seeds and parameters to generate class independent proposals. Both of these methods produced binary foreground and background segments. Both of the above methods learned to predict the segments that cover complete objects and rank proposals accordingly. The window scoring methods are very different, with each window score calculated according to how likely it is to contain an object. The objectness [21] is a window-based approach in which each candidate window scored on different image cues. It stands as one of one of the earliest object proposal methods and is capable of measuring

the likelihood that objects are present in an image. This method used in saliency, color contrast, edge density and super pixel straddling cues to obtain characteristics of images and adopts Bayesian's framework to combine several cues. The newly merged cues outperform the state of art saliency measure. Rahtu et al. [22] used a large number of randomly sampled boxes from an objectness and multiplied them with proposal generated from single, pair and triplet superpixels segmentations.

Feature extraction and representation is a critical step from multimedia data. It is used that how to extract meaningful features that can reflect the inner content of the images. Though, there is a big gap in research needs to fill to target the issue of useful feature extraction. The HOG provides a classification for these features. The basic hypothesis is local object appearance. it can often be characterized relatively well by the distribution of local intensity gradients. In addition, the HOG features are invariant to changes in illuminations or shadowing [23]. Zhu et al. [24] proposed a hierarchical structural learning method based on HOG features for object detection with two or three layers. Felzenszwalb et al. [25] have also shown a successful combination of HOG with the part-based model. However, despite that the above methods use an exhaustive search, the HOG features with a linear classifier is the only feasible choice from a computational perspective.

Yu and War [26] proposed the model which obtain the HOG features that emphasis on angle points. The HOG mined for numerous object detection scheme in which single or multiple stirring objects are categorized and marked. The outcomes showed that the proposed technique outperforms the existing results. Korkmaz et al. [27] developed the model for recognition of the stomach cancer images with probabilistic HOG feature vector histograms. The features' vectors obtained by plotting features on normal, benign, and malign original stomach images. Using these vectors, histograms of normal, benign, and malignant, the stomach images were plotted. The BOW [12] proposed used for image processing, with the help of the SIFT feature extraction algorithm. SIFT is also a well-known algorithm that represents the critical points of images and converts them into a features descriptor. It uses the K-means clustering method to get a cluster of features and finally classifies the images by SVM [28]. Liu et al. [29] used to improve the classification accuracy by mining meaningful SIFT features. Initially, high discriminative SIFT features are extracted with the correlation coefficients. Then, feature pairs are selected by using a minimum spanning tree. After that, high discriminative SIFT features and feature pairs are manipulated to paradigm the visual word dictionary and visual phrase dictionary, respectively.

In the last decade, classification accuracy is increased using deep learning method. The deep convolutional network developed AlexNet in 2012 by Krizhevshy et al. [30]. It needs high computational devices (i.e., GPU, a very deep network with 60 million and 650,000 neurons, etc.). It significantly outperformed all of the prior competitors and won the challenge by reducing the top-5 error to 15.3%. The second place top-5 error rate, on the other hand, which was not a CNN variation, was approximately 26.2%. With the popularity of AlexNet, Zeiler and Fergus [31] developed a model to visualize and understand the convolutional network, attempting to outdo the model developed by Krizhevsky et al. [30]. After the visualize model of Krizhevsky et al. [30] observed that the small changes in architecture improved classification performance. The only disadvantage of AlexNet is that model had too many parameters. The NIN [32] team developed a network which utilized a fewer number of the parameter: NIN's model had 7.5 million parameters, as compared to AlexNet's 60 million parameters. The Google team purposed a new, deep convolutional network model, called the Inception model [33]. This model reduced the network parameters even further: 4 million parameters as compared AlexNet's 60 million parameters. As for object detection, the Inception model used a similar

approach to R-CNN, but, for region proposal, the model combined selective search and multi-box approaches, with 50% of proposals taken from the selective search and 200 proposals taken from the multi-box [34]. Girshick et al. [35] introduced their R-CNN method which defines object detection in a two-step process. This method generates a set of category independent proposals using bottom-up grouping (i.e., selective search). Girshick et al. [35] then used a deep convolutional neural network on those generated proposals. This method dramatically improves the performance proposal generation, proposal classification, and overall object detection by replacing the traditional sliding window approach with object proposals, thus achieving a state-of-the-art object detection performance. Fast R-CNN [36] is an improvement of Girshick et al's previous work and allows for faster object detection. The VGG [37] team developed an even deeper convolutional network. The team observed that the depth of the convolutional network has a great deal of impact on image detection. For their mode, the VGG team utilized small 3×3 convolutional filters and set the convolutional stride to one, therefore no information got lost, all while utilizing has 19 weighted layers.

In this paper, an efficient combined approach to object detection and classification is proposed. This method generates high-quality object proposals following scoring and ranking measures, in order to increase the detection performance. Then refined BOW technique with a combination of SIFT and GIST is used to extract features. Lastly SVM is used for classification task to achieve a robust performance.

## 3. Proposed Methodology

The proposed method consists of four stages: (i) object proposals generation, (ii) feature extraction, (iii) CGSF integration, and (iv) classification. The details of the method are given in the subsections below. The flowchart of the overall framework is shown in Fig. 1.

### 3.1 Image Proposals

This section presents the detailed description of the image proposal's generation. The proposed method combines hierarchical segmentations [5] and window scoring method [6]. First, we generate object proposals on images through the use of the agglomerative clustering grouping method. For example, Fig. 2 is the original image and Fig. 3 contains the achieved proposals. We then score the boxes according to the sums of the magnitude of the all the edges in each edge group, minus the edge groups of the contours that straddle the bounding box. Finally, we rank the object proposals according to the score of the boxes, as shown in Fig. 4. The significant steps are as follows.

#### 3.1.1 Segmentation

The traditional approach to grouping methods uses segmentation to obtain a small set of starting regions. However, we also take a graph-based approach to segmentation, as the graph-based segmentation of Felzenszwalb and Huttenlocher [18] approaches also get a small set of initial regions, but with more appropriate rates and accuracy. It converts images into a graph—pixels, which are the vertices, and neighboring pixels connected on their edges. We then manipulate the graph to segment the image.

### 3.1.2 Hierarchical clustering

We group initial segments obtained from the above step according to the similarity measure between neighboring regions. This process continues until the whole image becomes one cluster/region. We use color, texture, size and gap similarities to measure. Agglomerative (bottom-up) clustering method is applied to different color spaces to cover a more diverse set of regions. Regions from each hierarchy are then combined, while duplicate regions are removed at the end.

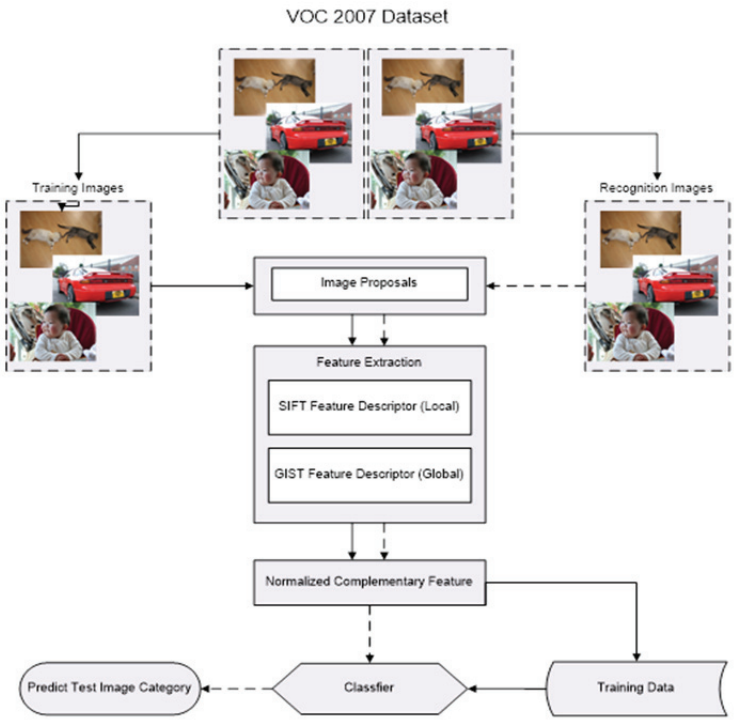


**Fig. 1**. Workflow of the proposed framework.

**Fig. 2.** Original image.

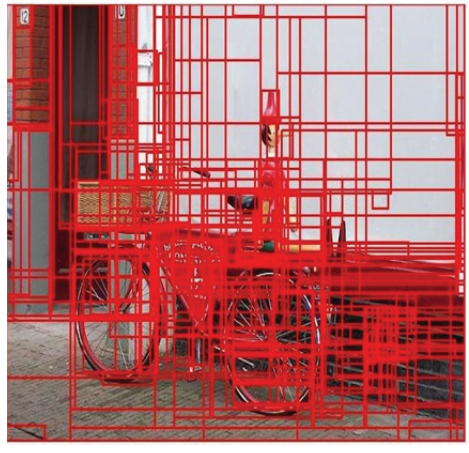

**Fig. 3.** Proposal evaluation on VOC.

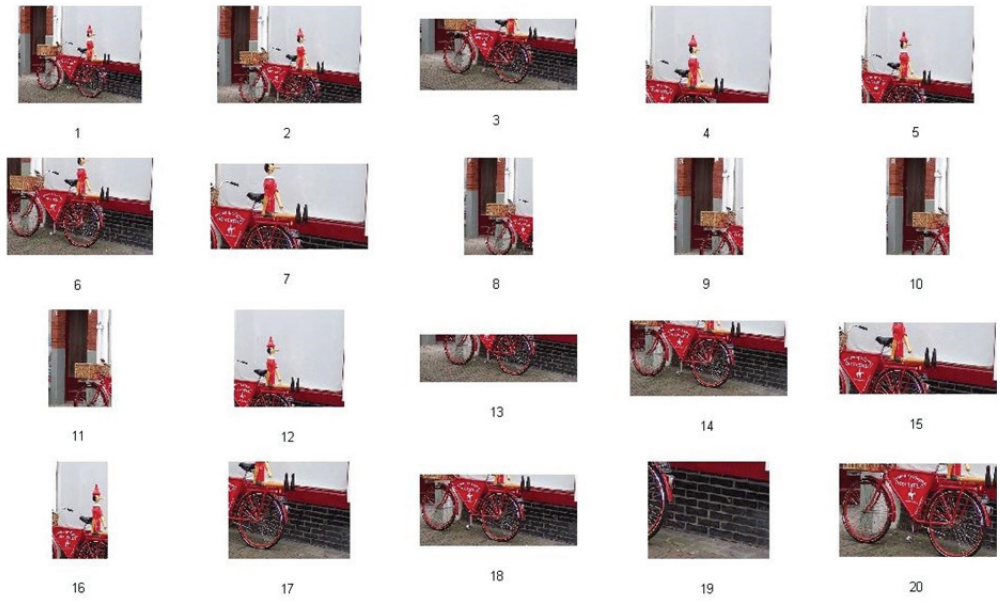


**Fig. 4.** Top ranked proposals on VOC.

### 3.1.3 Edge detection & edge group

For edge detection, we use structured edge detection. The structure-forest extracts image patches from the image convert each image patch into vectors, extracts the image features for each patch, and finally predicts scores for the patches on the edge. Edges obtained from the detector are then combined into eight connected neighboring edges with similar orientation until the orientation differences are above pi/2 to form the edge groups. This method shows greater accuracy and speed as compared to traditional edge detectors.

### 3.1.4 Score regions

Given a set of object proposals obtained from hierarchical clustering, we calculate the score of each object proposal. This is accomplished by summing the magnitude of every wholly enclosed edge in the group within each region and subtracting the magnitude of every edge in the group which straddles the object region. The value of $w_b(s_i)$ is calculated for each edge group to check if the group is wholly enclosed in the region. When an edge group is not entirely closed in the box, then $w_b(s_i) = 0$. If an edge group wholly enclosed in the box $w_b(s_i)$ calculated as below:

$$w_b(\mathrm{s}_i) = 1 - \max_t \prod_j^{|\mathrm{T}|-1} a(t_j - t_{j+1}) \tag{1}$$

where a is the affinity and "$t$" is the order path, so the above equation finds the order path with the max affinity between the groups. We then compute the score using the formula:

$$h(\mathrm{b}) = \frac{\sum_i w_b(s_i) m_i}{2(b_w + b_h)^k} \tag{2}$$

where $b_w$ and $b_n$ are the box width and height and $k$ is the bias value for larger boxes.

### 3.1.5 Ranking

We rank objects proposed according to the score obtained from Eq. (2), where a few thousands object proposals are passed for the classification task.

## 3.2 Feature Extraction

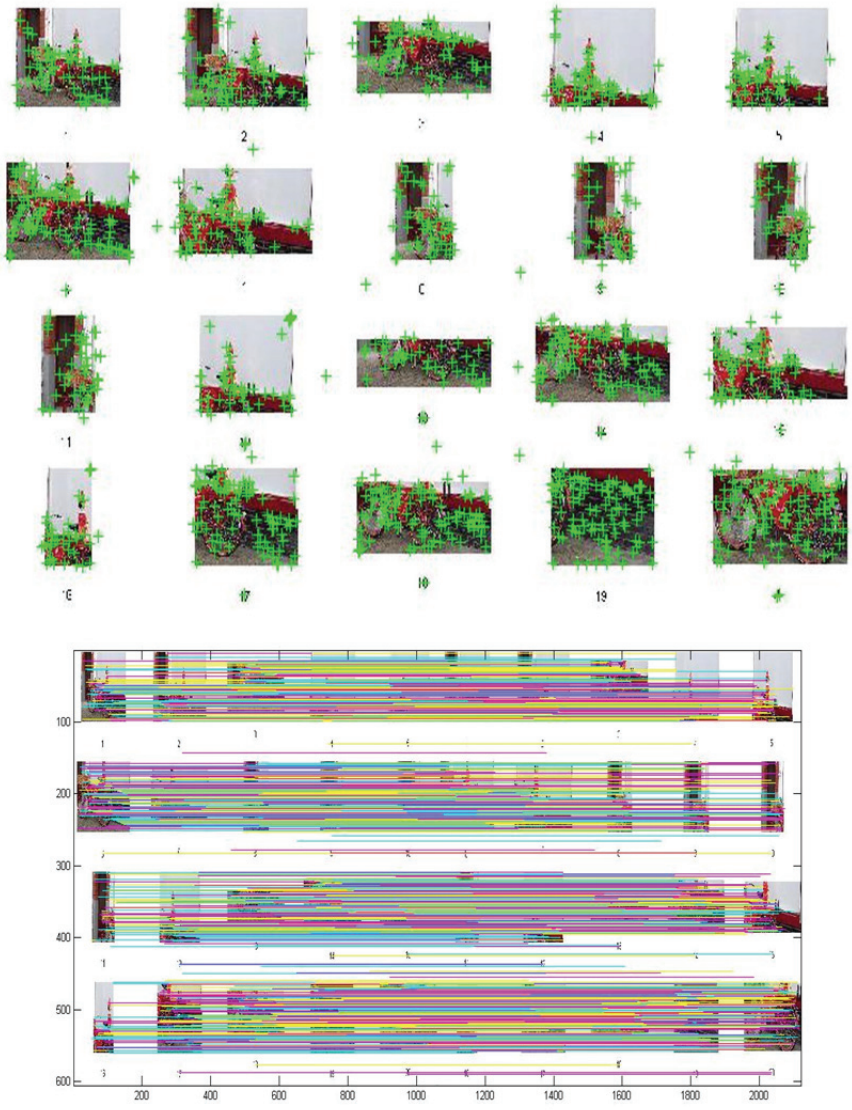


**Fig. 5.** Local feature detection and description using SIFT.

Features play a very significant role in image processing the image features are associated with 'exciting' scene elements in the image formation process and can be described by a feature vector.

Feature extraction is the method of extracting relevant information from image data to form feature vectors for classification purpose. There have been numerous features extraction techniques developed to construct feature vectors. In this paper, SIFT and GIST techniques are used to extract local and global features from image proposals shown in Figs. 5 and 6 (the feature extraction is an essential step in the construction of any pattern). Finally, the features obtained are combined to achieve high classification performances.

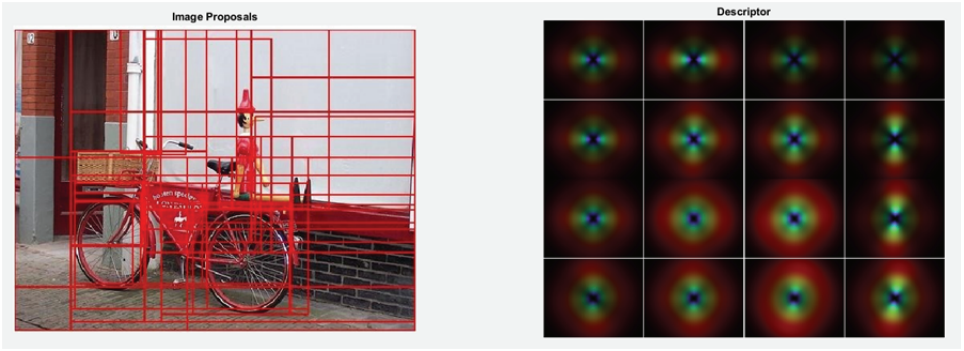


**Fig. 6.** Global feature detection and description using GSIT.

### 3.2.1 SIFT

Our goal was to find some local vital points that give us information about the object in an image. Interest point descriptors are formed at BOW using SIFT [38]. The primary task of SIFT is to obtain local features of an image and display them into various translation invariant patches [39]. First, it detects an interesting patch with an interest operator. Then, the SIFT feature detector converts these patches into 128×128 vector representations. There are primarily four steps involved in SIFT. In the first step, the scale and rotation invariant interest points (i.e., critical points) are extracted. The scale space of an image is a function $L(x,y,\sigma)$ that is produced from the convolution of a Gaussian kernel (at different scales) with the input image, as shown in Eq. (3).

$$L(x,y,\sigma) = G(x,y,\sigma) \cdot I(x,y) \tag{3}$$

In the second step, SIFT detects maxima and minima of difference-of-Gaussian in scale space. Each pixel of an image is compared with its eight neighboring pixels at identical scale. Furthermore, if the subsequent value of a pixel is the minimum or maximum among all corresponding pixels, it is anticipated to be a key point. In the third step, the crucial inaccurate point localization is dealt, pointed with low contrast (sensitive to noise) and localized along an edge are eliminated using quadratic Taylor expansion as shown in Eq. (4).

$$D(\mathrm{x}) = \mathrm{D} + \frac{\partial D^t}{\partial x} X + \frac{1}{2} X^t \frac{\partial^2 D}{\partial X^2} X \tag{4}$$

Further, in the last step, absolute extrema that high-level location nearest crucial point is calculated using the following formula.Further, in the last step, absolute extrema that high-level location nearest crucial point is calculated using the following formula.

$$D(\mathrm{X}^*) = \mathrm{D} + \frac{1}{2}\frac{\partial D^T}{\partial X} X^* \tag{5}$$

### 3.2.2 GIST

The GIST descriptor was first proposed for recognition of real-world scenes and has been proven to work well for scene classification [40]. It is used to obtain the global feature vector with a holistic representation of an image. GIST descriptors for an image are computed using spatial pyramid technique, which convolves the image with 32 Gabor filters at eight orientations and four scales, producing 32 feature maps of the same size each comprising of 16 (4×4) points. In the end, we have a 512-dimensional feature vector for each image, which is obtained by divided each feature map into 16 regions. The GIST descriptor for each image is normalized to have unit *L1* as the norm. The Gabor filter obtained using the following Eq. (6).

$$z_\theta^K = l \exp^{\left(\frac{x_\theta^2 + y_\theta^2}{2\sigma^{2(k-1)}}\right)} * \exp^{(2\pi j(ux_\theta + vy_\theta))} \tag{6}$$

where $l$ is the scale of the filter, $K$ is a positive constant, $\sigma$ is the variance of the Gaussian function and is the number of direction under the scale of $l$. The filtered image $F$ where $k$ represents the level of filters and shows the variance in Gaussian function. Hence the filtered image $F$ is expressed by Eq. (7)

$$F_\theta^K = Z_\theta^K L \tag{7}$$

### 3.2.3 CGSF integration

The SIFT [38] and GIST [40] are combined to improve classification accuracy; the feature fusion scheme is chosen to connect both SIFT Local and GIST Global feature vectors by weight parameter $W$, as seen in Eq. (8). Furthermore, CGSF is normalized by the image of the dataset to unity 512×512-pixel size.

$$CGSF = WL + (1 - W)G \tag{8}$$

where $L$ denotes the Local feature vector, $G$ denotes the global feature vector, and $W$ means the integration weight factor.

$$G(X, Y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2 + y^2}{2\sigma^2}} \tag{9}$$

### 3.2.4 Classification

The VLFeat toolbox is used for regression and classification tasks [28]. This SVM train function provides many different default parameters to improve effectiveness adequately. Through extensive range testing, we determine that keeping the size of proposals up to 1500 allows the classification results are more efficient and robust.

# 4. Evaluation and Results

To establish that the proposed method has a respectable performance for present purposes, a comparison between the proposed technique and recently reported results are presented in this section. The experimental setup was executed using MATLAB 2015 with an Intel 2.66 GHz CPU 4.0 GB RAM. We evaluated the performance of our proposed method on the most famous image repository, the Pascal Visual Object Classes Challenge 2007 (VOC2007) [41], used for detection and classification tasks.

Pascal VOC2007 provides standardized images with a variety of variations—such as scales, illuminations, viewpoints, and positions—making this database ideal for object recognition and classification. The dataset contains 9963 images, with a training set comprising of 2501 images, a validation set containing 2510 images, and a test set containing 4952 images, all within 20 object classes in four broad categories: person, animal, vehicle, and indoor. Training images are labeled with the ground truth from 20 object classes. Every picture has an annotation that contains the bounding box information and difficulty level of the object. Furthermore, the quality of our proposed approach was evaluated using the following measures: average best overlap (ABO), mean average best overlap (MABO), average precision (AP), and mean average precision (MAP). The ABO and MABO are used to assess the quality of the image proposals generated. The AP and MAP are used to evaluate the performance of the classification.

## 4.1 Average Best Overlap

ABO, for any class, is achieved by calculating the best overlap on Ground Truth of class and the proposed object region of said class and then taking its average. Overlap is the intersection of the proposed region with the ground truth over an area of their union.

$$IoU\ (box, gtruth) = \frac{area(box) \cap area\ (gtruth)}{area(box)\ \cup area\ (gtruth)} \tag{10}$$

## 4.2 Mean Average Best Overlap

Mean of the average best overlap for all the classes ABO.

## 4.3 Average Precision

It is analogous to ABO.

## 4.4 Mean Average Precision

Mean of the average precision for all the class.

**Table 1.** MABO on VOC2007 dataset on top 1500 proposals

| Method | Test images | Proposals | MABO |
|---|---|---|---|
| Edge box | 4952 | 1500 | 0.799 |
| Selective search | 4952 | 1500 | 0.820 |
| Our approach | 4952 | 1500 | 0.833 |

**Table 2.** ABO for 20 classes of VOC on top 1500 proposals

| VOC class | Edge box | Selective search | Our approach |
|---|---|---|---|
| Plane | 0.771 | 0.796 | 0.807 |
| BiCycle | 0.824 | 0.844 | 0.861 |
| Bird | 0.796 | 0.812 | 0.812 |
| Boat | 0.779 | 0.768 | 0.784 |
| Bottle | 0.692 | 0.660 | 0.673 |
| Bus | 0.841 | 0.864 | 0.868 |
| Car | 0.788 | 0.783 | 0.808 |
| Cat | 0.827 | 0.906 | 0.909 |
| Chair | 0.783 | 0.798 | 0.808 |
| Cow | 0.827 | 0.829 | 0.854 |
| Table | 0.817 | 0.891 | 0.894 |
| Dog | 0.837 | 0.895 | 0.900 |
| Horse | 0.815 | 0.828 | 0.841 |
| Bike | 0.815 | 0.829 | 0.846 |
| Person | 0.755 | 0.754 | 0.766 |
| Pottedplant | 0.746 | 0.740 | 0.758 |
| Sheep | 0.814 | 0.797 | 0.828 |
| Sofa | 0.828 | 0.904 | 0.907 |
| Train | 0.801 | 0.856 | 0.863 |
| Tvmonitor | 0.821 | 0.842 | 0.868 |

In Tables 1 and 2, the comparison of our approach with existing methods to generate high-quality proposals is shown. As can be seen in Fig. 7, in contrast to other methods, our approach's results have a high MABO score of 0.833 at a similar number of locations. Moreover, our technique yields the best ABO for 20 classes of VOC, on top the 1500 proposals shown in Fig. 8.

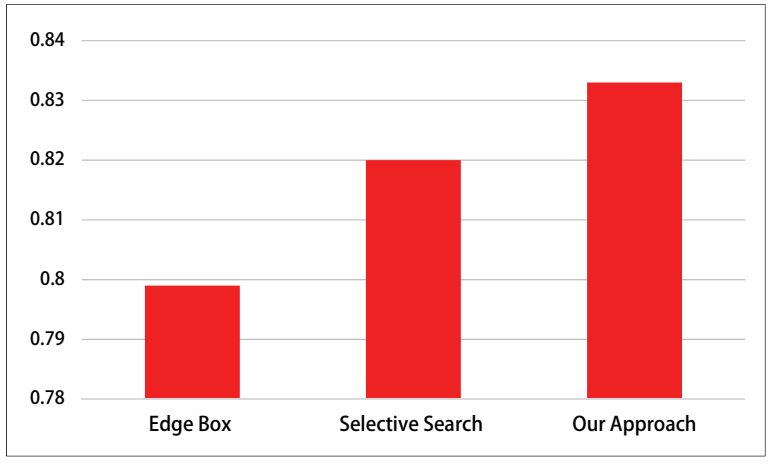


**Fig. 7.** MABO performance of our approach.

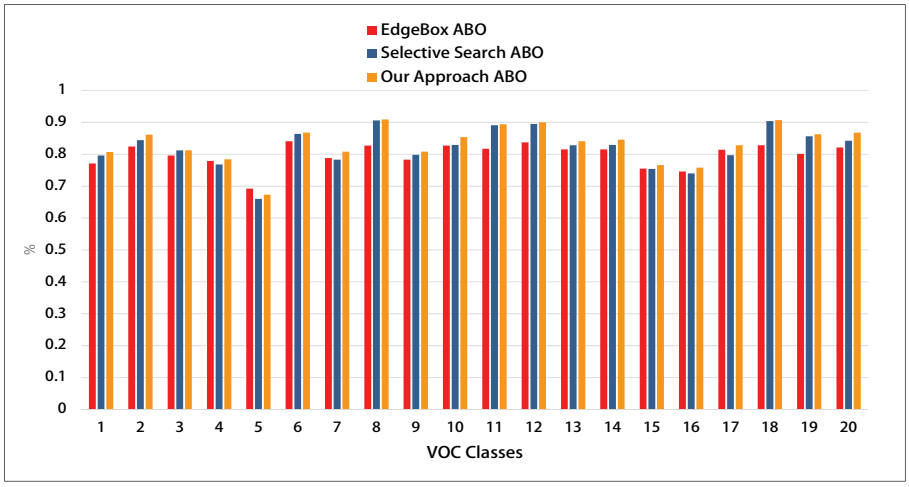


**Fig. 8.** Comparison of our approach with other methods for 20 classes of VOC on top 1500 proposals.

**Table 3.** MAP on VOC2007 dataset on top 1500 proposals

| Methods | Test images | Proposals | MAP |
|---|---|---|---|
| HOG-based [24] | 4952 | 1500 | 0.347 |
| HOG-based [25] | 4592 | 1500 | 0.342 |
| Selective search [5] | 4952 | 1500 | 0.354 |
| Our approach | 4952 | 1500 | 0.394 |

**Table 4.** Classification result: comparison of our approach with traditional approaches for 20 classes of VOC on top 1500 proposals

| VOC classes | HOG-based [24] | HOG-based [25] | Selective search [5] | Our approach |
|---|---|---|---|---|
| Plane | 0.544 | 0.532 | 0.556 | 0.563 |
| BiCycle | 0.565 | 0.560 | 0.539 | 0.570 |
| Bird | 0.162 | 0.140 | 0.155 | 0.160 |
| Boat | 0.194 | 0.160 | 0. 132 | 0.191 |
| Bottle | 0.293 | 0.355 | 0.228 | 0.353 |
| Bus | 0.559 | 0.548 | 0.499 | 0.557 |
| Car | 0.440 | 0.497 | 0.384 | 0.492 |
| Cat | 0.429 | 0.322 | 0.472 | 0.479 |
| Chair | 0.176 | 0.160 | 0.142 | 0.173 |
| Cow | 0. 280 | 0.267 | 0.323 | 0.325 |
| Table | 0.269 | 0.141 | 0.302 | 0.309 |
| Dog | 0.319 | 0.221 | 0.369 | 0.375 |
| Horse | 0.487 | 0.461 | 0.446 | 0.484 |
| Bike | 0.554 | 0.521 | 0.522 | 0.551 |
| Person | 0.411 | 0.479 | 0.340 | 0.476 |
| Pottedplant | 0.099 | 0.089 | 0.157 | 0.161 |
| Sheep | 0.564 | 0.353 | 0.413 | 0.417 |
| Sofa | 0. 311 | 0.195 | 0.321 | 0.325 |
| Train | 0.477 | 0.469 | 0.474 | 0.474 |
| Tvmonitor | 0.411 | 0.383 | 0.453 | 0.456 |

In summary, Tables 3 and 4 show the classification accuracy of the proposed method. It demonstrates that our approach yields the best results for non-rigid classes: plane, cat, cow, dog, plant, and sheep. Our technique scores 6 for non-rigid classes and scores 3 for rigid categories. The use of BOW, by combining SIFT and GIFT features, makes it better suited for non-rigid classes than HOG-features. The HOG-based methods perform better for the rigid classes such as, bike, bottle, bus, car, person, and train, as shown in Fig. 9. Furthermore, our approach slightly performs better than selective search [5] or the BOW method. Undoubtedly, this is due to our proposal generation methods resulting in higher ABO (for all classes) than selective search. Lastly, results from the Pascal VOC2007 detection task test set are shown in Fig. 10. To conclude, our approach is not only highly practical for generating object proposals but also more efficient than other methods.

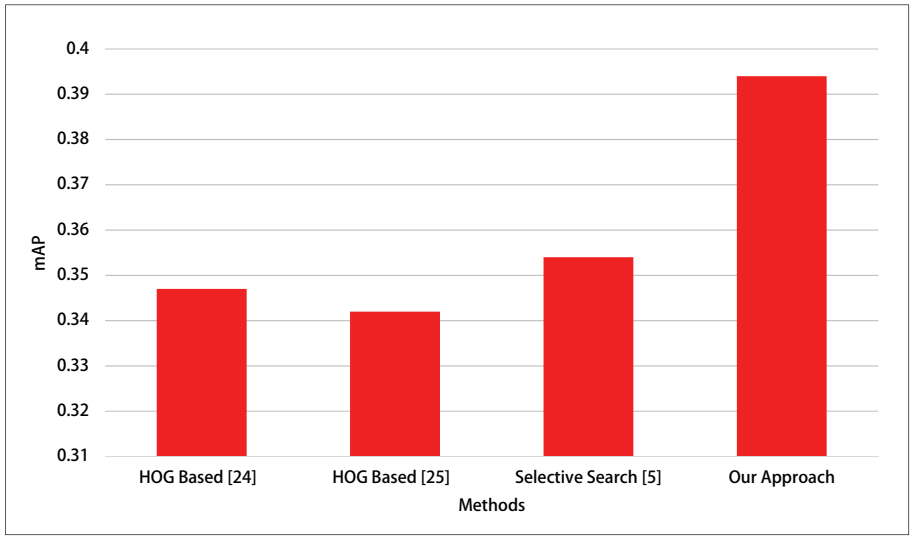


**Fig. 9.** Map performance of our approach.

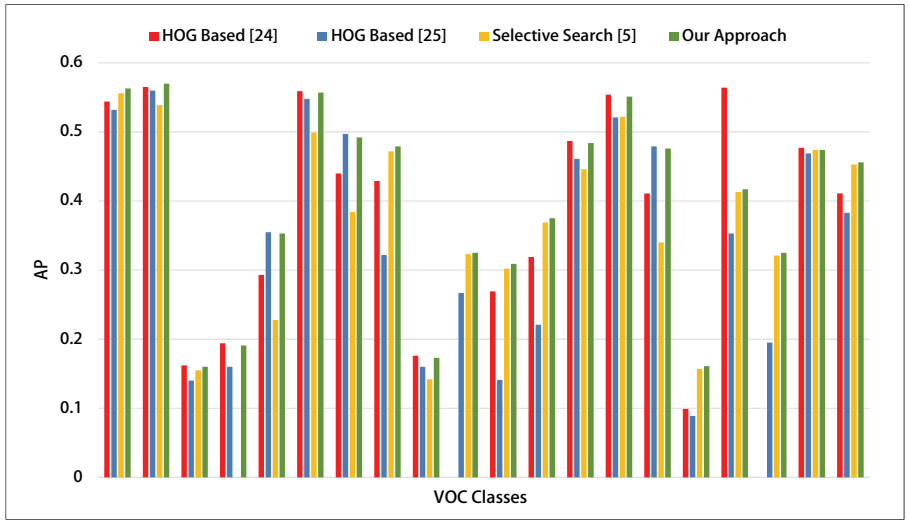


**Fig. 10.** Classification: AP performance of our approach on VOC classes for 1500 proposals.

## 5. Conclusion

In Summary, an efficient hybrid approach based on a combination of selective search, edge box, and SIFT algorithm with GIST proposed. Firstly, this method results in adequate detection rates for object detection tasks compared to object detection solely utilizing selective search that matches the accuracy of selective search, with only 25% the number of the proposal after ranking said proposals. Our method results in high-quality class independent object locations, with a MABO of 0.833 at 1500 proposals; which are close to optimal for our version of BOW based object classification. Secondly, our method results in satisfactory classification rates based on combine SIFT algorithm with GIST (with local SIFT with MAP 0.394). This hybrid approach provides more description of the features of the image as compared to traditional methods; therefore, it is much efficient in classification. The similarity between the image features measured by BOW paradigm by using the VLFeat linear SVM classifier. The experimental results show that CGSF approach is much efficient than traditional methods. Moreover, it can be seen in the tables that our system performs best for the non-rigid classes and achieved results are slightly better than a conventional approach based on BOW. Our approach is also theoretically better suited for the rigid classes than the HOG-features based methods which perform better for the rigid categories.

## 6. Future Work

In the future, the score function can be further optimized by penalizing the portion of edge groups that overlap the region boundary, instead of subtracting the strength of edges present in the edge group. The edge box generates redundant object proposals in each scale. Therefore, by reducing unnecessary object proposals, edge box performance can also be further improved. Furthermore, we can use a high post classification deep learning method, which increases classification efficiency with the advancement of deep learning of convolutional neural networks.

## Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 61571312), Academic and Technical Leaders Support Foundation of Sichuan province (No. 2016-183-5), National Key Research and Development Program Foundation of China (No. 2017YFB0802300). The authors would like to thank Ms. Siobhan Kathryn He for constructive criticism of the manuscript.

## References

[1] Z. Rahman, Y. F. Pu, M. Aamir, and F. Ullah, "A framework for fast automatic image cropping based on deep saliency map detection and Gaussian filter," *International Journal of Computers and Applications*, 2018. https://doi.org/10.1080/1206212X.2017.1422358.

[2] M. Aamir, Y. F. Pu, W. A. Abro, H. Naeem, and Z. Rahman, "A hybrid approach for object proposal generation," in *The Proceedings of* the *International Conference on Sensing and Imaging*. Cham: Springer, 2017, pp. 251-259.

[3] D. Phan, C. M. Oh, S. H. Kim, I. S. Na, and C. W. Lee, "Object recognition by combining binary local invariant features and color histogram," in *Proceedings of 2013 2nd IAPR Asian Conference on Pattern Recognition (ACPR),* Naha, Japan, 2013, pp. 466-470.

[4] N. Najva, "SIFT and tensor based object classification in images using Deep Neural Networks," in *Proceedings of International Conference on Information Science (ICIS),* Kochi, India, 2016, pp. 32-37.

[5] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154-171, 2013.

[6] C. L. Zitnick and P. Dollar, "Edge boxes: locating object proposals from edges," in *Computer Vision-ECCV 2014*. Cham: Springer, 2014, pp. 391-405.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[8] Y. Ke and R. Sukthankar, "PCA-SIFT: a more distinctive representation for local image descriptors," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, 2004, pp. 506-513.

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.

[10] B. A. Tama and K. H. Rhee, "A detailed analysis of classifier ensembles for intrusion detection in wireless network," *Journal of Information Processing Systems*, vol. 13, no. 5, pp. 1203-1212, 2017.

[11] P. Iswarya and V. Radha, "Speech query recognition for Tamil language using wavelet and wavelet packets," *Journal of Information Processing Systems*, vol. 13, no. 5, pp. 1135-1148, 2017.

[12] F. Ciompi, C. Jacobs, E. T. Scholten, M. M. Wille, P. A. De Jong, M. Prokop, and B. van Ginneken, "Bag-of-frequencies: a descriptor of pulmonary nodules in computed tomography images," *IEEE Transactions on Medical Imaging*, vol. 34, no. 4, pp. 962-973, 2015.

[13] T. Li and W. Zhang, "Classification of brain disease from magnetic resonance images based on multi-level brain partitions," in *Proceedings of 2016 IEEE 38th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC),* Orlando, FL, 2016, pp. 5933-5936.

[14] E. Tola, V. Lepetit, and P. Fua, "A fast local descriptor for dense matching," in *Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 2008, pp. 1-8.

[15] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346-359, 2008.

[16] W. Tahir, A. Majeed, and T. Rehman, "Indoor/outdoor image classification using gist image features and neural network classifiers," in *Proceedings of 2015 12th International Conference on High-Capacity Optical Networks and Enabling/Emerging Technologies (HONET),* Islamabad, Pakistan, 2015, pp. 1-5.

[17] H. T. Manh and G. Lee, "Small object segmentation based on visual saliency in natural images," *Journal of Information Processing Systems*, vol. 9, no. 4, pp. 592-601, 2013.

[18] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167-181, 2004.

[19] J. Carreira and C. Sminchisescu, "CPMC: automatic object segmentation using constrained parametric min-cuts," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 34, no. 7, pp. 1312-1328, 2011.

[20] I. Endres and D. Hoiem, "Category-independent object proposals with diverse ranking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 222-234, 2014.

[21] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189-2202, 2012.

[22] E. Rahtu, J. Kannala, and M. Blaschko, "Learning a category independent object detection cascade," in *Proceedings of 2011 IEEE International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 1052-1059.

[23] D. Ghimire and J. Lee, "Extreme learning machine ensemble using bagging for facial expression recognition," *Journal of Information Processing Systems*, vol. 10, no. 3, pp. 443-458, 2014.

[24] L. L. Zhu, Y. Chen, A. Yuille, and W. Freeman, "Latent hierarchical structural learning for object detection," in *Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010.

[25] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627-1645, 2010.

[26] T. T. Yu and N. War, "Condensed object representation with corner HOG features for object classification in outdoor scenes," in *Proceedings of 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, Kanazawa, Japan, 2017, pp. 77-82.

[27] S. A. Korkmaz, A. Akcicek, H. Binol, and M. F. Korkmaz, "Recognition of the stomach cancer images with probabilistic HOG feature vector histograms by using HOG features," in *Proceedings of 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY),* Subotica, Serbia, 2017, pp. 339-342.

[28] VLFeat algorithm [Online]. Available: http://www.vlfeat.org/.

[29] L. Liu, Y. Ma, X., Zhang, Y. Zhang, and S. Li, "High discriminative SIFT feature and feature pair selection to improve the bag of visual words model," *IET Image Processing*, vol. 11, no. 11, pp. 994-1001, 2017.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097-1105, 2012.

[31] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision-ECCV 2014*. Cham: Springer, 2014, pp. 818-833.

[32] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013; https://arxiv.org/abs/1312.4400.

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhouche, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 1-9.

[34] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 2147-2154.

[35] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 580-587.

[36] Fast R-CNN [Online]. Available: https://github.com/rbgirshick/fast-rcnn.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014; https://arxiv.org/abs/1409.1556.

[38] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[39] H. Naeem, G. Bing, M. R. Naeem, M. Aamir, and M. S. Javed, "A new approach for image detection based on refined Bag of Words algorithm," *Optik-International Journal for Light and Electron Optics*, vol. 140, pp. 823-832, 2017.

[40] 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, no. 3, pp. 145-175, 2001.

[41] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: a retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98-136, 2015.

**Muhammad Aamir** https://orcid.org/0000-0002-7679-0980

He received his bachelor of engineering degree in Computer Systems Engineering from the Mehran University of Engineering & Technology, Jamshoro, Sindh, Pakistan in 2008. And master of engineering degree in Software Engineering from Chongqing University, China in 2014. Currently, he is a PhD research student in Sichuan University, China. His research interest includes pattern recognition, computer vision, image processing, deep learning and fractional calculus.

**Yi-Fei Pu** https://orcid.org/0000-0003-2975-4976

He received the Ph.D. degree from the College of Electronics and Information Engineering, Sichuan University, in 2006. He is currently a Full Professor and a Doctoral Supervisor with the College of Computer Science, Sichuan University, a Chief Technology Officer with Chengdu PU Chip Science and Technology Company, Ltd., and is elected into the Thousand Talents Program of Sichuan Province and the Academic and Technical Leader of Sichuan Province. He has first-authored about 20 papers indexed by SCI in journals, such as *International Journal of Neural Systems, IEEE Transactions on Image Processing, IEEE Transactions on Neural Networks and Learning Systems, IEEE Access, Mathematical Methods in Applied Sciences, Science in China Series F: Information Sciences,* and *Science China Information Sciences.* He held several research projects, such as the National Nature Science Foundation of China and the Returned Overseas Chinese Scholars Project of Education Ministry of China, and holds 13 China Inventive Patents, as the first or single inventor. He focuses on the application of fractional calculus and fractional partial differential equation to signal analysis, signal processing, image processing, circuits and systems, and machine intelligence.

**Ziaur Rahman** https://orcid.org/0000-0001-8233-567X

He received an M.S. degree in software engineering in 2017 from Chongqing University, Chongqing, China. Currently, He is pursuing PhD degree from Sichuan University, Chengdu, China. His research includes are image processing, deep learning, and fractional calculus.

**Waheed Ahmed Abro** https://orcid.org/0000-0001-5878-3448

He received his bachelor of engineering degree in Computer System Engineering from the Mehran University of Engineering & Technology, Jamshoro, Sindh, Pakistan in 2008. And master of science in a Computer System from National University of Computer and Emerging Sciences, Islamabad, Pakistan in 2015. Currently, he is PhD research student in Southeast University, China. His research interest includes natural language processing, spoken language understanding, computer vision and pattern recognition.

**Hamad Naeem** https://orcid.org/0000-0003-1511-218X

He graduated from NFC IET, Multan, Pakistan, and Chongqing University, Chongqing, China in 2012 and 2016, and received B.S. and M.E. degrees, respectively. He received excellent master student award from Chongqing University, China in 2016. Currently, he is pursuing a PhD degree in Software Engineering at Sichuan University, China. He has published various articles in reputed SCIE journals. His research interest includes internet security, malware detection, image processing and machine learning. Mr. Hamad is a recipient of one belt one road Scholarship from Sichuan University, Chengdu, China.

**Farhan Ullah** https://orcid.org/0000-0003-2422-575X

He received his M.S. degree in Computer Science from CECOS University Peshawar, Pakistan, in 2012 and B.S. degree in Computer Science from University of Peshawar, Pakistan, in 2008. He is currently pursuing PhD degree in computer science from School of Computer Science, Sichuan University, Chengdu, China. He is on PhD study leave from Lecturer position in Department of Computer Science, COMSATS University Islamabad, Sahiwal Campus, Pakistan. He has authored/co-authored 13 publications including 6 SCI-indexed journals. His research interests include software similarity, software piracy, machine learning and data science.

**Aymen Mudheher Badr** https://orcid.org/0000-0002-1373-9950

He received his bachelor's degree in Computer Science from Baghdad University, Iraq, in 2002. And master's degree of Computer Science and Technology from Chongqing University, China, in 2015. Currently, he is a PhD researcher at Sichuan University, China. His research interest in the information society and deep machine learning.