# *Image object Classification using Optimal Bag-of-Words*

**D.Priyanka (PG Student)[1],Mrs.S.Vanitha Sivagami[2],Dr.K.Muneeswaran[3]**

**Department of Computer Science and Engineering,**

**Mepco Schlenk Engineering College, Sivakasi, India.**

**priyankadhamodar@gmail.com[1] , svanitha@mepcoeng.ac.in[2], kmuni@mepcoeng.ac.in[3]**

***Abstract-Image Classification plays an important role in applications like 3D modeling, image search and analytics, remote sensing and travel guide recommendation. Lot of works has been implemented to achieve accuracy but still it remains unsolved due to its large intra-class variance. Hence, this proposes the extraction of ROI of an image using interest points. ROI of an image are then used for k-means and self organizing map(SOM) clustering algorithm to form bag-of-words for image classification which achieves high classification accuracy. The visual codebooks are constructed for an image using SOM and K-means based bag-of-words model. Multi label classifier is used for developing model.. Experiments were conducted on MSRC-21class database.***

***Key Terms—Image Classification, K-means, Self Organizing Map, Bag-Of-Words.***

## I. INTRODUCTION

Image classification accurately aims to categorize a query image into a discrete category from training images. Image classification plays a major role in image search analytics. Image classification is an important and challenging task in various application domains, including biomedical imaging, biometry, video surveillance, vehicle navigation, industrial visual inspection, robot navigation, and remote sensing. Even now the major reason is that images contain high diverse visual contents. Among various object recognition methods, the methods based on BoW have been paid much attention due to its low computation cost, robustness against illumination variation, partial occlusion and clutter background. In this paper, we are construction the classifier using bag of words model by taken region of an image. Initially the corner of the points are taken from the image, then saliency map values also taken. These corner and saliency values are then further used for extracting the object region from the images by calculating the interesting points. The isolated interest points are neglected from the images by setting the specific threshold(based on study analysis). SIFT features are then extracted from the regions. K-means and SOM clustering are then used for the constructing the visual codebook. Bag of words model is used for generating the image feature vector for the training and testing images. Finally Multi-class classifier is used for building the model for each category of the training images.

The main contributions of this paper can be summarized as follows:

1) ROI is extracted to select the object of interest from the image.
2) The SIFT feature descriptors are detected and described in the interest images. These highly distinctive features can distinguish the foreground from background of the image, and reduce the effects of the cluttered backgrounds.
3) A visual codebook is generated using K-means and self organizing map.
4) The similarities between each visual word and corresponding local feature are computed to construct image feature vector.
5) The Support vector machine (SVM) is used to perform image classification using the image feature vector.

## II. RELATED WORK

Most of the existing methods uses image search, to retrieve images represented by bag-of-visual- words [3]. Zhu et al. [4] proposed object recognition approach in the natural scenes and generate the effective descriptor for an computationally expensive training phase. Deselaers et al. [5] introduced principal component analysis to reduce the dimensions of the SIFT descriptors, to create visual vocabulary, and employed unsupervised training of Gaussian mixture model (GMM). Wu et al. [6] proposed the Histogram Intersection Kernel could also be used in an unsupervised manner to significantly improve the generation of visual codebooks based on SIFT descriptors. Histogram kernel k-means algorithm which is easy to implement and runs almost as fast as k-means. Since entire image has been used in constructing the visual codebook , it provides less accuracy in recognition. Based on SIFT descriptors, Lienhart et al. [7] used probabilistic Latent Semantic Analysis (PLSA) for images recognition. Wang et al. [8] proposed It uses appearance variation of each visual word and employs the max-min posterior pseudo-probabilities discriminative learning method to estimate GMMs of visual words. The drawback of this method is that more features have been used ,hence takes more training time. D. Nister et. al [9] demonstrated the hierarchical k-means clustering approach where the local region descriptors are hierarchically quantized in a vocabulary tree. The vocabulary tree allows a larger and more discriminatory vocabulary to be used efficiently. The drawback of this paper is that no prior knowledge of cluster size and it takes more time for processing. Harris et al. [10] proposed a robust interest points detection method, which could reduce the effect of image rotation, translation, illumination variation. Itti et al.[11] explains that saliency is a measure of difference of the image regions from their surroundings in terms of elementary features such as color, orientation, movement or distance from the eye. MacQueen et al.[12] describes a process for partitioning an N-dimensional population into k sets based on the number of samples. Mikolajczyk et al. compares the various descriptors among which the performance of SIFT descriptors gives efficient results is widely used in object recognition.

## III. PROPOSED WORK

In our proposed work, object classification is done after extracting Region of Interest (ROI) and then using Bag of Words model. The SIFT features are then extracted from the ROI of an image. Finally, Bag of Words are constructed by Clustering the features using K-means and self Organized map.

The Fig 1 explains about the framework of our proposed method. In training phase, the corner points and saliency value features are extracted from all the training images. ROI are extracted by combining the corner and saliency values. More Isolated corners are removed by specifying the certain threshold. SIFT Features are extracted from ROI of an image. Bag-of-words used for constructing a codebook using K-means and SOM Clustering. Finally, SVM Classifier is used for constructing model for each category of images.

In testing phase, the corner points and saliency value features are extracted from all the training images. ROI are extracted by combining the corner and saliency values. More Isolated corners are removed by specifying the certain threshold. SIFT Features are extracted from ROI of an image. Finally, SVM Classifier is used for constructing model for each category of images. The class label can be predicted by comparing the model of each category.

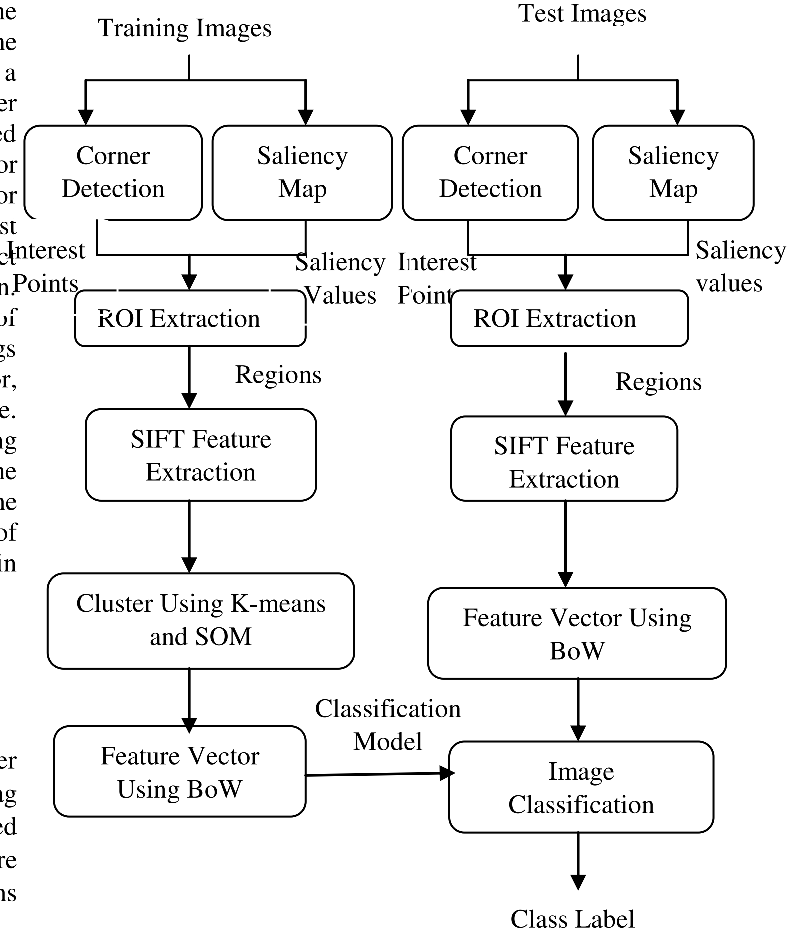


Fig 1.Framework of our proposed method

### *A. Feature Extraction*

In feature extraction, corners and saliency map are extracted. These features are described below:

*Harris Corner Detection:* The Harris Corner Detector is a mathematical operator that finds features in image. The Harris Corner Detector is just a mathematical way of determining which windows produce large variations when moved in any direction. With each window, a score R is Harris Corner Detection is calculated as,

$$R=\lambda_1\lambda_2 - k\,(\lambda_1+\lambda_2) \qquad (1)$$

Where$\lambda_1$,$\lambda_2$ are eigen values.

*Itti-Koch Saliency Map:* Saliency map is an image that shows each pixel's unique quality. It calculates the Color, orientation and intensity values. Normalization of the calculated three values will provide saliency value of a pixel. The I-K model outputs a list of image coordinates, each one corresponding to a point of attention.

$$S = \frac{1}{3}\big(N(I) + N(C) + N(O)\big) \qquad (2)$$

Where N(C), N(I), N(O) are normalized values of color, intensity and Orientation. S is the final input to Saliency Map. The maximum saliency map defines the most salient region of an image.

*Interest Point Detection and ROI Extraction:* Interest points are detected using Corner points and saliency Map values. For exact detection of corner of an ROI image, calculate the average of saliency values around each corners in the image be a Saliency Threshold (ST), and only retain the point that saliency value are greater than Saliency Threshold as interest point.

$$ST_i = \frac{1}{3n}\sum_{p=1}^{n}\left(2S_p + \frac{1}{8}\sum_{m=1}^{8} S_{pm}\right) \qquad (3)$$

Assume n is the number of corners in the image. Here the corner points Cp are taken, the saliency value of an image is $S_p$. The surrounding pixel saliency values are represented as $S_{pm}$. This will eliminate the isolated corners and greatly reduce the error and missing detection. Now choose the saliency points which are S $>ST_i$

The Euclidean distance measure is calculated to remove the interest points which more isolated from one another. The distance which are above 100 are considered as the more isolated (based on study analysis). The points that are below threshold will be taken. Using those points, the comprised fromof the rectangular region will be taken as ROI of an image

*SIFT Feature Extraction:* SIFT feature extraction used for describing image patches. The SIFT feature extraction is used as it is invariant to rotation, scaling and illumination changes. The four steps involved in SIFT feature extraction are Scale-space Extrema Detection, Keypoint Localization, Orientation Assignment, Keypoint Matching. The dimension of a SIFT descriptor is 128.

### B. Clustering Approach

For constructing the dictionary, SIFT features extracted from ROI have been used. The tw clustering algorithm has been used here:

1. K-means Clustering
2. Self Organizing Map

*K-Means Clustering:*

K-means is one of the simplest unsupervised learning algorithms. The main idea is to define k centers, one for each cluster. K-means clustering is used for codebook generation. The K-means clustering can be obtained using,

$$J \;= \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2 \qquad (4)$$

Where,

$\left\|x_i^{(j)} - c_j\right\|^2$ is the Euclidean distance between $x_i$ and $c_j$.

n is the number of data points in the $i^{th}$ cluster.

k is the number of cluster centers.

The algorithm works as follows:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

*Self Organizing Map:*

The Self-Organizing Map is the popular one among the neural network models. The Self-Organizing Map is based on unsupervised learning. But little needs to be known about the characteristics of the input data. If we use SOM for clustering data without knowing the class memberships of the input data, then it detects features inherent to the problem. Hence it is also called as Self-Organizing Feature Map. It provides a preserving mapping from the high dimensional space to map units. Map units, or neurons, usually form a two-dimensional lattice and thus the mapping is a mapping from high dimensional space onto a plane. SOM maintains the relative distance between the points. Points that are nearer in input space are mapped to nearby map units.

### C. Image Feature Vector Formation

K-means clustering and SOM clustering approach is used for codebook generation. Now each SIFT feature of an image is matched with each codeword in the codebook and the frequency of the closest matching codeword is formed as a histogram for that image instance. The feature vector dimension of each image is 100 since the number of code words considered is 100.

### D. Multiclass Classification

Support Vector Machines(SVM) is used for training and classifying images in multi label image classification. Here multiple binary SVMs are used to classify each label. SVM are based on the concept of decision planes that define decision boundaries. Support Vector Machine is primarily a classier method that performs classification tasks by constructing hyper planes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. For categorical variables, a dummy variable is created with case values as either 0 or 1 Support Vector Machine is a supervised machine learning algorithm which can be used for both classification and regression challenges. SVM maximizes the distance between hyper plane and the closest sample point. It constructs a model for each category.

The equation of the separating hyperplane is:

$$w_t.\alpha+b=0 \qquad (12)$$

The testing image is predicted with the help of SVM model. The testing image will be classified to the model it fits into. By performing voting based method, the labels are integrated by combining multiple binary SVM classifiers. The predicted labels are checked with manually interpreted labels. The threshold will be taken for voting based method as 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0.

## IV. DATA PREPARATION

The dataset used in this project is MSRC-21 CLASS DATABASE. The MSRC v2 dataset is an extension of the MSRC v1 dataset from Microsoft Research in Cambridge. The dataset is commonly used for full scene segmentation, and may also be used for object instance segmentation, as the current annotation also contains individual object instances next to pure class annotation. Evaluation is done using an average pixel-wise, class-average and PASCAL class-wise accuracy. The dataset includes twenty classes which denote the following landmarks:

- Objects: Aero plane, Building, Book, Bird, Bicycle, Boat, Cat, Car, Cow, Chair, Dog, Flower, Grass, Horse, Human, Mountain, Road, Sky, Street Sign, Sheep, Tree, Water.

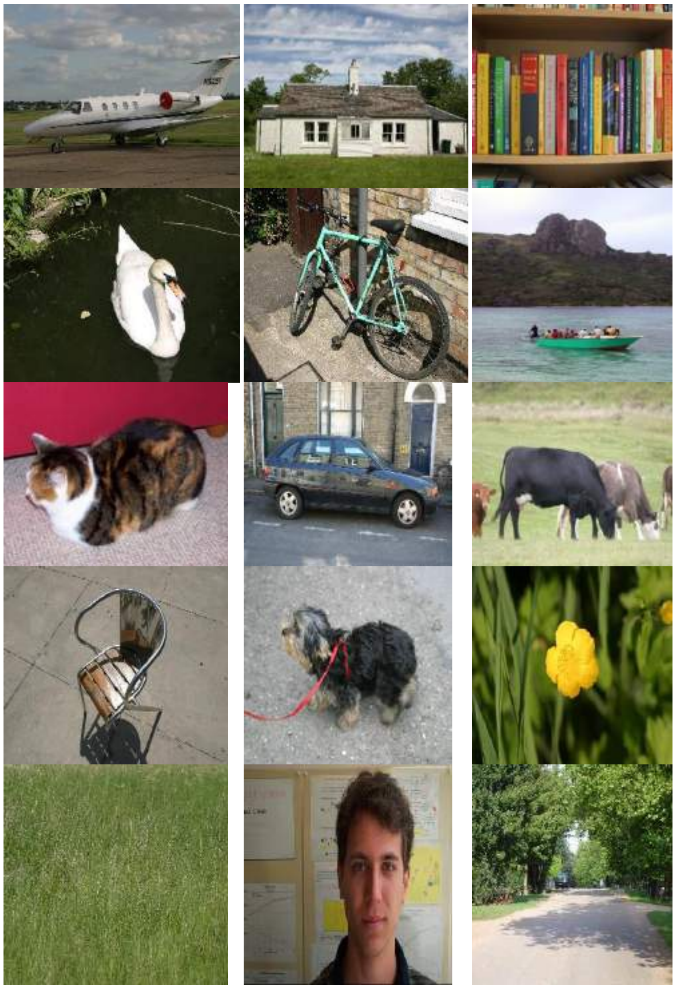


Fig. 2 Examples of MSRC-21 Class Database

## V. RESULTS AND DISCUSSIONS

*Harris Corner Detection:*

The corners are detected using Harris Corner Though detection. Harris corner detector could accurately and effectively detect the corners. Almost all the Corners of the image are detected.

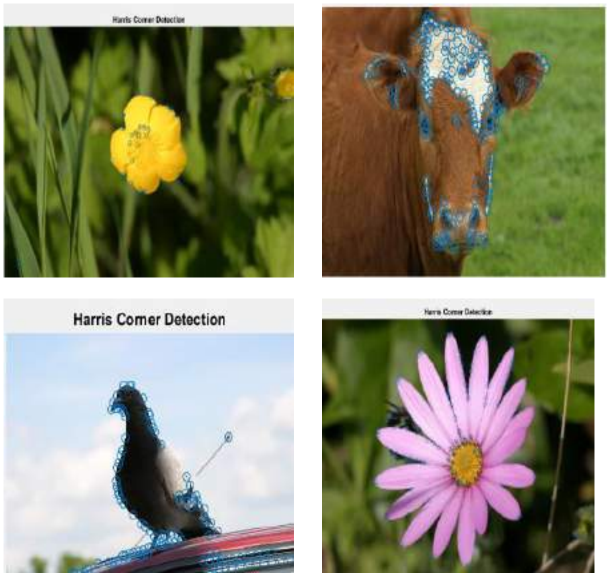


Fig 3.Examples Harris Corner Detection

*Itti-Koch Saliency Map:*

Though Harris corner detector could accurately and effectively detect the corners, some isolated corners are also be detected. To further improve the efficiency, saliency values have been taken. The salient regions provides more information to detect ROI of theimage

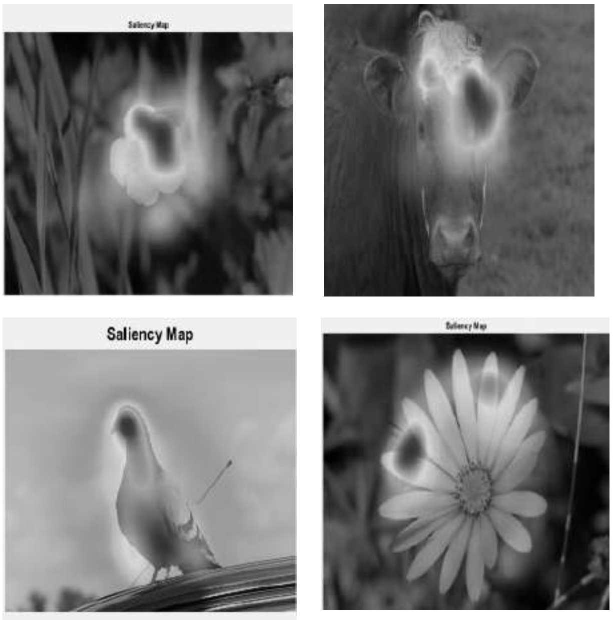
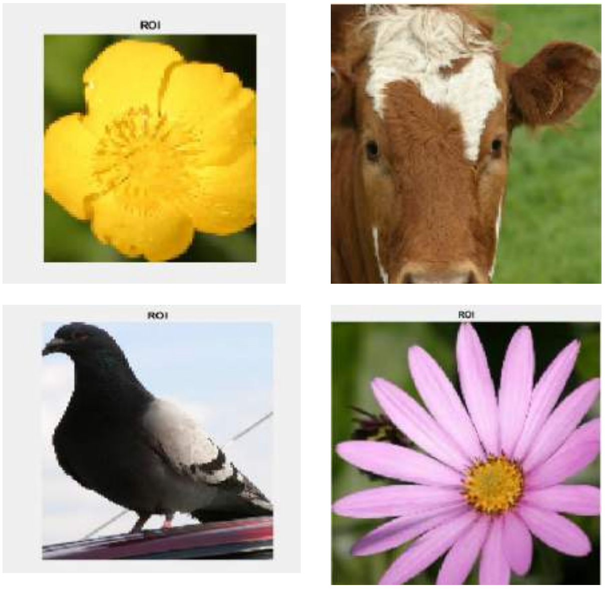


Fig 4. Saliency Map

*ROI Extraction:*

The results obtained from the Corner detection and saliency values provides more information for ROI detection. The combined features of the saliency and Corner detection provide efficient results.



. Fig 5. ROI Extraction

**Comparing the accuracy for classification using K-means and SOM based codebook generation:**

The following TABLE 1 shows the comparison between k-means and SOM clustering approach. Accuracy between k-means and SOM algorithm shows that K-means provides better results than SOM. The voting threshold varies between 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0. For each threshold, k-means shows good performance and high accuracy than SOM

| **Voting Thresh**<br>**Accuracy** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1.0** |
|---|---|---|---|---|---|---|
| **K-means** | **100** | **96.89** | **84.44** | **42.02** | **10.90** | **0.39** |
| **SOM** | **100** | **95.33** | **82.88** | **44.36** | **10.90** | **0** |

TABLEI. Comparison of K-means and SOM.

## VI. FUTURE WORK

In our proposed work, the number of representative regions is fixed for each category. In future, the proposed work maybe enhanced with the clustering the features by constructing the decision trees.

Classification of images is still a research topic which can be used for improving the performance of the proposed system.

## REFERENCES

[1] Y.-T. Zheng et al., “Tour the world: Building a web-scale landmark recognition engine,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog., Jun. 2009, pp. 1085–1092.

[2] J. Hays and A. Efros, “IM2GPS: Estimating geographic information from a single image,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.(CVPR), Jun. 2008, pp. 1–8.

[3] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog., Jun. 2007, pp. 1–8.

[4] Zhu et al., “Visual Object Recognition Using Daisy Descriptor,” in Proc. IEEE Int. Conf. Multimedia Expo., Jun. 2011, pp. 1–6.

[5] T.Deselaers, L.Pimenidis, H.Ney, “Bag-of-visual words models for adult image classification and filtering”, in:Proceedings of the 2008 IEEE 19 th International Conferenceon Pattern Recognition, 2008, pp.1–4.

[6] Wu et al., “Beyond the Euclidean distance: Creating effective visual codebooks using the histogram intersection kernel,” in Proc. IEEE Int. Conf. Comput. Vis., 2009, pp. 630–637.

[7] R.Lienhart, R.Hauke, “Filtering adult image content with topic models”, in: Proceedings of the 2009 IEEE International Conference on Multimedia and Expo (ICME), 2009, pp.1472–1475.

[8] Wang et. al, ”Visual word soft-histogram for image representation”, J.Softw, 2012, pp.1787–1795.

[9] D.Nister, H.Stewenius, “Scalable recognition with a vocabulary tree”, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp.2161–2168.

[10] C. Harris, M. Stephens, A combined corner and edge detector, in: Proceedings of the Alvey Vision Conference, 1988, 15, pp. 50.

[11] L.Itti, C.Koch, Computational modeling of visual attention, Nat. Rev. Neurosci. 2 (3) (2001) 194–203.

[12] D.G.Lowe, “Object recognition from local scale-invariant features”, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999: pp.1150–1157.