Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 2, Issue 5, 2024

HISTON-BASED SEGMENTATION FOR SEMANTIC SCENE CLASSIFICATION AND OBJECT LABELING USING DYNAMIC FEATURES MODELING

Adnan Ahmed Rafique[*1], **Yasir Javaid**[2]

[*1,2]*Dept. of Computer Sciences & IT, University of Poonch Rawalakot*

[*1]adnanrafique@upr.edu.pk, [2]yasi.javaid@gmail.com

DOI: https://doi.org/10.5281/zenodo.15469768

Keywords
Difference of Gaussian, Histon-based segmentation, Ridge detection, Visual Interpretation.

Article History
Received on 23 November 2024
Accepted on 23 December 2024
Published on 30 December 2024

Corresponding Author: *
Adnan Ahmed Rafique

Abstract
Visual interpretation of the scenes precisely requires visionary information on the detection of objects and scene types. The visual interpretation and understanding are fundamental tasks for many applications, such as robotic vision, image-based modeling and augmented reality scene integration. In this paper, a histon based segmentation model is designed that can partition each object into separate regions. Then, the dynamic features of those regions are extracted that extends the key points features, ridge detection and difference of Gaussian histograms. The combination of these features provide a reasonable fact to detect the objects in the scene. Labeling of these objects are performed on the basis of similar features. Finally, the recognizer engine is used to recognize different complex scenes. The experimental results show a better performance on 15 Scene and PASCAL VOC datasets.

INTRODUCTION

Scene understanding is a challenging and most demanding task at present in complex pattern recognition and machine learning applications. Accurate scene understanding is much harder due to inter-dependability of the objects in the scene. These objects are detected using visual features. The quality of the detection is dependent on segmentation of the objects and semantics. To improve the quality scene understanding, there is a need to improve the segmentation and detection process for different applications like robotic vision, navigation, autonomous driving, visual surveillance, image-based modeling and security. These applications demand not only the identification of objects within a scene but also a comprehensive understanding of spatial and semantic relationships among them. Achieving robust scene recognition hinges on the ability to detect, distinguish, and label objects effectively, especially in complex and cluttered environments.

Traditional approaches to scene understanding have primarily relied on global features or handcrafted descriptors, often failing to capture the nuanced dynamics of real-world scenes. To address these limitations, recent efforts have shifted towards more localized and segmented analyses that incorporate both spatial and appearance-based information for precise scene classification.

Object labeling is then conducted by grouping regions with similar dynamic features, allowing for coherent categorization of scene elements. Finally, a recognition engine is applied to classify complex scenes based on the labeled objects, providing a high-level semantic understanding of the visual content. A number of experiments are conducted on

benchmark datasets such as 15-Scene and PASCAL VOC demonstrate the effectiveness and robustness of the proposed method, showcasing improved segmentation accuracy and scene classification performance when compared to conventional approaches.

The remainder of this paper is arranged as follows: Section II provides a review of the existing literature in the field of study. Section III outlines the architectural flow of the proposed method, detailing the stages of preprocessing, segmentation, object detection, feature extraction, labeling, and scene classification. Section IV provides a comparative performance analysis of the proposed approach against existing state-of-the-art scene recognition systems using the 15-Scene and additional benchmark datasets. Finally, Section V summarizes the key findings and concludes the study.

Literature Review

Scene understanding is explored by different researchers using unique models. In [1], P. Espinace et al. proposed a generative probabilistic hierarchical model which include adaptive object search to improve the mutual information. It adaptively searches by converting its probability distribution into an approximation. In [2], J. Shotton et al. proposed a discriminative model which incorporating texture, layout and context information well using conditional random fields. Their model may be used to automatically understand visual information and semantic segmentation. A hierarchical semantic segmentation method [3] is developed by Q. Li et al. that can be used for both scene regions and objects in the image. They used saliency map for object-level labeling and graph-cut for scene-level labeling. Probability to assign a label is predicted, in their approach, on the basis of one vs all classification models. In [4], R. Kachouri et al. employed an unsupervised image segmentation approach to divide images into semantically significant regions. They utilized a merging strategy based on color and texture features to identify different objects and extract scene-related information. Jurio et al. [5] conducted a comparative analysis of various color spaces within a cluster-based segmentation framework, aiming to identify the most effective color representation model. The study evaluated four color spaces—HSV, CMY, RGB, and YUV—and while HSV demonstrated satisfactory performance, the CMY model yielded the highest segmentation accuracy. Beunestado et al. [6] introduced an image segmentation approach that integrates statistical confidence intervals with the conventional Otsu method to enhance segmentation performance. Their technique involves first refining the image using the proposed statistical interval framework, followed by applying the Otsu algorithm, resulting in improved outcomes compared to the traditional Otsu method alone

In this paper, we proposed a framework that uses dynamic features, ridge detection, Difference of Gaussian (DoG) histograms and Support Vector Machine (SVM) to understand and recognize a scene. The proposed model is processed into a sequence of steps i.e. smoothing of image, calculating the histograms of RGB color channels, computing histon, roughness index and then merging of regions to segment the objects in the scene. After object detection through key points and unique dynamic features, labels are assigned to the regions with similar features and one vs all classifier is applied to recognize and understand the scene. Moreover, to evaluate the model, public datasets 15 Scene and PASCAL VOC 2012 are used and significant improvement of scene understanding is achieved over state of the art methods.

System Methodology

The proposed system comprised of number of steps as shown in the Fig. 1. Initially, input images is acquired from the datasets then resizing and smoothing of image/scene is performed. Gaussian filter is used to smooth the image. In the second step, segmentation is performed using histon based thresholding [7-9] and roughness index is measured to produce multiple regions. Then, merging of regions provide clear regions [10] for objecs localization. To extract the features of the objects, at third step, dynamic features like key points calculation [11-12], ridge detection and histogram of DoG is accomplished. Finally, objects are detected on the basis of these extracted features in the fourth step. These objects are then labeled [13] accordingly to regions based having similar regions.

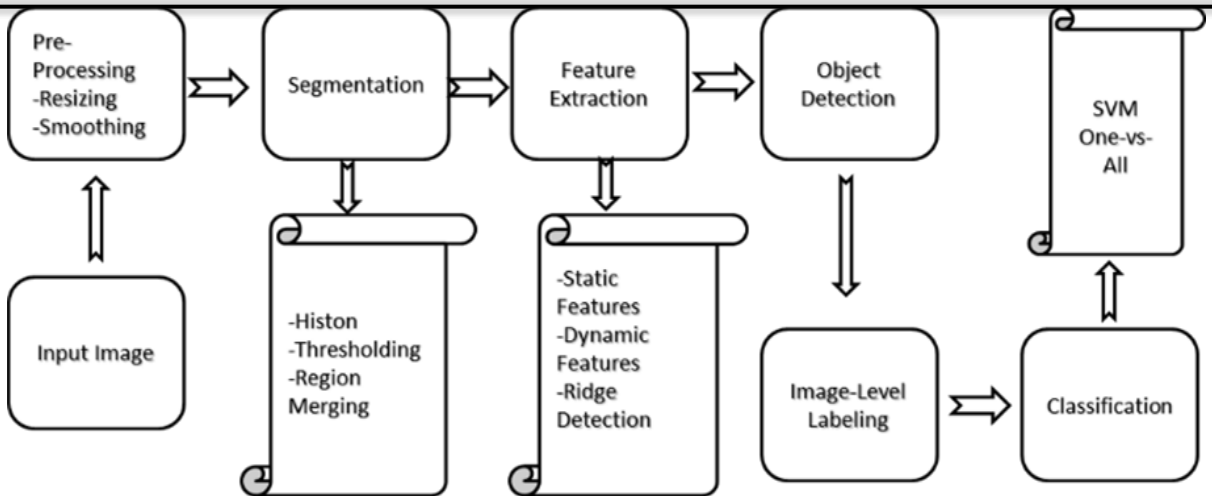


Fig. 1. Scene Classification Framework via Object Detection and Dynamic Features Modeling

A. Pre-processing

The system accepts multiple input formats (e.g. .gif, .png, .jpg, .tiff, .bmp). The images are reshaped into 300 x 300 pixels to process and smoothened using a Gaussian filter with a value of sigma 2, for further processing as shown in the Fig 2.

Fig. 2. Original image on the left and filtered images on the right side

B. Histon-Based Segmentation Model

Histon-based segmentation model [7-9] is applied on the acquired image to detect, localize and recognize the objects in the image. Histograms of the color components RGB are computed as well as the histon. Fig. 3 shows a few examples of histogram of the original image.

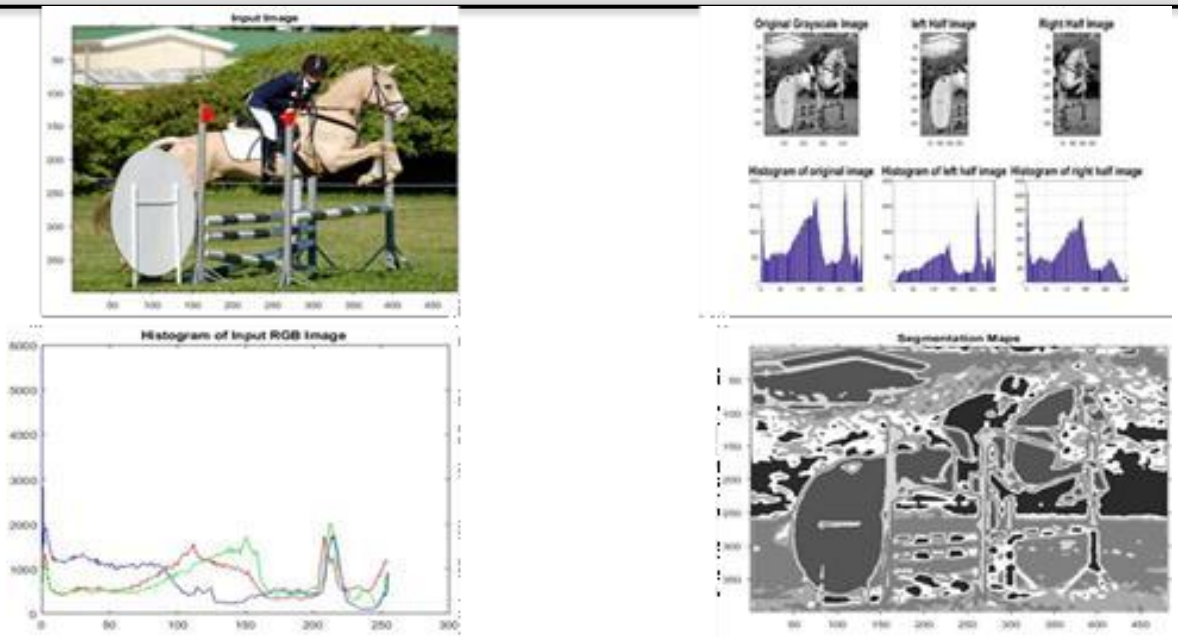


Fig. 3. Segmented Model. Top left shows original image, top right shows color histogram of original image, bottom left shows left half and right half of the image and bottom right shows segmented image

Let a RGB image $\mathcal{J}$ having size $\mathcal{M} \times \mathcal{N}$ comprised of channels red, green and blue, respectively. Compute the histogram of these channels R,G and B using following formulation as;

$$h_i(g) = \sum_{m-1}^{M} \sum_{n-1}^{N} \delta(\mathcal{I}(m,n,i) - g) \text{ for } 0 \leq g \leq \mathcal{L} - 1 \quad (1)$$

and $i = \{R, G, B\}$

where δ is the impulse function and $\mathcal{L}$ represents the color channels intensity levels of each channel. The pixels with intensity g represents the value of each bin. Note that Eucleadian metric between pixel $\mathcal{J}(m, n)$ and piexel $\mathcal{J}(p, q)$ may be calculated as;

$$d(\mathcal{I}(m,n), \mathcal{I}(p,q)) = \sqrt{\begin{matrix}(\mathcal{I}(m,n,R) - \mathcal{I}(p,q,R))^2 + \\ (\mathcal{I}(m,n,G) - \mathcal{I}(p,q,G))^2 + \\ (\mathcal{I}(m,n,B) - \mathcal{I}(p,q,B))^2\end{matrix}} \quad (2)$$

While, process of histon calculation includes neumerous steps between selection of neighbourhood pixels and expanse. If the pixel $\mathcal{J}(m,n)$ is surrounded by $\mathcal{P}\times\mathcal{Q}$ neighbourhood then the total distance of all pixels can be calculated as;

$$d_{\mathcal{I}}(m,n) = \sum_{p \in P} \sum_{q \in Q} d(\mathcal{I}(m,n), \mathcal{I}(p,q)) \quad (3)$$

Similarly, if expanse is greater than , the domain or region will be the sam of the neighbourhood pixels. Now, a matrix $\mathbb{X}$ of size $\mathcal{M} \times \mathcal{N}$ is defined in such a way that (m, n) is written as;

$$\mathbb{X}(m,n) = \{1, d(m,n) < \text{expanse} \quad (4)$$

0 otherwise

Finally, the Histon is defined as;

$$\mathcal{H}(g) = \sum_{m=1}^{M} \sum_{n=1}^{N} (1 + \mathbb{X}((m,n)) \delta(\mathcal{I}(m,n,i) - g) \quad (5)$$

$for\ 0 \leq g \leq \mathcal{L} - 1$ and $i = \{R, G, B\}$

The neighborhood pixels are selected in the form of 3x3 or 5x5 or even bigger windows size. These pixels are used in the computation of histon by using following formulae as;

$$\rho = nm - n \quad (6)$$

where ρ is the total number of pixels, n is the number of rows of window and m represents the number of columns of the window for neighborhood pixels. The expanse is the span of similar color

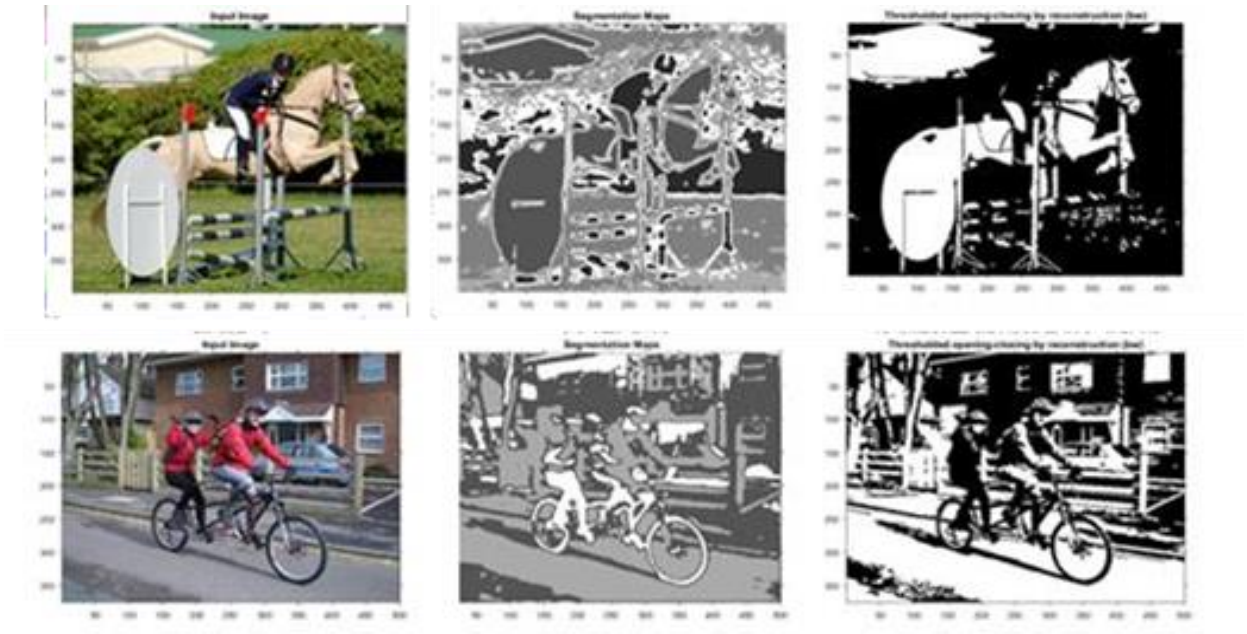
domain or region. This span or radius may be calculated by the given equation as;

$$(a - R)^2 + (b - G)^2 + (c - B)^2 = \gamma^2 \quad (7)$$

where γ is the radius of the neighborhood region with color intensities R (Red), G (Green) and B (Blue). After the computation of expanse, the roughness index is measured as;

$$\mathcal{P}_i(g) = 1 - \frac{\hbar_i(g)}{\mathcal{H}_i(g)} \; for \; 0 \leq g \leq \mathcal{L} - 1 \quad (8)$$

The roughness index [7] highlights the maxima and minima within the graph. However, segmentation is strongly dependent on the selection of these maxima and minima. Therefore, to select a maxima, the distance between two maxima's ideally should be greater than 10. While, the height of maxima should be greater than ⅕ of the roughness index. After the significant maxima's are selected, the minima are obtained by finding the minimum values between two maxima's. During segmentation process, region merging is performed after obtaining clusters from maxima's and minima. As a result of clustering, several small regions are generated which is known as over-segmentation. These small regions are merged, where the number of pixels in a region are a few, with the closest large regions. Region merging consists of two steps. In the first step, merging is performed on those small regions whose number of pixels are small enough are merged with their nearest large regions. The second step merges the nearest regions whose distance is less than a specific threshold to form a single region. Fig. 4 demonstrates the segmented images.



(a) (b) (c)

Fig. 4. (a) Original image, (b) segmented and (c) region merging

C. Dynamic Features Modeling

To extract the useful cues, we computed the dynamics of the segmented objects. Each segmented object is figured with some key points. Then, Euclidean distance is calculated by using these key points. Now, perimeter and area of the polygons (triangles) formed by connecting these key points is calculated. Fig. 5 represents the key point's computation of segmented objects.

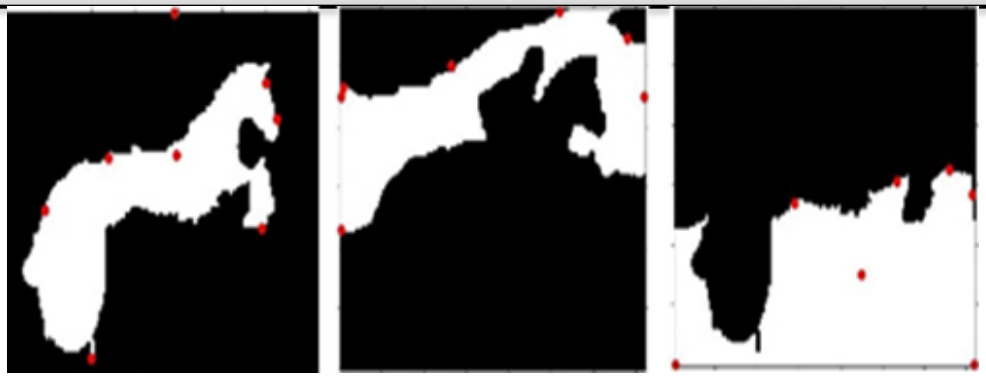


Fig. 5. Few examples of Key point's calculation

The Euclidian distance D between any of two points on the segmented object may be calculated mathematically as:

$$\|D\| = \sqrt{(p_2 - q_2)^2 + (p_1 - q_1)^2} \quad (9)$$

where p, q are points taken from extreme points to calculate Euclidian distance D between these points.

The perimeter P of the polygons (triangles), triangles formed by connecting any three points, may be formulated mathematically as:

$$P = \overline{XY} + \overline{YZ} + \overline{XZ} \quad (10)$$

where P symbolizes perimeter, $\overline{XY}$ is the measure of the length of one side of the triangle, $\overline{YZ}$, $\overline{XZ}$ are the measures of the lengths of the other two sides of the triangle under consideration.

The area inside the interconnected points (three) and forming a triangle $\triangle XYZ$ may be computed mathematically as:

$$\text{Area} = \sqrt{\bar{p}\,(\bar{p} - \bar{XY})(\bar{p} - \bar{YZ})(\bar{p} - \bar{ZX})} \quad (11)$$

where $Area$ means the area of the triangle, p is half of the perimeter P, $\overline{ZX}$, $\overline{YZ}$ *and* $\overline{ZX}$ are the three sides of the triangle, respectively.

To further improve feature strength, a Difference of Gaussian (DoG) filtered image is obtained after applying the difference of Gaussian. Difference of Gaussian [11] is obtained by taking the difference of two Gaussian Functions. These DoG functions have two different values for σ, which control the degree of smoothness in images. Mathematically, DoG may be expressed as:

$$h(k, l) = h_1(k, l) - h_2(k, l) \quad (12)$$

where $h_1(k, l)$ and $h_2(k, l)$ are Gaussian functions and these functions can be formulated as:

$$h_1(k,l) = e^{x^2}/2\,{\sigma_1}^2$$
$$h_2(k,l) = e^{x^2}/2\,{\sigma_2}^2$$

where $\sigma_2 < \sigma_1$

by putting values of (13) and (14) in (12)

$$h(k,l) = \frac{e^{x^2}}{2\,{\sigma_1}^2} - \frac{e^{x^2}}{2\,{\sigma_2}^2}$$

Then, a histogram of the filtered image is plotted and shown in Fig. 6.

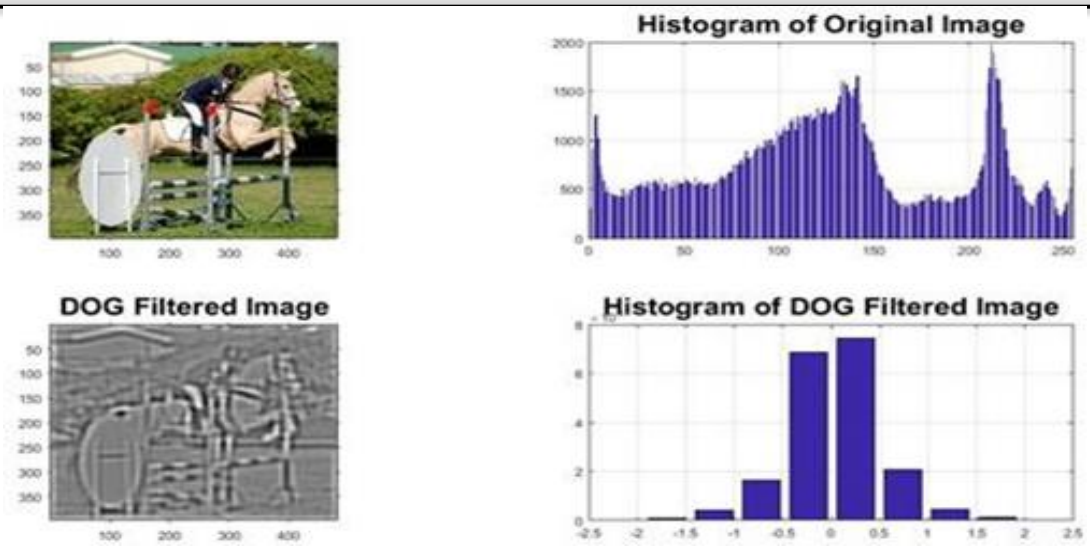


Fig. 6. Histogram of DoG filtered image vs original image

D. Object Detection & Labeling

After the computation of dynamic features and ridge feature detection, object detection is performed. On the basis of these features [12-16] , the similar regions are labeled as shown in Fig. 7. The similar regions with similar features are assigned with the same label.

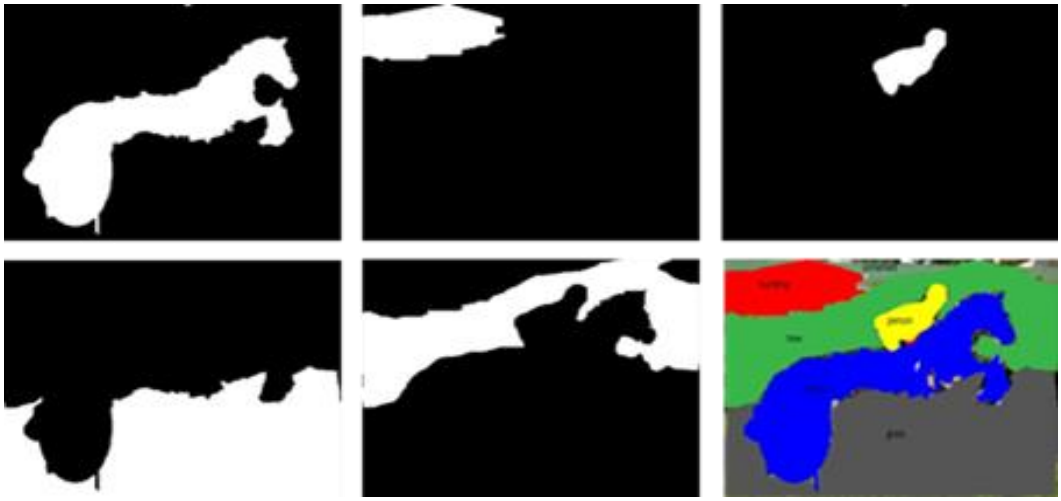


Fig. 7. Examples of multiple objects detection and labeling

E. Scene Recognition via SVM

Support Vector Machine (SVM) is a supervised [20-21] machine learning technique that analyze the data for classification. There are different variants or kernels works under SVM technique. Here One-vs-All approach is followed to recognize the class of the scene/objects in the image that divides the multiclass problem into multiple binary class problems [17]. The object/scene is classified under a specific class label unless the features are accepted for the same class SVM and rejected by all the other class SVM's. Therefore, it leaves the objects or regions unnamed if there is an ambiguous situation for the objects features. The flow of SVM is explained in Fig. 8. Feature vector is computed from the dynamic features, ridge detection and DoG is forwarded to SVM for assigning the class labels to each class.

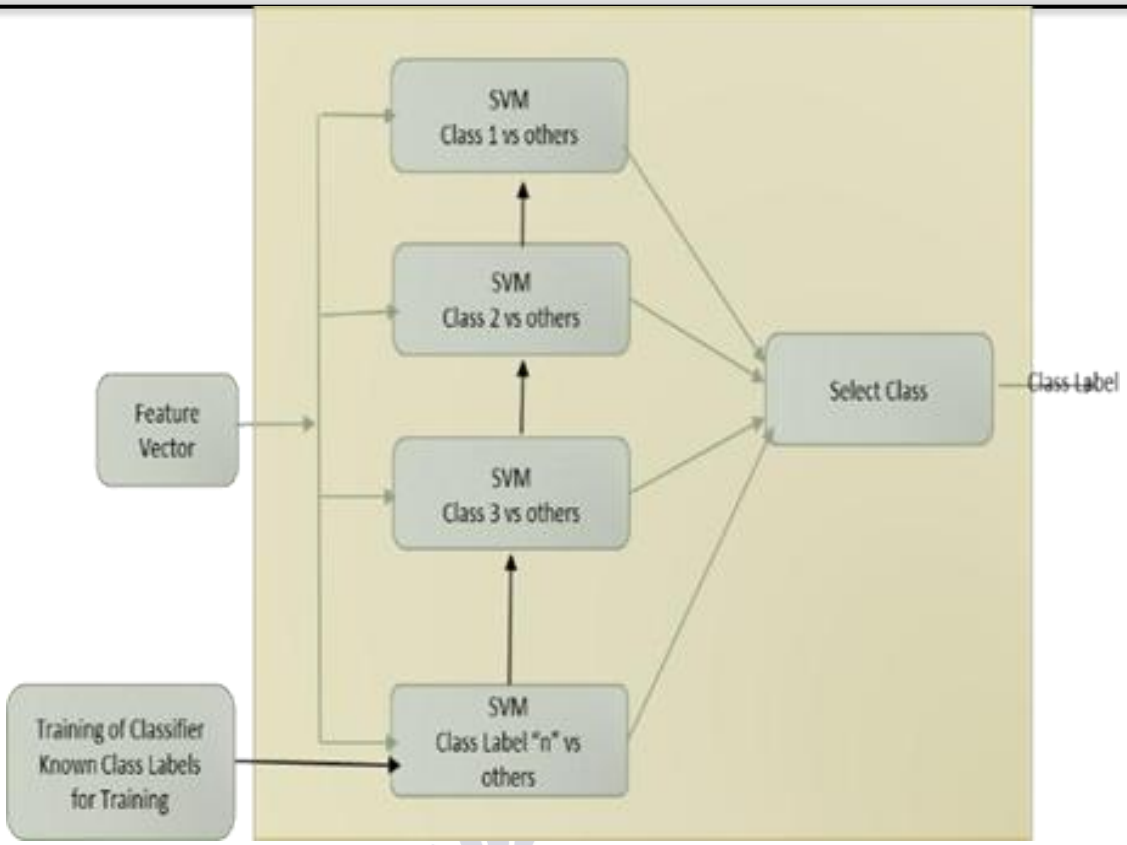


Fig. 8. Flowchart of SVM working

Performance Evaluation

All experiments were conducted on a system equipped with an Intel Pentium Core i3 processor (2.0 GHz) and 6 GB of RAM. Training and testing were carried out using a cross-validation approach on the PASCAL VOC dataset as well as the 15-Scene natural and indoor dataset. Table I presents the segmentation outcomes, benchmarked against the annotated ground truth of the PASCAL VOC and 15-Scene natural and outdoor datasets. The results demonstrate notably strong performance.

A. Dataset Description of PASCAL VOC Dataset

The PASCAL VOC dataset [22] comprises 20 object categories, including: person, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, and TV/monitor. The training set consists of 11,530 images, encompassing 27,450 annotated objects within defined Regions of Interest (ROIs), along with 6,929 corresponding segmentation annotations.

B. Dataset Description of 15-Scene Dataset

The 15-Scene dataset was initially developed by Oliva and Torralba [18], introducing eight scene categories. Subsequently, Fei-Fei and Perona [17] contributed five additional classes, and the final two categories were incorporated by Lazebnik et al. [21]. The complete set includes the following scene types: mountain, forest, kitchen, living room, office, bedroom, store, industrial, tall building, inside city, street, highway, coast, open country, and suburb. The images have an approximate resolution of 250×300 pixels, with each category containing between 210 and 410 images. This dataset encompasses both indoor and outdoor scenes, offering a diverse representation of environments.

C. Performance Evaluation of PASCAL VOC and 15-Scene datasets

Table 1 illustrates the segmentation performance against ground truth annotations from the 15-Scene and PASCAL VOC datasets.

TABLE I. COMPARISON OF SEGMENTATION WITH ANNOTATED IMAGES

Class Labels	Segmentation Accuracy % (PASCAL VOC)	Class Labels	Segmentation Accuracy % (15 SCENE)
Horse	85.6	Bedroom	81.3
Bird	86.7	Forest	83.8
Person	82.5	Mountain	85.6
Cow	87.2	Street	82.8
Sheep	87.5	Tall building	86.3
Aeroplane	78.4	Industrial	79.5
Cat	75.7	Highway	86.2
Dog	76.1	Living room	69.3
Boat	82.5	Office	74.1
Bus	70.8	Store	77.5
Train	81.2	Inside city	62.3
Car	86.2	Coast	58.5
Motorbike	67.5	Open country	62.0
Bicycle	68.6		
Bottle	71.8		
Chair	75.4		
Dining table	62.8		
Potted plant	60.3		
Sofa	62.1		
Monitor/TV	80.4		
Overall Accuracy	**76.46 %**	**Overall Accuracy**	**76.09 %**

Table II demonstrates the scene classification accuracy of the PASCAL VOC datasets, having a mean accuracy of 74.86 %. While, recognition accuracy of 75.26% on the 15 Scene dataset, respectively.

TABLE II. EVALUATION OF THE PROPOSED MODEL USING SVM

Class Labels	Classification Accuracy % (PASCAL VOC)	Class Labels	Classification Accuracy % (15 SCENE)
Horse	84.12	Bedroom	82.50
Bird	85.28	Forest	75.90
Person	83.56	Mountain	83.85
Cow	87.80	Street	83.65
Sheep	89.25	Tall building	85.00
Aeroplane	76.01	Industrial	80.90
Cat	71.66	Highway	87.75
Dog	76.10	Living room	70.21
Boat	72.50	Office	75.50
Bus	72.75	Store	78.70
Train	72.50	Inside city	61.28
Car	85.95	Coast	55.50

Motorbike	66.00	Open country	57.68
Bicycle	65.66		
Bottle	75.50		
Chair	66.65		
Dining table	61.78		
Potted plant	60.80		
Sofa	64.25		
Monitor/TV	79.10		
Overall Accuracy	**74.86 %**	**Overall Accuracy**	**75.26 %**

Conclusion

An efficient scene recognition method is presented in this study, utilizing a Histon-segmented model that incorporates color histogram analysis in a multi-step framework. The histogram thresholding and roughness index are applied to segment the objects in different regions. Then, region merging is applied to merge similar and closest regions with similar features. Segmented objects are labeled having similar features and then SVM classifier is executed to recognize the scene. We validated our system on PASCAL VOC dataset and 15 Scenes dataset as well.

REFERENCES

P. Espinace, T. F. Kollar and N. Roy, "Indoor Scene Recognition Through Object Detection", in proc. of IEEE International Conference on Robotics and Automation, pp. 1-9, Jun. 2010.

J. Shotton, J. Winn and C. Rother, "TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context", in Int. J Comput Vis, pp. 1-22, Dec. 2007.

Q. Li, A. Liang, H. Liu, "Hierarchical semantic segmentation of image scene with object labeling", EURASIP Journal on Image and Video Processing, pp. 1-10, Mar. 2018.

R. Kachouri. M. Soua. M. Akil, "Unsupervised Image Segmentation Based on Local pixel Clustering and Low-Level Region Merging", in proc. of 2nd international conference on advanced technologies for signal and image processing, pp. 1-7, Mar. 2016.

Jurio, A., Pagola, M., Galar, M., Lopez-Molina, C., & Paternain, D. (2010). A comparison study of different color spaces in clustering based image segmentation. In Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications: 13th International Conference, IPMU 2010, Dortmund, Germany, June 28–July 2, 2010. Proceedings, Part II 13 (pp. 532-541). Springer Berlin Heidelberg.

Buenestado, P., & Acho, L. (2018). Image segmentation based on statistical confidence intervals. Entropy, 20(1), 46.

M. M. Mushrif and A. K. Ray, "Color image segmentation: Rough-set theoretic approach", in proc. of Pattern Recognition Letters, pp. 483-493, Nov. 2007.

M. M. Mushrif and A. K. Ray, "A-IFS Histon Based Multithresholding Algorithm for Color Image Segmentation", IEEE Signal processing letters, pp.168-171, Mar. 2009.

X. D. Yue, D. Q. Miao, N. Zhang, Q. Wu and L. B. Cao, "Multiscale roughness measure for color image segmentation", in proc. of Information Sciences, pp. 93-112, Mar. 2012.

M. Yan, J. Cai, J. Gao, L. Luo, "K-means Cluster Algorithm Based on Color Image Enhancement for Cell Segmentation ", in proc. of 5th international conference on biomedical engineering and informatics, pp. 295-299, Oct. 2012.

Z. Niu, G. Hua, X. Gao, and Q. Tian, "Context Aware Topic Model for Scene Recognition", in proc. of IEEE computer society conference on computer vision and pattern recognition, pp.1-8, Jun. 2012.

K.-K. Maninis, S. Caelles, J. P.-Tuset and L. Van Gool, "Deep Extreme Cut: From Extreme Points to Object Segmentation", in proc. of computer vision and pattern recognition (CVPR), pp. 616 – 625, 2018.

K. P. Risha, C. Kumar, C.S Sindhu, "Difference of Gaussian on Frame Differenced Image", in proc. of "International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering", pp. 92-95, Feb. 2016.

S. Liu, S. Yan, T. Zhang et al., "Weakly Supervised Graph Propagation Towards Collective Image Parsing", in proc. of IEEE transactions on multimedia, pp. 361-373, Apr. 2012.

T. Bagautdinov, A. Alahi, F. Fleuret1, P. Fua1 and S. Savarese, "Social Scene Understanding:End-to-End Multi-Person Action Localization and Collective ActivityRecognition" , in proc. of computer vision and pattern recognition, pp. 1-10, Nov. 2016.

X. Li and Y. Guo, "Latent Semantic Representation Learning for Scene Classification", in proc. of 31st internatiioanl conference on machine loearning, pp. 1-9, 2014.

L. F.-Fei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories", in proc. of IEEE computer society conference on computer vision and patteren recognition, pp. 1-8, Jul. 2005.

Ruiz-Sarmiento, Jose-Raul, Cipriano Galindo, and Javier Gonzalez- Jimenez, "A survey on learning approaches for Undirected Graphical Models. Application to scene object recognition," International Journal of Approximate Reasoning83 , pp. 434-451 2017.

Y. Lee. Y. Lin and G. Wahba, "Multicategory Support Vector Machines", in proc. of NSF, pp-1-15, 1999.

M. Baji, Dr. I. S. Prabha, "Implementation of Optical Flow, Sliding Window and SVM for Vehicle Detection and Tracking", in proc. of International Journal of Innovative Research in Science, Engineering and Technology, pp. 18452-18458, Sep. 2017.

S. Lazebnik, C. Schmid and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories", in poc. Of of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Oct. 2006.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International journal of computer vision, 88, 303-338.