Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-13717-y

Maximum entropy scaled super pixels segmentation for multi-object detection and scene recognition via deep belief network

Adnan Ahmed Rafique[1,2] · Munkhjargal Gochoo[3] · Ahmad Jalal[1] · Kibum Kim[4]

Received: 25 January 2021 / Revised: 3 March 2022 / Accepted: 24 August 2022

Abstract
Recent advances in visionary technologies impacted multi-object recognition and scene understanding. Such scene-understanding tasks are a demanding part of several technologies such as augmented reality based scene integration, robotic navigation, autonomous driving and tourist guide applications. Incorporating visual information in contextually unified segments, super-pixel-based approaches significantly mitigate the clutter, which is normal in pixel wise frameworks during scene understanding. Super-pixels allow customized shapes and variable size patches of connected components to be obtained. Furthermore, the computational time for these segmentation approaches can significantly decreased due to the reduced number of super-pixel target clusters. Hence, the super pixel-based approaches are more commonly used in robotics, computer vision and other intelligent systems. In this paper, we propose a Maximum Entropy scaled Super-Pixels (MEsSP) Segmentation method that encapsulates super-pixel segmentation based on an Entropy Model and utilizes local energy terms to label the pixels. Initially, after acquisition and pre-processing, image is segmented by two different methods: Fuzzy C-Means (FCM) and MEsSP. Then, to extract the features from these segmented objects, the dynamic geometrical features, fast Fourier transform (FFT), blob extraction, Maximally Stable Extremal Regions (MSER) and KAZE features are extracted using the bag of features approach. Then, to categorize the objects, multiple kernel learning is applied. Finally, a deep belief network (DBN) assigns the relevant labels to the scenes based on the categorized objects, intersection over union scores and dice similarity coefficient. The experimental results regarding multiple objects recognition accuracy, precision, recall and F1 scores over PASCAL VOC, Caltech 101 and UIUC Sports datasets show a remarkable performance. In addition, the evaluation of proposed scene recognition method over these benchmark datasets outperforms the state of the art (SOTA) methods.

Keywords Bag of features · Deep belief network · Entropy-scaled segmentation · Super-pixels

✉ Kibum Kim
kibum@hanyang.ac.kr

Extended author information available on the last page of the article.

Published online: 20 September 2022

1 Introduction

Scene understanding in intelligent systems [39] is very challenging task and it has gained much interest in the research community particularly in relation to vision based systems. These systems have achieved remarkable attention due to their potential in real-world applications [20] like autonomous driving [6], drone targeting systems [50], healthcare [55], artificial eye for blind people [4], global positioning system based location finder [44], human activity recognition [29] and unmanned aerial vehicles. Understanding scenes correctly and analyzing behaviors within them are still critical tasks and challenging which demand reasoning from low-level features to high-level scene labels. These scene labels are used to classify items out of large numbers of scene classes as they are described directly by using scene descriptors such as GIST, and CENTRIST. Therefore, scene understanding and recognition methodology are generally divided into two stages: i) converting features into objects and ii) converting objects to scenes. Such methodology recognizes the scenes by concatenating the low-level features with high-level object detectors. Then, detected objects are analyzed for their semantic relationships. Finally, the scenes are labeled by the information attained from semantic relationship analysis in order to better discriminate each scene class.

The key task of the object detection and scene recognition system (ODSR) is to recognize and label the scene based on the recognition of all the objects present in the scene. Numerous researchers have proposed different models to understand and accurately label the scene, however, still there are several challenges such as complex backgrounds, occlusion, scaling, and illumination variations, that need to be addressed to get the best possible recognition accuracy. Therefore, we proposed a robust model that deals with these challenges by segmenting images via unique Maximum Entropy scaled Super-Pixel (MEsSP), merging multiple local and global features by incorporating the bag of features technique fused with multiple kernel learning to categorize the objects. Hence, compared to conventional approaches, our novel approach produces superior performance results.

In this paper, we proposed a novel idea that integrates the MEsSP Segmentation with a bag of features to categorize the objects and then by employing deep belief network (DBN), predicts the scene labels. Our method is comprised of the following steps: Primarily, we remove the noise, smooth the image and apply filters using the ideal low pass filter. In the next step, two different segmentation techniques such as FCM and MEsSP are applied. The segments obtained are further examined for the extraction of dynamic multiple features using a bag of features approach in the third step. Then, we incorporate multiple kernel learning for the categorization of objects. Finally, based on object's categories, IoU scores and dice similarity coefficient (DSC), DBN predicts the scene label.

Our contributions are as follows:

1. Two approaches namely, FCM and MEsSP segmentation for single or multi-objects are applied. The MEsSP method performs better compared to FCM, so MEsSP based results are forwarded for the extraction of features using a bag of features technique.
2. To extract geometrical features, an algorithm is proposed that extracts the key points of an object and then connects these points to form different geometrical shapes for geometrical features extraction.
3. Multiple kernel learning is applied to the features extracted from the bag of features technique for the categorization of the objects.
4. DBN is applied to predict a scene based on categorized objects, their IoU scores and DSCs.

Table 1 Analysis of different techniques and limitations in the literature

Ref.	Datasets	Methods	EM	Limitations
Ahmed et al. [1]	MSRC, Corel-10K, CVPR 67 Indoor Scene Dataset	Fuzzy C-Means, Mean Shift, and Multi-Class Logistic Regression	Classification Accuracy	Comparison with the traditional models is not presented, Computational time and its comparisons are missing.
Asif et al. [7]	WRGB-D Object Dataset, Cornell Grasp Object Dataset	STEM-CaRF Framework, CNN Features Extraction, Hierarchical Cascaded Forests	Instance and Category recognition accuracy	Computationally complex and computational time not compared with other state-of-the-art methods.
Rashid et al. [48]	Caltech-101, Birds Database, Butterflies Database, CIFAR-100	Multi-Layer Deep Features Fusion, VGG19, and Inception V3	Accuracy and FNR	The quality of features is compromised when inputting low-quality images.
Zia et al. [67]	WRGB-D Dataset	VGGNet, 3D CNN and VGG3D	Accuracy	Comparison with the traditional models is not presented, Computational time and its comparisons are missing.
Hussain et al. [26]	Caltech 101	HOG, LBP, Inception V3, JEKNN	Accuracy, FNR, Time	A gape of 18% accuracy between SVM and Co-KNN classifiers.
Xia et al. [58]	MIT Indoor 67, Scene 15, and UIUC Sports	Grad-CAM and WS-AM	Accuracy	Computational time is not compared with other existing methods.

References (Ref.), evaluation metrics (EM)

The Remainder of the paper is organized as follows: Related work on scene understanding and recognition is discussed in Section 2. Section 3 addresses the proposed methodology of single/multi object detection for scene understanding using bag of features. Section 4 presents experimental evaluation of segmentation and recognition accuracy using DBN compared with other state of the art (SOTA) methods. Section 5 presents the conclusion and future work.

2 Related work

ODSR are demanding tasks which have gained a lot of interest over the last couple of decades in multimedia and visual technologies. The current era of these studies and technologies has manipulated the ODSR in diverse environments such as action recognition, GAIT-recognition, pedestrian tracking and the diagnosis of diseases using biomedical imaging.

Multi-object recognition is more complicated since one image consists of several instances with a cluttered environment and complicated backgrounds in various locations. Numerous researchers have focused on multi-object recognition while some others have used traditional systems to explore scene understanding and labeling. These traditional systems compute different features to recognize the objects and to classify scenes. A detailed overview of these systems is described in Tables 1 and 2.

In the above table, multiple authors did not consider the traditional models while comparing their proposed approach. A few researchers used low-quality features while building their model for classification or recognition. Similarly, some of the researchers did not compute the computational time to verify the effectiveness of their proposed model while some others only consider the accuracy as an evaluation metric. Based on the above analysis, we have proposed a novel method that incorporates the effective segmentation mechanism integrated with a diverse set of features to categorize the objects and recognize the scenes. We also evaluated our model on several metrics including accuracy, precision, recall, F1 Score, computational time, and compare the results with other well-known SOTA methods as well as traditional models. The comparison results demonstrated the effectiveness of our model.

3 Design framework for object detection and scene recognition

Initially, the proposed ODSR System acquires data as input for the pre-processing steps. Secondly, for segmentation, two approaches, MEsSP and FCM, are employed. To extract features effectively, using bag of features, different features are computed such as geometrical features, FFT, blobs, MSER, and KAZE. Then, the computed feature vector via bag of features is forwarded to MKL for the categorization of the objects. After categorization, IoU scores and DSC are computed. Finally, DBN is applied to these object categories, IoU score and DSC for scene recognition and understanding. The schematic view of our proposed system is presented in Fig. 1.

3.1 Data acquisition and preprocessing

During preprocessing, an image is acquired as an RGB color image with different file formats. Initially, all the images are converted into 213x320 pixels. These resized images are then transformed from RGB to grayscale. Then, the median filter [3, 46] is used to remove

Table 2 Main contributions of various researchers on mulit-object and scene recognition

Reference	Main Contributions
A. Ahmed et al. [1]	They proposed a novel method to recognize multi-objects in a scene based on object categorization. They segmented the image by employing improved FCM and mean-shift segmentation techniques. Later, local descriptors are extracted and multiple kernel learning is applied for object categorization. Additionally, they incorporated intersection over union (IoU) scores and multi-class logistic regression for scene classification.
U. Asif et al. [7]	They presented a hierarchical cascade forests model that uses computed probabilities at different phases of an image based on which unknown objects and classes are recognized. They introduced an objective function that extracts features from the point clouds of RGB-D objects for object recognition and grasp detection.
M. Rashid et al. [48]	They designed a model based on a Deep Learning (DL) technique that recognizes objects by selecting multi-layer deep features. They used three steps namely, Inception V3 for feature extraction, DL architecture for feature fusion, and Deep CNN for recognition, to accurately recognize the objects. Moreover, they also optimized their system by applying logistic regression controlled entropy variance.
S. Zia eta al. [67]	Suggested a solution for object recognition using a deep convolutional neural network (CNN). They designed a hybrid 2D/3D CNN that used a pre-trained network. They train their CNN over a small RGB-D dataset. They combined the features extracted from both RGB only and depth only models, in their hybrid model, to produce more accurate results.
L. Zhang et al. [62]	They proposed a framework that relies on the traversing of total pixels to make the status of objects understandable as an object bank. Scene classification is performed by a normalized sum match kernel.
X. Song et al. [52]	Following the semantic manifold model, they extended the model to recognize the scenes by employing semantic illustrations and confined contextual relationships. They trained the model to discriminate multiple complex scenes, which have numerous classes.
R. Kachouri et al. [33]	They introduced a segmentation method that depends on the unsupervised learning technique (i.e., clustering) to identify the significant regions. They used color and textures as features to separate out the clusters or regions for the retrieval of scene information.
H. Zhao et al. [26]	They proposed a Pyramid Scene Parsing Network that takes an image as input and extracts the features by using a pre-trained convolutional neural network. Then, by using pyramid pooling, they accumulate visual features via a multiple-level pyramid. Finally, they predict scene labels with the help of a convolution layer.
N. Hussain et al. [64]	Introduced a hybrid mechanism that combines features based on both classical and deep learning techniques. After making the database balanced, two classical features pyramid histograms of gradients and central symmetric local binary patterns are fused with the deep learning features which are extracted from the pre-trained Convolutional Neural Network model. To choose the best features, they used joint entropy with the K-nearest neighbor. J. Feng et al. proposed a probabilistic topic model that uses latent Dirichlet allocation to extract the features to make scene semantic recognition. They used deeper training latent Dirichlet allocation for features training.
S. Xia et al. [58]	They discussed unique regions rendering using neural nets and designed a combined architecture in which they build a semantic manifold on top of multi-scale CNN. Then, they employed Markov random fields to concatenate various features such as multi scales and spatial relations to detect the appropriate scenes.

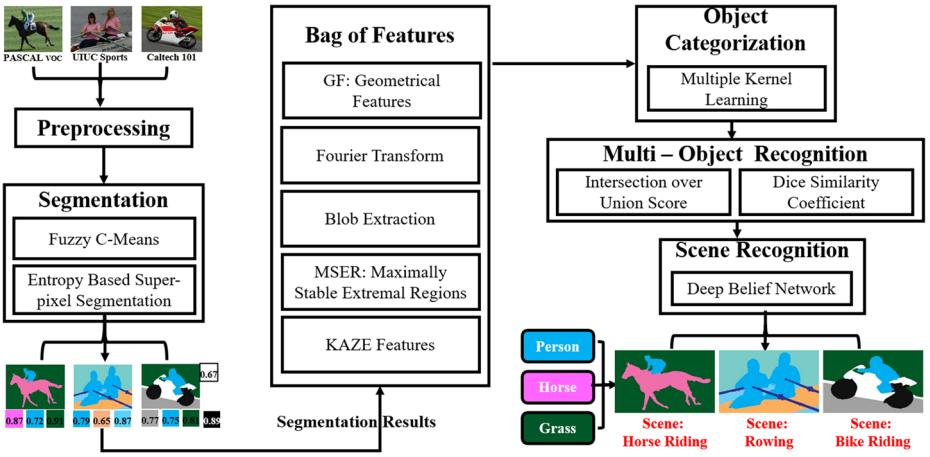

Fig. 1 Flow architecture of the proposed ODSR Model

the noise from the image and an ideal low pass filter is applied to smooth the image. The following transfer function of ideal low pass filter is used for smoothing:

$$F(u, v) = \begin{cases} 1 & \text{if } S(u, v) \leq S_0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $S(u, v)$ denotes the distance of (u, v) to the center of pixels in an object.

3.2 Segmentation framework

Numerous visual technologies exploit image segmentation as a fundamental step to analyze scenes. To perform the segmentation process, images are partitioned into homogeneous regions. Thus, the role of segmentation is significant in different applications including autonomous vehicles, surveillance systems [27], medical imaging, and virtual reality. Various algorithms for object segmentation have been proposed by different researchers, such as region-growing [31], watershed transform [31], k-means clustering [15], graph-cuts [66], conditional random fields [63] and other advanced deep learning (DL) methods. To find the optimal solutions of objects for segmentation during scene recognition, we propose two different segmentation methods for the object detection, scene understanding and recognition.

3.2.1 Maximum entropy scaled super-pixels segmentation

As a graph partitioning problem, we take into consideration clustering. To divide the image into K super-pixels, we seek a graph configuration that has K associated sub-graphs and boosts the objective function implied [37].

To represent an image over a graph ($\mathbb{G} = V, E$) having vertices that denote the pixels, and edge weights represent similarities in terms of a similarity matrix. We need to pick out group of edges ($\mathfrak{A}$) to a limit of E so that the resulting equation ($\mathbb{G} = V, \mathfrak{A}$). We note that a self-loop is maintained by each edge within the graph, but it is not sufficient to sort out the problem of graph partition. When the edge of the related vertices is not included in $\mathfrak{A}$, we

improve the edge weight for the self-loop to such a degree that each vertex has a constant incidence.

To get dense and consistent clusters, we use the entropy scale of the random walk (RW) criteria on the graph. The distribution of RW remains unchanged and probability set functions $prob_{x,y}$ are symbolized as follows:

$$prob_{x,y}(\mathfrak{A}) = \begin{cases} \frac{W_{x,y}}{W_x} & \text{if } x \neq y \text{ and } e_{x,y} \in A \\ 0 & \text{if } x \neq y \text{ and } e_{x,y} \notin A \\ 1 - \frac{\sum_{y:e_{x,y} \in A} W_{x,y}}{W} & \text{if } x = y \end{cases} \tag{2}$$

Therefore, the entropy scale of the RW on $(\mathbb{G} = V, \mathfrak{A})$ may be formulated as a function:

$$\mathfrak{H}(\mathfrak{A}) = -\sum_x \mu_x \sum_y prob_{i,j}(\mathfrak{A}) log(prob_{x,y}(\mathfrak{A})) \tag{3}$$

We employ a balancing function that inspires clusters with equivalent dimensions. Let $\mathfrak{A}$ set of edges , $N_{\mathfrak{A}}$ as linked pixels (connected components), and $Z_{\mathfrak{A}}$ is the membership of diffused clusters. For example, suppose the graph partitioning for set of edges A be $S_{\mathfrak{A}} = S_1, S_2, S_3, \ldots, S_{N_{\mathfrak{A}}}$. Then the diffusion of $Z_{\mathfrak{A}}$ can be expressed as:

$$p_{Z_{\mathfrak{A}}}(i) = \frac{|S_i|}{|V|}, i = 1, ..., N_{\mathfrak{A}} \tag{4}$$

and the balancing function may be written as:

$$B(\mathfrak{A}) = \mathfrak{H}(\mathfrak{A}) - N_{\mathfrak{A}} = -\sum_i p_{Z_{\mathfrak{A}}}(i) log(p_{Z_{\mathfrak{A}}}(i)) - N_{\mathfrak{A}} \tag{5}$$

Entropy $\mathfrak{H}(\mathfrak{A})$ supports clusters having similar dimensions; while $N_{\mathfrak{A}}$ support a couple of clusters.

The objective function based on composition of rate of entropy and balancing function, encourages versatile, symmetric, and organized clusters. By optimizing the objective function with respect to set of edges, the clustering is attained:

$$max_{\mathfrak{A}} H(\mathfrak{A}) + B(\mathfrak{A}) \tag{6}$$

subject to $\mathfrak{A}$ is subset of E and $N_{\mathfrak{A}}$ is greater than or equal to K, where weight of balancing term is considered as 0. Figure 2 illustrates the results of MEsSP.

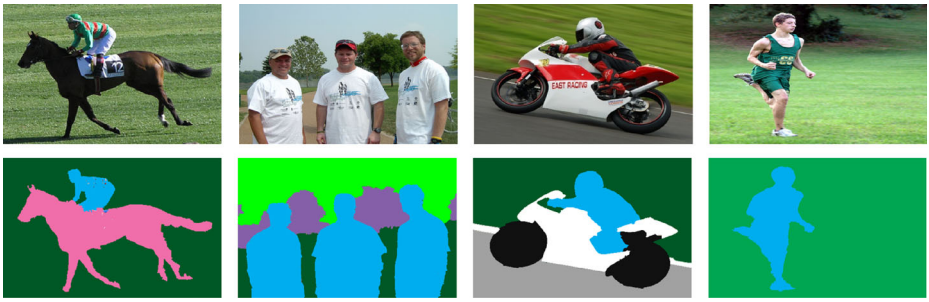

Fig. 2 A few examples of MEsSP Segmentation on PASCAL VOC 2012

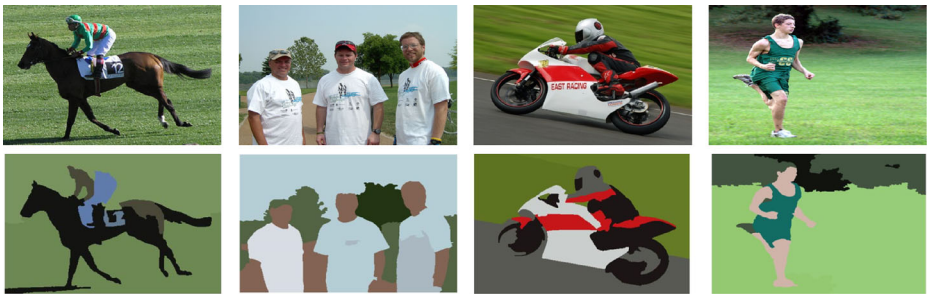

Fig. 3 FCM segmentation on PASCAL VOC 2012. The upper row represents the original images and segmentation results are illustrated in the lower row

3.2.2 Fuzzy c-means segmentation

FCM is a technique where clustering is performed in such a way that a single element belongs to two or more clusters [1, 41]. Hence, based on these coinciding elements, pixels that belong to more than one cluster are supposed to depict the fuzzy logic. The FCM segmentation process is accomplished by optimization of the objective function through iterations. In addition, the membership $p_k l^R$ and cluster center b_k are updated accordingly. The weighted sum of the distance between the cluster center and the elements of the corresponding fuzzy cluster is computed by performance index $F_M(P, Q)$ and it is formulated as follows:

$$F_M(P, Q) = \sum_{k=1}^{c} \sum_{l=1}^{m} p_{kl}^R |a_l - b_k|^2 \tag{7}$$

Table 3 Computational time (CT) comparison for MEsSP and FCM segmentation algorithms over PASCAL VOC 2012

Class	MEsSP	FCM	Class	MEsSP	FCM
HS	91.0	101.2	TN	103.6	141.5
BD	107.1	123.5	CR	109.1	139.2
PR	95.7	99.0	MB	131.3	157.9
CW	98.4	107.3	BC	111.5	123.1
SH	132.0	155.2	BL	137.8	154.0
AP	75.5	91.1	CH	81.0	98.5
CT	127.2	129.3	DT	94.5	115.6
DG	83.1	97.9	PP	117.9	129.9
BT	119.2	135.3	SF	106.2	137.0
BS	160.9	175.2	TV	127.0	151.7
Average CT of the MEsSP = **107.13 s**					
Average CT of the FCM algorithm = **131.69 s**					

HS = horse; BD = bird; PR = person; CW = cow; SH = sheep; AP = aeroplane; CT = cat; DG = dog; BT = boat;

BS = bus; TN = train; CR = car; MB = motorbike; BC = bicycle; BL = bottle; CH = chair; DT = dining table;

PP = potted plant; SF = sofa; TV = television

where the number of clusters is represented by c, data points m and real numbers R in k^{th} cluster illustrating the fuzziness of resulting cluster, the association of a_l pixels in k^{th} cluster can be expressed p_{kl}^R, and b_k is cluster center;

$$p_{kl}^R = \frac{1}{\sum_{h=1}^{c} (\frac{|a_l - b_k|}{|a_l - b_k|}^2)^{\frac{1}{R-1}}} \tag{8}$$

the p_{kl} belongs to [0, 1] for $k = [1, ..., c]$

$$b_k = \frac{\sum_l = 1^m p_{kl}^R X_j}{\sum_{k=1}^{m} p_{kl}^R} \tag{9}$$

Hence, if $F_M(P, Q)$ results in the smallest distance from the relevant pixel to the cluster center, the higher membership value is assigned to the relevant pixel. The FCM results are shown in Fig. 3.

The two segmentation techniques namely, MEsSP and FCM algorithms are compared and evaluated based on segmentation accuracies along with computational time. FCM execution time is more than that of our proposed MEsSP. Additionally, our proposed algorithm produces comparatively better segmentation results than FCM, therefore, we used MEsSP results for ensuring the experiments. Table 3 describes the comparison of computational time between the MEsSP and FCM on PASCAL VOC 2012. While, Tables 4 and 5 demonstrates segmentation accuracies over PASCAL VOC 2012 and Caltech 101 datasets respectively. Based on the computational time and accuracy of both the segmentation techniques, MEsSP is more effective and accurate. Therefore, the results from MEsSP are considered for further processing i.e. feature extraction, object categorization and scene recognition.

3.3 Bag of features methods

To describe the image, multiple features are extensively used in object detection, recognition and classification models. Various researchers considered effective feature extraction techniques including ensemble machine learning approach [24]. Here, an idea referred to as bag of features [43] has been used in various works. This has been derived from a widely known idea used in text retrieval techniques known as bag-of-words [61]. This phenomenon can be carried out in the case of images by splitting the image into meaningful patches, as

Table 4 Segmentation accuracy comparison for MEsSP and FCM algorithms over PASCAL VOC 2012 dataset

	HS	BD	PR	CW	SH	AP	CT	DG	BT	BS
MEsSP	95.23	92.71	93.88	94.54	91.63	88.87	89.91	94.19	86.55	88.25
FCM	93.65	91.11	93.06	93.88	90.91	87.25	90.15	93.75	83.57	82.66
	TN	**CR**	**MB**	**BC**	**BL**	**CH**	**DT**	**PP**	**SF**	**TV**
MEsSP	84.25	91.51	89.46	85.73	90.19	94.45	90.87	95.65	82.15	90.95
FCM	80.55	91.68	86.79	84.25	87.45	92.56	89.32	94.92	80.78	90.11

HS = horse; BD = bird; PR = person; CW = cow; SH = sheep; AP = aeroplane; CT = cat; DG = dog; BT = boat; BS = bus; TN = train; CR = car; MB = motorbike; BC = bicycle; BL = bottle; CH = chair; DT = dining table; PP = potted plant; SF = sofa; TV = television

Table 5 Segmentation accuracy comparison for MEsSP and FCM algorithms over Caltech 101 dataset

	HS	KR	AN	MB	KT	CN	CM	DK	FY	HW
MEsSP	89.35	84.81	95.27	96.36	86.75	94.11	92.50	94.22	94.68	92.67
FCM	88.71	81.05	90.89	92.67	85.19	91.95	91.40	93.29	90.88	91.56
	AP	**ET**	**ZB**	**LP**	**HC**	**TG**	**TR**	**DR**	**BR**	**DN**
MEsSP	89.80	88.75	85.25	89.67	91.70	93.35	91.55	92.84	90.07	87.66
FCM	85.57	87.21	81.37	86.97	89.65	92.77	90.19	89.46	85.97	80.25

HS=Horse, KR=Kangaroo, AN=Anchor, MB=Motorbike, KT=Ketch, CM=Camel, DK=Duck, FY=Ferry, HW=Hawksbill, AP=Airplane, ET=Elephant, ZB=Zebra, LP=Lamp,HC=Helicopter, TG=Tiger, TR= Tree, DR=Deer, BR=Bear, DN=Dolphin

performed in segmentation. Many techniques can perform the segmentation process. Moreover, the detection and extraction of features have been commonly used in problems like segmentation. Here, we have extracted multiple features to employ a bag of features technique such as fast Fourier transform (FFT), blob extraction, geometrical features, MSER, and KAZE features.

3.3.1 Fast fourier transform

To represent an image into a frequency domain, the image is converted into real and imaginary parts by applying FFT to the image [42]. We can compute the FFT of the image mathematically as follows:

$$F(u,v) = \sum_{x=0}^{x-1}\sum_{y=1}^{y-1} f(x,y)e^{-}(i*2*\pi(u\frac{x}{X}+v\frac{y}{Y})) \tag{10}$$

$$f(u,v) = \frac{1}{X.Y}\sum_{x=0}^{x-1}\sum_{n=1}^{N-1} F(x,y)e^{-}(i*2*\pi(u\frac{x}{X}+v\frac{y}{Y})) \tag{11}$$

where $f(x,y)$ represents the pixel at position (x,y), whereas $F(u,v)$ is the function to represent the image in the frequency domain pertaining to the position u and v. Here $X \times Y$ represents dimension of the image, and i is $sqrt(-1)$. Figure 4 depicts the FFT on some images from the Pascal VOC 2012 dataset.

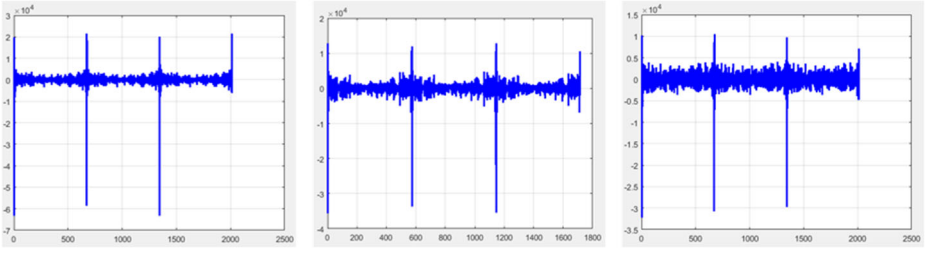

Fig. 4 FFT features extraction over PASCAL VOC 2012 images

3.3.2 Blob extraction

A connected cluster of pixels in a specific shape is known as a blob [53, 59]. To group these pixels into a cluster, a morphological closing operation is performed that results in the detection of the object as a combination of blobs (See Fig. 5). The most popular blob detection technique uses Laplacian of Gaussian (LoG). An image I(u,v) after convolving with LoG is formulated as:

$$g(u, v, t) = \frac{1}{2\pi t} e^{-} \frac{u^2 + v^2}{2t} \tag{12}$$

a scale space representation on a explicit scale t is furnished $S(u, v, t) = g(u, v, t) * I(u, v)$. After applying the Laplacian operator,

$$\nabla^2 S = S_{uu} + S_{vv} \tag{13}$$

is determined, that returns robust dark blobs of radius $r = \sqrt{2t}$ (for a two-dimensional image,$r = \sqrt{dt}$ for a d-dimensional image) and returns powerful negative bright blobs of similar dimensions. However, a problem is observed when operator is applied at single scale. Therefore, to overcome the problem, to dynamically capture the blobs in the iamge, multi-scale blob detection is performed. So, scale-normalized Laplacian operator is employed to achieve the dynamic blobs that can be expressed mathematically as:

$$\nabla^2_{norm} S = t(S_{uu} + S_{vv}) \tag{14}$$

where $\nabla^2_{norm} S$ is to detect scale-space max/min. Hence, to determine a three dimensional scale-space volume $S(u, v, t)$ based on given two dimensional image $I(u, v)$. The pixel is considered for bright blob if the value is greater than all the pixels surrounded by that specific pixel and similarly the pixel is considered for dark blob if the value is smaller than the other pixels. Accordingly, concurrent interest points $(\widehat{u}, \widehat{v})$ and scales $\widehat{t}$ are picked up by

$$\widehat{u}, \widehat{v}, \widehat{t} = argminmax_{(u,v,t)}(\nabla^2_{norm} S)(u, v, t) \tag{15}$$

3.3.3 Geometrical features

Local geometrical features are computed over the segmented objects. Initially, four extreme points are computed for each of the detected and segmented objects. These four extreme points are extreme top, extreme bottom, extreme left and extreme right points. These extreme points are further used to calculate the local geometrical features Euclidean distance [28]. Primarily, four extreme points of each segmented object are computed and plotted on four extreme locations of the foreground object as illustrated in Algorithm 1.

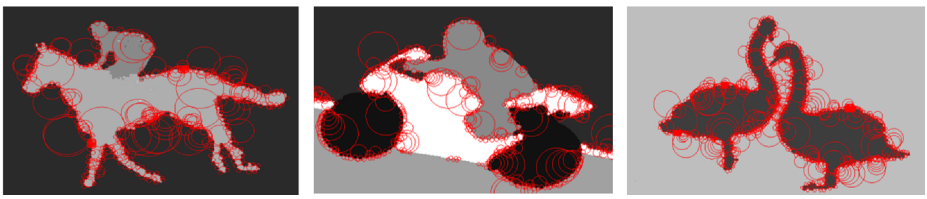

Fig. 5 Blob features extraction on horse riding, bike riding and ducks examples

```
Input: Segmented objects
Output: Dynamic geometric features of objects.
[ROW COL] = size (seg_object) /* returns number of rows and number of col. in the object*/
/* Find these extrempoints */
[extreme_top , extreme_right, extreme_bottom, extreme_left, mid_pixel]=extremes(seg_object)
for 1: ROW   /* read all pixel's pix in the scene/image */ and mark spot points
    for 1: COL   /* Extract points*/
    if pix (u,v) = mid_pixel Then spot point on pix(u,v) end
  else-if pix (u,v) = extreme_left Then spot point on pix (u,v) end
  else-if pix (u,v) = extreme_right Then spot point on pix (u,v) end
  else-if pix (u,v) = extreme_top Then spot point on pix (u,v) end
  else-if pix (u,v) = extreme_bottom Then spot point on pix (u,v) end
  else Then pix (u,v) is other than extreme point pixels end
     end
  end
/* compute distance and drawing lines between these extreme points by using array of points
for all images */
ArrayPoints = [extreme_top, extreme_right, extreme_bottom, extreme_left]
mid_pixel;
for j = 1 : ArrayPoints []
    if (ArrayPoints [j] => 4) Then ArrayPoints [j+1]=0; end
    dist (ArrayPoints [j], ArrayPoints [j+1])   /*1st iteration show distance between extreme top
and right points */
   draw_line(ArrayPoints [j], ArrayPoints [j+1]) end
for j = 1 : ArrayPoints []
    dist (ArrayPoints [j], mid_ pixel)   /* 1st iteration show distance between extreme top and
mid_ pixel */
   draw_line(ArrayPoints [j], mid_ pixel)
  end
return extreme_left, extreme_right, extreme_top, extreme_bottom, mid_ pixel, and distance
between these points
```

Algorithm 1 Dynamic geometric features computation.

The Euclidean distance d is measured between each pair of extreme points. If ext_{top} and ext_{bot} are two points for the computation of Euclidean distance, then the Euclidean distance may be formulated as:

$$\|d\| = \sqrt{(ext_{top_x} - ext_{bot_x})^2 + (ext_{top_y} - ext_{bot_y})^2} \tag{16}$$

where d is the Euclidean distance from $extreme_{top}$ to $extreme_{bot}$, the x-coordinate of the $extreme_{top}$ is denoted by ext_{top_x}, y-coordinate of $extreme_{top}$ is represented by ext_{top_y} while x, y-coordinates of $extreme_{bot}$ are ext_{bot_x} and ext_{bot_y} respectively. The Euclidean distance between different extreme points for the objects in PASCAL VOC 2012 dataset are shown in Fig. 6.

3.3.4 KAZE features

To extract KAZE features [2], an imput image is convolved with Gaussian kernel for noise reduction. An image gradient histogram is computed from the convolved image and a

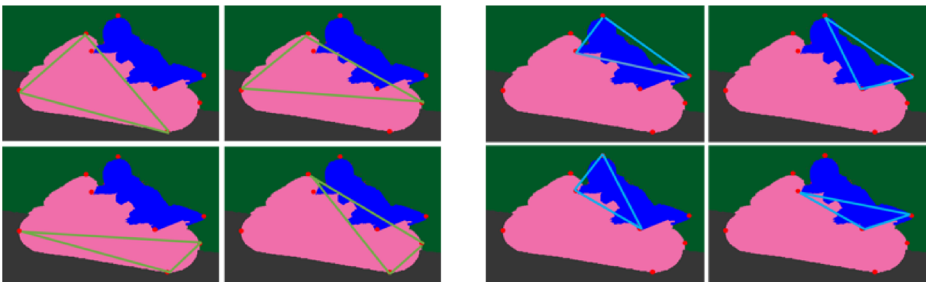

Fig. 6 Geometrical Features extraction on the bike riding example from the PASCAL VOC 2012 dataset

contrast parameter p is attained through a programmed process. With the known values of contrast parameter and evolution time, the nonlinear scale space is established using:

$$d^{j+1} = (I - t_j + 1 - t_j \sum_{i=1}^{m} A_i(d^j))^{-1} d^j \tag{17}$$

Furthermore, to detect points of interest, the response of scale normalized determinant of Hessian at several levels is computed:

$$S_{Hess} = \sigma^2 (d_{uu} d_{vv} - d_{uv}^2) \tag{18}$$

where the 2^{nd} order horizontal derivatives is represented by d_{uu} and the vertical derivative is illustrated as d_{vv}, while the 2^{nd} order cross derivative can be expressed as d_{uv}. Figure 7 demonstrates the KAZE features.

3.3.5 Maximally Stable Extremal Regions (MSER) features

In MSER [40], the unique regions are demarcated by the feature of the intensity function in the region and on the contour of the objects in the image, called extremal property. Figure 8 represents the MSER extracted features on some examples of PASCAL dataset.

Image I is a mapping $I : D$ belongs to $Z^2 \rightarrow S$ external region and it can be defined if:

- S is fully ordered
- there exists a proximity $A \subset D \times D$

Segment Q is a component which is connected within D and $\forall p, q$ belongs to $Q \exists$ a sequence $p, a_1, a_2, \ldots, a_n, q$ which fulfills $pAa_1, a_1Aa_2, \ldots, a_nAq$

The external edges of a segment may be expressed $\partial Q = q \in D \setminus Q : \exists p \in Q : qAp$ extremal region $Q \subset D$ is a region for every p belongs to Q, q belongs to $\partial Q : I(p) > I(q)$

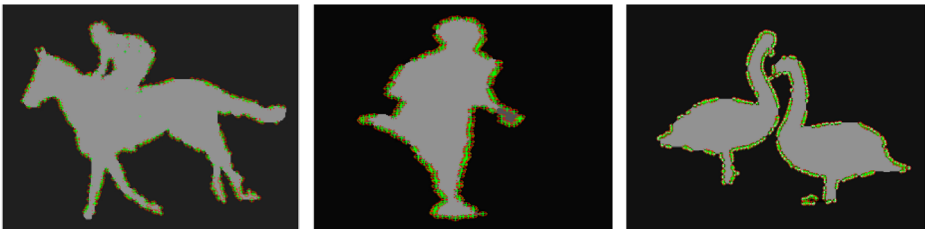

Fig. 7 KAZE Features extraction from horse riding, person running and ducks example

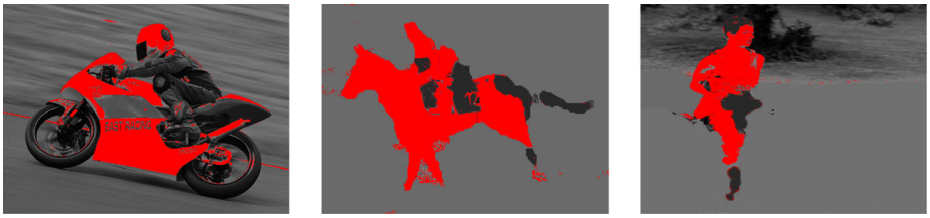

Fig. 8 MSER Features extraction from bike riding, horse riding and person running from the PASCAL VOC 2012 dataset

or $I(p) < I(q)$ Let $Q_1, \ldots, Q_{i-1}, Q_i, \ldots$ is a pattern of nested extremal regions i.e. Q_i is a subset of Q_{i+1}). Extremal region Q_{i*} is maximally stable

if: $q(i) = \frac{Q_{i+\nabla} \setminus Q_{i-\nabla}}{Q_i}$ has a local minimum in $i*$. ∇ belongs to S is the parameter of the method.

4 Object categorization

To categorize the objects in the complex scene based on extracted multiple features via bag of features, a MKL [60] technique is employed. Object categorization in our model is depicted in Fig. 9. During the process of categorization, i image having multiple objects with the number of n clusters of unique colors, is considered for bag of features and region R_g of the particular image i. Now, to compute the descriptor from the features set F_i as $f_{(Rg)} : F_i$. We can formulate $f_{(Rg)}$ as follows:

$$CEN_n = \frac{1}{|n|} \sum_i \sum_j F_{jni} \tag{19}$$

$$M_n = \frac{1}{|n|} \sum_i \sum_j (F_{jni} - CEN_n)(F_{jni} - CEN_n)^T \tag{20}$$

$$M_{i,n} = \sum_j (F_{jni} - CEN_n)(F_{jni} - CEN_n)^T - M_n \tag{21}$$

where a cluster center is denoted by CEN_n,total features in n number of clusters for all the images belongs to a class are expressed as $|n|$, F_{jni} are features that belongs to cluster n for an image i while M_n is the mean of the bag of features that belongs to clusters n. Mi, n computes the descriptors of an image i. Then $M_{i,n}$ is transformed into $V_{i,n}$ vector. The combination of vectors $V_{i,n}$ is used to compute the descriptor of image i for all clusters n.

$$V_i = (V_1, V_2, V_3, ..., V_n) \tag{22}$$

To obtain the multiple regions with bounding boxes for extraction of descriptors, deformable part model is incorporated. Then, based on maximum scores assumed by the detector, various regions of foreground objects are extracted. After the extraction of specific region R_g from the object, a kernel function Ker_{Rg} is used to measure the similarities between extracted region R_g and images i, j . The kernel function is defined as:

$$Ker_{Rg}(i, j) = (f_{(Rg)}(F_{Rgi}), f_{(Rg)}(F_{Rgj})) \tag{23}$$

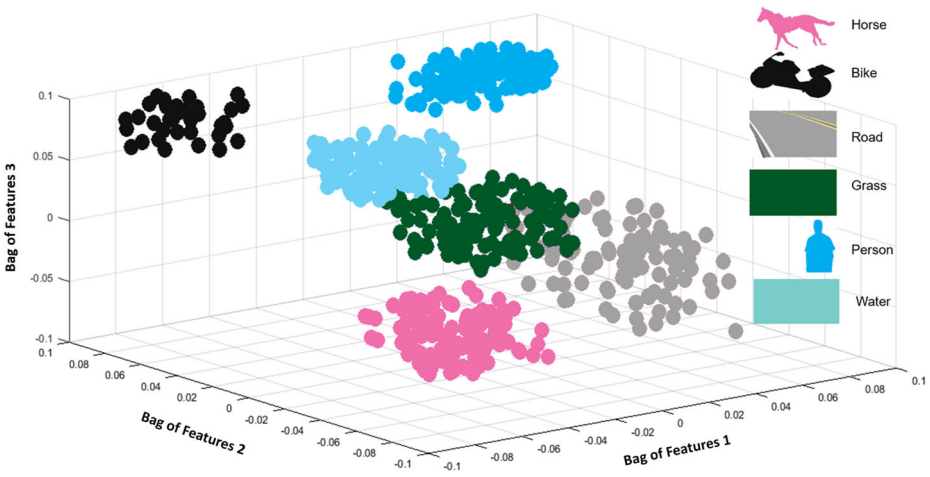

Fig. 9 Multiple kernel learning for object categorization on the PASCAL VOC 2012 dataset

Although, to attain the similarity upon whole the image, multiple regions of the image are considered. Hence, the similarity for all the regions may be formulated as:

$$Ker(i, j) = \sum W_{Rg} Ker_{Rg}(i, j) \tag{24}$$

where weights of regions are represented by W_{Rg} .

4.1 Intersection over Union (IoU) score

The IoU score is to specify the level of accuracy regarding how perfectly the objects are predicted by our model [32, 49]. The proposed system assigns an IoU score to every object in the complex scene. The IoU function is examined based on multiple objects , the locations of the objects and finally predicted objects . Mathematically we can express this as follows:

$$IoU_{score} = \frac{1}{nc} \sum_{nc=1}^{nc} Score_{iou}^{nc}(O_p, O_i) \tag{25}$$

where C denotes the number of classes and $Score_{iou}^{nC}$ is described as:

$$Score_{iou}^{nc}(O_p, O_i) = \frac{\sum_j \in V^1[O_p - k \wedge O_i - nc]}{\sum_j \in V^1[O_p = k \wedge O_i - nc]} \tag{26}$$

where $\forall i \in 1_t \ldots tCVi \in V$ and pixels in the images are represented by V. The indicator function $1_{[O_p=k \wedge O_i - nc]}$ returns 1 if $[O_p = k \wedge O_i - nc]$ is true, otherwise it returns 0. The ratio of sum of pixels represents the value of $Score_{iou}^{nC}$ as the IoU score of the object. The computed IoU score of the objects over three benchmark datasets are graphically represented in Fig. 10.

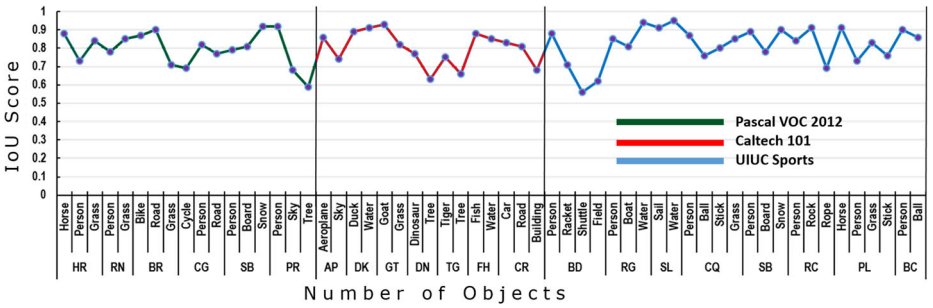

Fig. 10 IoU score for three benchmark datasets

4.2 Dice Similarity Coefficient (DSC)

To measure the performance of a model, DSC is an evaluation metric that computes the overlapping area of ground truth images with the images produced after applying the proposed method [16, 57]. If A and B are two sets of objects such that A belongs to the ground truth objects and B belongs to detected objects using proposed model, then, the DSC can be formulated as follows:

$$D_{sc} = 2 \times \frac{|A \cap B|}{|A| + |B|} \tag{27}$$

where $|A|$ and $|B|$ are the number of objects in sets A and B and $|A \cap B|$ represents common objects of both the sets. The DSC for three benchmark datasets is shown in Table 6.

4.3 Scene recognition

After multiple object categorization, the labeled information is further used for scene understanding and recognition. To achieve scene recognition, the following significant approaches are employed, (1) IoU scores, (2) DSC and (3) categories of the objects. IoU and DSC are computed to analyze the foreground objects, while in order to recognize the multi-objects in a scene, categories are fused with IoU and DSC scores to accomplish scene recognition. In order to recognize the scene, various researchers have applied different classifiers. Every classifier has its merits and demerits according to the problem under consideration. In [21], a novel PCA-whale optimization-based Deep CNN is used for classification. Multi-view PointNet is proposed in [30] for 3D scene understanding. Another multiclass classifier is used for the Malware classification in [56] that uses CNN architecture to detect Malwares based on those color images that are extracted through the raw malware binaries.

4.3.1 Deep belief network

In machine learning, a DBN [34, 38] belongs to one of the deep neural networks. It is comprised of multiple hidden layers or latent variables. These layers are connected to each other, whereas the units of the same layer are not connected with one another. DBN fine-tune the weights of units by employing the following equation:

$$W_{i,j}(t+1) = W_{i,j}(t) + n\frac{log(p(v))}{W_{i,j}} \tag{28}$$

Table 6 DSC for all the objects of the three benchmark datasets

Pascal VOC 2012		Caltech 101		UIUC Sports	
Object	DSC	Object	DSC	Object	DSC
Horse	0.912	Horse	0.799	Racket	0.670
Bird	0.769	Kangaroo	0.775	Shuttle	0.592
Person	0.871	Anchor	0.601	Person	0.831
Cow	0.927	Motorbike	0.881	Boat	0.688
Sheep	0.883	Ketch	0.722	Water	0.930
Airplane	0.750	Cannon	0.763	Sky	0.895
Cat	0.895	Camel	0.850	Tree	0.755
Dog	0.871	Duck	0.826	Snow	0.682
Boat	0.694	Ferry	0.667	Board	0.655
Bus	0.717	Hawksbill	0.612	Ball	0.728
Train	0.562	Airplane	0.755	Horse	0.912
Car	0.735	Elephant	0.637	Sail	0.717
Motorbike	0.881	Zebra	0.681	Stick	0.561
Bicycle	0.570	Lamp	0.754	Rock	0.457
Bottle	0.531	Helicopter	0.562	Grass	0.671
Chair	0.703	Tiger	0.728		
Dining table	0.624	Tree	0.639		
Potted plant	0.511	Deer	0.680		
Sofa	0.618	Bear	0.657		
Television	0.626	Dolphin	0.527		

where the probability of a visible vector is expressed as $(p(v)$ and produced by

$$p(v) = \frac{1}{Z} \sum_{h} e^{-} E(v, h) \tag{29}$$

Z is the partition function, and $E(v, h)$ is the energy function assigned to the state of the network. Scene recognition results over few image of three benchmark datasets are shown in Fig. 11.

The structure of our DBN is comprised of various layers. In addition to visible and one output layer, there are also three hidden layers. Here Genetic Algorithm (GA) is used to compute the optimum units used in each hidden layer as well as total epochs for hidden layers. The output layer is responsible to generate the probabilities of classes under consideration. The visible layer takes categories of objects, IoU scores, and DSI as input. To maximize the correlation, a stochastic operation is executed on the input vector. Two different types of sampling methods are used in DBN: Contrastive Divergence (CD) and Persistent Contrastive Divergence (PCD). The scene recognition is performed by using the Softmax function. The training process continues until one of the three conditions is met: i) the highest possible epoch is achieved which is 200, ii) minimum gradient is approached, or iii) meet the mean squared error benchmark for performance. As we are using GA for the optimization of hidden units, therefore fitness function is most important in this process.

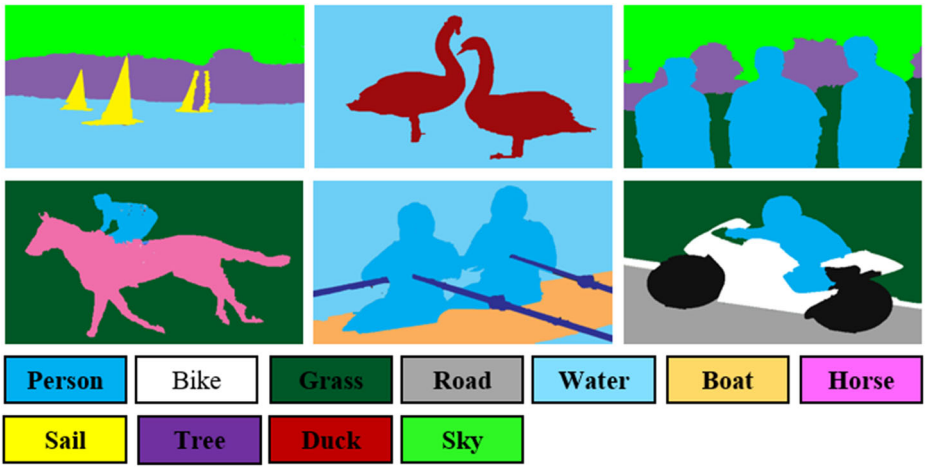

Fig. 11 Scene recognition results over various images from the benchmark datasets

Now, fitness function may be defined to decrease the error rate for scene recognition and reduce the training time of DBN, as follows:

$$Fit_{Fun} = 1000 * E + (T_T + T_B)/40 \tag{30}$$

where E denotes the misclassification rate, T_T represents the total training time of DNB before backpropagation and T_B is the time taken by the DBN for the fine-tuning of the parameters during backpropagation. The performance may be judged as smaller the error and time, smaller the fitness.

A parallel GA is incorporated to provide improved quality scene recognition as it divides the population into three subgroups called subpopulations. These three subpopulations used three different rates of mutations as demonstrated in the parameters table. The convergence of the fitness function is considered best when the training time and error rates are minimum. Once the generations reached 25, evolution is stopped as the chromosomes of individuals appeared almost identical. The optimized neurons in the three hidden layers are 984, 560, 1624 respectively while the epochs generated by the genetic algorithm are reported as 128, 112, and 148. These hidden layer neurons and epochs generated are analyzed and the best ones are used for scene recognition. The DBN classifier generates the probability values for every input feature vector. The classifier is trained against each dataset with its corresponding number of classes. For instance, the UIUC Sports dataset comprised 8 scene classes, the output generates eight probability values. The nearest probability value is assigned with the scene label. The parameters used are verified during our experiments. We have also tested different mutation rates, however, the best suitable rates are used and described. The parameters used in this research are described as follows: sub-populations = 2, individuals = 15 and 20, crossover probability = 0.85, rate of mutation = 0.01 and 0.03, migration rate = 0.1, termination condition = 50 generations.

Fig. 12 Example images of from the PASCAL VOC 2012 dataset

5 Experimental analysis

To evaluate the performance of designed system, we adapted the leave-one-single-out (LOSO) cross validation method. The three benchmark datasets, PASCAL VOC 2012, Caltech 101 and UIUC Sports are considered for training/testing performance evaluation. These datasets comprised of dynamic and challenging activities captured in divergent environments like public places, sports areas, and indoor-outdoor scenarios.

5.1 Dataset description

In this section, the comprehensive overview of the three-benchmark datasets that we used in this research is given.

5.1.1 The PASCAL VOC 2012 dataset

The PASCAL VOC dataset [17] is introduced for the recognition of objects from a number of real scenes. There are 20 object classes: airplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, sofa, train, TV/monitor which belongs to four categories namely, Person, Animal, Vehicle and Indoor. The training/validation data consists of 11,530 images, 27,450 regions of interest annotated objects and 6,929 segmentations. Figure 12 represents some examples from the PASCAL VOC 2012 dataset.

5.1.2 The Caltech 101 dataset

In the Caltech 101 dataset [18], there are 9144 images (RGB and gray-scale) having objects belonging to 101 categories where each category contains different classes. Some categories have less than 50 samples while some of them consists of more than 800 images. Fei-Fei Li, Marco Andreetto, and Marc 'Aurelio Ranzato collected these images in 2003. They performed some preprocessing on these images and flipped them manually to orient them in the same direction. They scaled the images to 300 x 200 pixels. We selected twenty classes from this dataset to perform our evaluation experiments: Horse, Kangaroo, Anchor, Motorbike, Ketch, Cannon, Camel, Duck, Lamp, Ferry, Hawksbill, Airplane, Elephant, Zebra, Helicopter, Tiger, Tree, Deer, Bear, and Dolphin. Figure 13 illustrates a few images from the Caltech 101 categories dataset.

Fig. 13 A few images from the Caltech 101 categories benchmark dataset

5.1.3 The UIUC sports dataset

The UIUC sports dataset introduced by Li and Fei-Fei [36] ,consists of eight sports event categories. They have captured the images in dynamic and challenging environments in high resolutions (minimum of 800x600 pixels). The UIUC dataset consists of the following categories: badminton, polo, rowing, croquet, bocce, snowboarding, sailing and rock climbing. Every category has a different number of images ranges form 100-240 images. All the images in each category were comprised of multiple foreground objects. Figure 14 shows a few examples from the dataset.

5.2 Experimental results and evaluation

In this section, we demonstrate the statistics of experimental results and evaluation. For performance evaluation, computational time of segmentation techniques is computed and compared. We used a Matlab environment to conduct the experiments using Intel 10th Generation Corei7 CPU of 2.0 GHz with 8GB of RAM and a dedicated 4GB graphics card. To evaluate the performance of the proposed ODSR over three benchmark datasets namely, PASCAL VOC 2012 , Caltech 101 and UIUC Sports, the experiments were repeated ten times and a LOSO cross validation method was used. The datasets are split into 1 and N-1 sample sets for testing and training respectively. Then, prediction weights are observed

Fig. 14 Some example images from the UIUC Sprots dataset

Table 7 Confusion matrix for recognition accuracy over the UIUC Sports dataset

	BAD	BOC	CRO	POL	RCL	ROW	SAL	SNB	Pr	Re	FS
BAD	**0.97**	0	0	0	0	0	0.03	0	0.89	0.82	0.85
BOC	0	**0.96**	0	0	0.04	0	0	0	0.87	0.86	0.86
CRO	0.05	0	**0.95**	0	0	0	0	0	0.88	0.78	0.83
POL	0.03	0	0	**0.93**	0	0.04	0	0	0.81	0.89	0.85
RCL	0	0.08	0	0	**0.90**	0	0	0	0.79	0.84	0.81
ROW	0	0	0	0	0	**0.87**	0	0	0.82	0.83	0.82
SAL	0.03	0	0	0	0	0	**0.83**	0.14	0.85	0.87	0.86
SNB	0	0	0	0	0	0.08	0	**0.92**	0.88	0.81	0.84
				Mean accuracy=**91.63%**					**0.848**	**0.837**	**0.842**

BAD=badminton, POL=polo, ROW=rowing, CRO=croquet, BOC=bocce, SNB=snowboarding, SAL=sailing RCL=rock-climbing, Pr=Precision, Re=Recall, FS= F1 Score

for each sample set. The experiments, accuracies, precision, recall and comparisons of the proposed ODSR system with other SOTA methods are described below:

5.2.1 Experiment 1: Using the UIUC sports dataset

In the experimental evaluation on the UIUC Sports dataset, our proposed system showed significant scene recognition results in terms of accuracy, precision, recall and F1 score. Table 7 illustrates the confusion matrix of recognition accuracies over eight different activities from the UIUC sports dataset with a mean accuracy of **91.63%**, along with precision, recall and F1 score for these sports activities. Table 10 demonstrates remarkable performance compared with other SOTA methods over the UIUC Sports dataset.

5.2.2 Experiment 2: Using the PASCAL VOC 2012 dataset

With regard to the PASCAL VOC 2012 dataset, the proposed model revealed better scene recognition results. Table 8 illustrates the confusion matrix of recognition accuracies, along with precision, recall and F1 score for twenty classes over the PASCAL VOC 2012 with an average accuracy of **87.67%**. The accuracy of PASCAL VOC 2012 is less than the accuracy for the UIUC Sports and Caltech 101 datasets due to multifarious background and the diverse range of cluttered scenes. However, Table 10 still reveals higher recognition accuracies over other SOTA techniques using the PASCAL VOC 2012.

5.2.3 Experiment 3: Using the Caltech 101 dataset

During the experiments, our proposed system obtained higher accuracy on 20 classes of the Caltech 101 dataset. Table 9 depicts the confusion matrix of recognition accuracies along with precision, recall and F1 score over the Caltech101 dataset with an average accuracy of **88.60%**. Similarly, Table 10 verifies excellent performance over other advanced scene recognition techniques.

Scene understanding is a hot topic for the last couple of decades and various researchers have worked on object detection and scene understanding by incorporating a diverse range of methods. A. Ucar et al. [54] used CNN with various layers for object recognition and

Table 8 Confusion matrix for recognition accuracy over the PASCAL VOC 2012 dataset

	HS	BD	PR	CW	SH	AP	CT	DG	BT	BS	TN	CR	MB	BC	BL	CH	DT	PP	SF	TV	Pr	Re	F1
HS	0.93	0	0	0.05	0	0	0	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0.89	0.86	0.875
BD	0	0.89	0	0	0	0	0.06	0	0	0	0	0	0	0	0	0	0	0.05	0	0	0.81	0.87	0.839
PR	0	0	0.91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.09	0	0	0.80	0.81	0.805
CW	0.04	0	0	0.93	0.02	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0.91	0.83	0.868
SH	0.01	0	0	0	0.90	0	0.04	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0.83	0.84	0.835
AP	0	0	0	0	0	0.87	0	0	0.07	0.06	0	0	0	0	0	0	0	0	0	0	0.87	0.82	0.844
CT	0.03	0	0	0	0	0	0.86	0.11	0	0	0	0	0	0	0	0	0	0	0	0	0.80	0.80	0.800
DG	0	0	0	0	0	0	0	0.91	0	0	0	0	0	0.03	0	0	0	0	0	0	0.81	0.85	0.829
BT	0	0	0	0	0	0	0	0	0.79	0.09	0.08	0.04	0	0	0	0	0	0	0	0	0.79	0.76	0.744
BS	0	0	0	0	0	0	0	0	0.04	0.85	0	0.11	0	0	0	0	0	0	0	0	0.77	0.77	0.770
TN	0	0	0	0	0	0	0	0	0.07	0.12	0.81	0	0	0	0	0	0	0	0	0	0.82	0.74	0.778
CR	0	0	0	0	0	0	0	0	0.08	0.02	0	0.90	0	0	0	0	0	0	0	0	0.85	0.78	0.813
MB	0	0	0	0	0	0	0	0	0	0	0	0	0.87	0.13	0	0	0	0	0	0	0.84	0.84	0.840
BC	0	0	0	0	0	0	0	0.03	0	0	0	0	0.17	0.83	0	0	0	0	0	0	0.78	0.83	0.804
BL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.88	0	0	0.12	0	0	0.83	0.82	0.825
CH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.91	0	0	0.09	0	0.81	0.80	0.805
DT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.87	0.04	0	0.09	0.85	0.81	0.829
PP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03	0	0	0.93	0	0.04	0.87	0.85	0.
SF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.13	0	0	0.81	0.06	0.79	0.86	0.823
TV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.06	0.07	0	0	0.87	0.77	0.76	0.764
												Mean accuracy=**87.67%**									**0.824**	**0.815**	**0.819**

HS = horse; BD = bird; PR = person; CW = cow; SH = sheep; AP = aeroplane; CT = cat; DG = dog; BT = boat; BS = bus; TN = train; CR = car; MB = motorbike; BC = bicycle; BL = bottle; CH = chair; DT = dining table; PP = potted plant; SF = sofa; TV = television; Pr = Precision; Re = Recall; F1 = F1 Score

Table 9 Confusion matrix for recognition accuracy over the Caltech 101 dataset

	HS	KR	AN	MB	KT	CN	CM	DK	FY	HW	AP	ET	ZB	LP	HC	TG	TR	DR	BR	DN	Pr	Re	F1
HS	0.87	0	0	0	0	0.09	0	0	0.02	0	0.01	0.01	0	0	0	0	0	0	0	0	0.91	0.89	0.899
KR	0.03	0.82	0	0	0	0.03	0	0	0.04	0	0.04	0.02	0	0	0	0	0	0.02	0	0	0.87	0.85	0.860
AN	0	0	0.93	0	0	0	0.07	0	0	0	0	0	0	0	0	0	0	0	0	0	0.88	0.84	0.859
MB	0	0	0	0.94	0.01	0	0	0	0	0	0	0	0.01	0	0	0.04	0	0	0	0	0.92	0.87	0.894
KT	0	0	0	0	0.85	0	0.07	0	0	0.03	0	0	0.05	0	0	0	0	0	0	0	0.86	0.81	0.834
CN	0.05	0	0	0	0	0.93	0	0	0.01	0	0.01	0	0	0	0	0	0	0	0	0	0.83	0.88	0.854
CM	0	0	0.03	0	0.05	0	0.90	0	0	0	0	0	0.02	0	0	0	0	0	0	0	0.84	0.86	0.849
DK	0	0	0	0.02	0	0	0	0.91	0	0	0	0	0	0.06	0	0.01	0	0	0	0	0.87	0.82	0.844
FY	0.01	0	0.02	0	0	0.03	0	0	0.89	0	0.04	0	0	0	0	0	0	0.01	0	0	0.81	0.79	0.799
HW	0	0	0.08	0	0.02	0	0	0	0	0.90	0	0	0	0	0	0	0	0	0	0	0.82	0.81	0.814
AP	0	0.01	0	0	0	0.01	0	0	0.06	0	0.88	0.04	0	0	0	0	0	0	0	0	0.80	0.77	0.784
ET	0	0.02	0	0	0	0	0	0	0.06	0	0.02	0.87	0	0	0	0	0	0.03	0	0	0.81	0.72	0.762
ZB	0	0	0	0.04	0.05	0	0.03	0	0	0.03	0	0	0.83	0	0	0	0	0	0	0.02	0.83	0.80	0.815
LP	0	0	0	0	0	0	0	0.07	0	0.02	0	0	0	0.86	0	0.02	0.03	0	0	0	0.77	0.81	0.789
HC	0.02	0	0	0	0	0.03	0	0	0.03	0	0	0.02	0	0	0.90	0	0	0	0	0	0.79	0.75	0.769
TG	0	0	0	0.06	0	0	0	0	0	0	0	0	0	0.02	0	0.92	0	0	0	0	0.80	0.83	0.815
TR	0	0	0	0.09	0	0	0	0	0	0	0	0	0.03	0	0	0	0.88	0	0	0	0.82	0.78	0.799
DR	0	0	0	0.05	0	0	0	0.02	0	0	0	0	0	0.02	0	0	0	0.91	0	0	0.88	0.83	0.854
BR	0	0	0	0	0	0	0.05	0	0	0	0	0	0	0	0	0.04	0	0	0.88	0.03	0.75	0.79	0.769
DN	0	0	0	0.09	0	0	0	0	0	0	0	0	0.02	0	0	0	0.03	0.01	0	0.85	0.78	0.80	0.789
									Mean accuracy=**88.60%**												**0.832**	**0.815**	**0.822**

HS=Horse, KR=Kangaroo, AN=Anchor, MB=Motorbike, KT=Ketch, NM=Camel, DK=Duck, LP=Lamp, FY=Ferry, HW=Hawksbill, AP=Airplane, ET=Elephant, ZB=Zebra, HC=Helicopter, TG=Tiger, TR= Tree, DR=Deer, BR=Bear, DN=Dolphin, Pr = Precision; Re = Recall; F1 = F1 Score

Table 10 Comparison of the proposed method's recognition accuracy with SOTA methods over PASCAL VOC 2012, Caltech 101 and UIUC Sports datasets

Methods	PASCAL VOC 2012 (%)	Caltech 101 (%)	UIUC Sports (%)
A. Ucar et al. [54]	–	82.80	–
H. Zhao et al.[64]	85.40	–	–
Z. Niu et al. [45]	–	–	78.00
S. Shetty et al. [51]	85.60	–	–
A. A. Rafiqeu et al. [47]	–	–	85.09
C. Zheng et al. [65]	–	–	80.30
S. Gupta et al.[23]	–	85.43	–
N Hussain et al. LSVM [26]	–	76.90	–
J. Guo et al. [60] [22]	70.7	–	–
L. Chen et al [61] [12]	72.9	–	–
J. Feng et al. [62] [19]	–	–	87.34
M. Bansal et al. [63] [9]	–	86.4	–
Proposed ODSR	**87.67**	**88.60**	**91.63**

pedestrian tracking. They consider features extraction techniques including CNN, BOW by taking into account HOG and SURF techniques followed by an SVM classifier. In [64], H. Zhao et al. proposed a novel framework that integrates pyramid pooling and PSPNet by incorporating contextual as well as spatial information for scene recognition. J. Feng et al. [19] introduced a hybrid approach by combining supervised and unsupervised techniques. They analyzed visual features by employing CNN while semantics are studied via topic model. P. Y. Chen et al. [11] proposed a SceneNet that considers the geometric structure of the object with the combination of semantic segmentation and produced competitive results for scene understating. M. Jaritz et al. [30] presented a Multi-View PointNet model to understand a 3D scene and predict the labels accurately. I. Armeni et al. [5] introduced a new dataset for scene understanding comprised 70,000 images having 13 object categories. The initially collected data from Matterport Camera is further processed to generate RGB-D data as well as to extract point clouds of the same images. The data is annotated based on these point clouds instead of annotating images. The key feature of the dataset is the coexistence of its modalities like RGB images, scene labels, depth information, surface normal, and annotated 3D meshes.

Some other well-known methods are very helpful in human detection and object recognition. N. D. Doulamis et al. [14] presented a re-adjustment framework that is effective for behavior classification and recognition. They incorporated users into the learning process to improve the modeling of behavior that leads towards enhanced classification. In [10], P.V.K. Borges et al., presented an extensive survey on different techniques including feature extraction, segmentation, and detection of objects for behavior understanding. Furthermore, the author highlighted efficient approaches and important future directions for used benchmark datasets. Chung et al. [13], discussed Hierarchical Context Hidden Markov Model (HC-HMM), which used videos from a nursing center to understand elderly behaviors. The key idea was to extract three contexts including spatial, activities, and temporal from the videos. Moreover, HC-HMM achieved high accuracy of 90% and 0% false alarm. Another Bayesian filter-based model [35], achieved remarkable recognition rates of behavior recognition and understanding over real-life complex scenarios. In this work, Hidden Markov

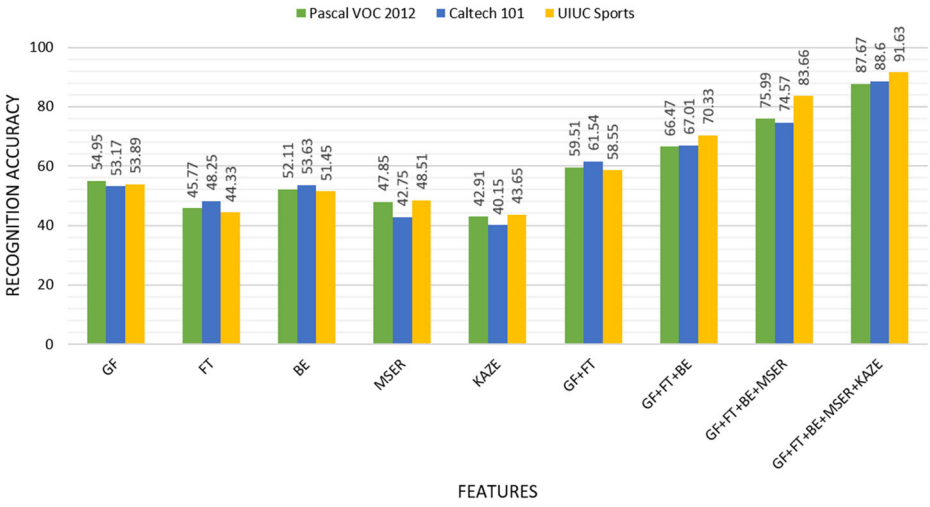

Fig. 15 Scene recognition results over various images from the benchmark datasets

models are used to support the filters for understanding the behaviors at an industrial plant. Bakalos et al. [8], used adaptive deep learning and multimodal fused data to monitor critical systems for the protection of water infrastructure from different types of attacks. It used visual surveillance, ICS sensor data, and channel state information (CSI) to detect different attacks including cyberattacks and human presence assaults.

5.3 Discussion

The proposed ODSR Model is intended to achieve better performance over other SOTA methods by incorporating our novel MEsSP segmentation technique. Initially, various images having a diverse range of objects and complex backgrounds are taken as input from benchmark datasets including Pascal VOC 2012, Caltech 101, and UIUC sports. The input images are processed for segmentation using our proposed MEsSP segmentation technique as well as FCM an existing segmentation technique. Both of these segmentation techniques performed well, however, our proposed segmentation method MEsSP demonstrated better results in terms of segmentation accuracy and computational time as shown in Tables 3, 4, and 5.

The overall multi-object detection and scene recognition is improved due to effective segmentation mechanism and combination of features including geometrical features (GF), Fourier Transform (FT), blob extraction (BE), maximally stable extremal regions (MSER), and KAZE features. We have conducted experiments by choosing features one by one, and then the combination of these features to note the changes in the scene recognition accuracy.

Table 11 The computational time (seconds) of the proposed ODSR model vs other well-known classifiers

Dataset	Proposed	Linear SVM	Decision Tree	ANN
Pascal VOC 2012	**6802.70**	8357.44	7911.75	8687.33
Caltech 101	**5120.64**	5791.50	5623.25	6158.19
UIUC Sports	**1072.10**	1390.09	1567.67	1871.95

Table 12 The recognition accuracies (%) of well-known classifiers over benchmark datasets

Dataset	Proposed	Linear SVM	Decision Tree	ANN
Pascal VOC 2012	**87.67**	74.95	77.54	82.11
Caltech 101	**88.60**	81.25	83.61	85.37
UIUC Sports	**91.63**	84.77	85.01	87.89

It is demonstrated that while using only a single feature, whatever the feature is, the scene recognition results are very low and below 55%. On the other hand, while we used an approach to add the features in incremental order, the performance of the model increased with the addition of each new feature. The detailed analysis of the individual as well as combinations of features is illustrated in Fig. 15. It depicts the effectiveness of the features selected for the ODSR model. Some of the features performed better on Pascal VOC 2012 dataset while some others performed on Caltech 101 and UIUC Sports datasets, however, the combination of these features has proven superior to only considering the individual features on all the three benchmark datasets.

It is clear from the previous paragraphs that based on the computational time and accuracy of both the segmentation techniques, MEsSP is more effective and accurate. Therefore, the results from MEsSP are considered for further processing i.e. feature extraction, object categorization, and scene recognition.

Moreover, we have conducted experiments by applying well-known classifiers and compared accuracy as well as the computational time with our proposed model under the same conditions and computational environment. The details of the computational time taken by each classifier over benchmark datasets are demonstrated in Table 11 while recognition accuracies comparison of well-known classifiers with proposed one is illustrated in Table 12 which shows the significance of the proposed model over other models in terms of computational time.

6 Conclusion and future work

In this paper, we formulated an entropy-scaled super-pixels segmentation method to recognize multiple objects in an image. Our proposed model can precisely recognize multiple objects in dynamic and challenging environments for the three-benchmark datasets which are the Pascal VOC 2012, Caltech 101 and UIUC sports datasets. Firstly, after preprocessing the input images, segmentation of these images is performed and a bag of features is extracted. These extracted features are comprised of blob extraction, geometrical features, MSER, KAZE and FFT. Then, MKL categorizes the objects into different classes. Finally, after computing IoU scores and DSC, a DBN is employed to predict the scene labels. Our proposed method significantly outperforms others in terms of accuracy, precision and recall.

In addition to the model's incredible performance, we also encountered several limitations while working with it.

6.1 Theoretical implications

The proposed ODSR model works in different and complex scenarios to classify various scenes. ODSR works with single as well as multi-object datasets, although the theoretical implications to find out the more complex application of the system in terms of scene

recognition in autonomous driving, sports, medical diagnosis, drone targeting, surveillance system, however, for these applications, we can apply the proposed ODSR system in a real-time environment.

6.2 Research limitations

The UIUC Sports dataset in the domain of scene understanding and recognition is a more complex dataset compared to the Caltech 101 dataset while Pascal VOC 2012 dataset is more challenging as it comprised 20 classes with diverse and cluttered backgrounds. Our findings are somewhat varied across the datasets due to the varying complexity and messy information of the datasets. We faced difficulties under conditions like occluded and similar objects. In the future, we will work on this problem by using deep learning feature extraction techniques fused with other classical features to devise a new method for outstanding scene recognition along with semantics.

In the future, we will adopt new feature extraction strategies using deep learning models along with classical features extraction techniques for the recognition of objects and scenes to overcome the challenges encountered during this research.

Acknowledgements This research was supported by the Ministry of Culture, Sports and Tourism and Korea Creative Content Agency (Project Number: R2021040093).

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Ahmed A, Jalal A, Kim K (2020) A novel statistical method for scene classification based on multi-object categorization and logistic regression. Sensors 20(14):3871
2. Alcantarilla PF, Bartoli A, Davison AJ (2012) KAZE features. In: European conference on computer vision (pp 214–227). Springer, Berlin, Heidelberg
3. Appiah O, Asante M, Hayfron-Acquah JB (2020) Improved approximated median filter algorithm for real-time computer vision applications. Journal of King Saud University-Computer and Information Sciences
4. Arasu B, Kumaran S (2014) Blind man's artificial EYE an innovative idea to help the blind. In: Conference proceeding of the international journal of engineering development and research (IJEDR), SRM university, Kattankulathur, pp 205–207
5. Armeni I, Sax S, Zamir AR, Savarese S (2017) Joint 2d-3d-semantic data for indoor scene understanding. arXiv:1702.01105
6. Arnold E, Al-Jarrah OY, Dianati M, Fallah S, Oxtoby D, Mouzakitis A (2019) A survey on 3d object detection methods for autonomous driving applications. IEEE Trans Intell Transp Syst 20(10):3782–3795
7. Asif U, Bennamoun M, Sohel FA (2017) RGB-D object recognition and grasp detection using hierarchical cascaded forests. IEEE Trans Robot 33(3):547–564
8. Bakalos N, Voulodimos A, Doulamis N, Doulamis A, Ostfeld A, Salomons E, Li P (2019) Protecting water infrastructure from cyber and physical threats: Using multimodal data fusion and adaptive deep learning to monitor critical systems. IEEE Signal Proc Mag 36(2):36–48
9. Bansal M, Kumar M, Kumar M, Kumar K (2021) An efficient technique for object recognition using the Shi-Tomasi corner detection algorithm. Soft Comput 25(6):4423–4432
10. Borges PVK, Conci N, Cavallaro A (2013) Video-based human behavior understanding: a survey. IEEE Trans Circuits Syst Vid Technol 23(11):1993–2008

11. Chen PY, Liu AH, Liu YC, Wang YCF (2019) Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp 2624–2632
12. Chen L, Zhan W, Tian W, He Y, Zou Q (2019) Deep integration: a multi-label architecture for road scene recognition. IEEE Trans Image Process 28(10):4883–4898
13. Chung PC, Liu CD (2008) A daily behavior enabled hidden Markov model for human behavior understanding. Pattern Recogn 41(5):1572–1580
14. Doulamis ND, Voulodimos AS, Kosmopoulos DI, Varvarigou TA (2010, October) Enhanced human behavior recognition using hmm and evaluative rectification. In: Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams, pp 39–44
15. Debelee TG, Schwenker F, Rahimeto S, Yohannes D (2019) Evaluation of modified adaptive k-means segmentation algorithm. Comput Vis Media 5(4):347–361
16. Dong X, Lei Y, Wang T, Thomas M, Tang L, Curran WJ, Yang X (2019) Automatic multiorgan segmentation in thorax CT images using U-net-GAN. Med Phys 46(5):2157–2168
17. Everingham M, Eslami SA, Van Gool L, Williams CK, Winn J, Zisserman A (2015) The pascal visual object classes challenge: A retrospective. Inter J Comput Vis 111(1):98–136
18. Fei-Fei L, Fergus R, Perona P (2004) Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: 2004 conference on computer vision and pattern recognition workshop (pp 178–178). IEEE
19. Feng J, Fu A (2018) Scene semantic recognition based on probability topic model. Information 9(4):97
20. Feng X, Jiang Y, Yang X, Du M, Li X (2019) Computer vision algorithms and hardware implementations: A survey. Integration 69:309–320
21. Gadekallu TR, Rajput DS, Reddy M, Lakshmanna K, Bhattacharya S, Singh S, Alazab M (2021) A novel PCA–whale optimization-based deep neural network model for classification of tomato plant diseases using GPU. J Real-Time Image Proc 18(4):1383–1396
22. Guo J, Gould S (2015) Deep CNN ensemble with data augmentation for object detection. arXiv:1506.07224
23. Gupta S, Kumar M, Garg A (2019) Improved object recognition results using SIFT and ORB feature detector. Multimed Tools Appl 78(23):34157–34171
24. Hakak S, Alazab M, Khan S, Gadekallu TR, Maddikunta PKR, Khan WZ (2021) An ensemble machine learning approach through effective feature extraction to classify fake news. Futur Gener Comput Syst 117:47–58
25. Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554
26. Hussain N, Khan MA, Sharif M, Khan SA, Albesher AA, Saba T, Armaghan A (2020) A deep neural network and classical features based scheme for objects recognition: an application for machine inspection. Multimed Tools Appl. https://doi.org/10.1007/s11042-020-08852-3
27. Jalal A, Batool M, Kim K (2020) Stochastic recognition of physical activity and healthcare using tri-axial inertial wearable sensors. Appl Sci 10(20):7122
28. Jalal A, Kim YH, Kim YJ, Kamal S, Kim D (2017) Robust human activity recognition from depth video using spatiotemporal multi-fused features. Pattern Recognit 61:295–308
29. Jalal A, Sarif N, Kim JT, Kim TS (2013) Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart home. Indoor Built Environ 22(1):271–279
30. Jaritz M, Gu J, Su H (2019) Multi-view pointnet for 3d scene understanding
31. Jiang X, Guo Y, Chen H, Zhang Y, Lu Y (2019) An adaptive region growing based on neutrosophic set in ultrasound domain for image segmentation. IEEE Access 7:60584–60593
32. Jiang B, Luo R, Mao J, Xiao T, Jiang Y (2018) Acquisition of localization confidence for accurate object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 784–799
33. Kachouri R, Soua M, Akil M (2016) Unsupervised image segmentation based on local pixel clustering and Low-Level region merging. In: 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP) (pp 177–182). IEEE
34. Kamada S, Ichimura T (2019) An object detection by using adaptive structural learning of deep belief network. In: 2019 international joint conference on neural networks (IJCNN) (pp 1–8). IEEE
35. Kosmopoulos DI, Doulamis ND, Voulodimos AS (2012) Bayesian filter based behavior recognition in workflows allowing for user feedback. Comput Vis Image Underst 116(3):422–434
36. Li LJ, Fei-Fei L (2007) What, where and who? classifying events by scene and object recognition. In: 2007 IEEE 11th international conference on computer vision (pp 1–8). IEEE
37. Liu MY, Tuzel O, Ramalingam S, Chellappa R (2011) Entropy rate superpixel segmentation. In: CVPR 2011 (pp 2097–2104). IEEE

38. Liu Y, Zhou S, Chen Q (2011) Discriminative deep belief networks for visual data classification. Pattern Recogn 44(10-11):2287–2296
39. Mahmood M, Jalal A, Kim K (2019) WHITE STAG Model: Wise human interaction tracking and estimation (WHITE) using spatio-temporal and angular-geometric (STAG) descriptors. Multimedia Tools and Applications, pp 1–32
40. Matas J, Chum O, Urban M, Pajdla T (2004) Robust wide-baseline stereo from maximally stable extremal regions. Image Vis Comput 22(10):761–767
41. Miao J, Zhou X, Huang TZ (2020) Local segmentation of images using an improved fuzzy C-means clustering algorithm based on self-adaptive dictionary learning. Applied Soft Computing, p 106200
42. Nair V, Chatterjee M, Tavakoli N, Namin AS, Snoeyink C (2020) Fast fourier transformation for optimizing convolutional neural networks in object recognition. arXiv:2010.04257
43. Nanni L, Lumini A (2013) Heterogeneous bag-of-features for object/scene recognition. Appl Soft Comput 13(4):2171–2178
44. Narain S, Ranganathan A, Noubir G (2019) Security of GPS/INS based on-road location tracking systems. In: 2019 IEEE Symposium on Security and Privacy (SP) (pp 587–601). IEEE
45. Niu Z, Hua G, Gao X, Tian Q (2012) Context aware topic model for scene recognition. In: 2012 IEEE Conference on computer vision and pattern recognition (pp 2743–2750). IEEE
46. Quaid MAK, Jalal A (2020) Wearable sensors based human behavioral pattern recognition using statistical features and reweighted genetic algorithm. Multimed Tools Appl 79(9):6061–6083
47. Rafique AA, Jalal A, Ahmed A (2019) Scene understanding and recognition: Statistical segmented model using geometrical features and gaussian naïve Bayes. In: IEEE conference on International Conference on Applied and Engineering Mathematics (vol 57)
48. Rashid M, Khan MA, Alhaisoni M, Wang SH, Naqvi SR, Rehman A, Saba T (2020) A sustainable deep learning framework for object recognition using multi-layers deep features fusion and selection. Sustainability 12(12):5037
49. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S (2019) Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 658–666
50. Rohan A, Rabah M, Kim SH (2019) Convolutional neural network-based real-time object detection and tracking for parrot AR drone 2. IEEE Access 7:69575–69584
51. Shetty S (2016) Application of convolutional neural network for image classification on Pascal VOC challenge 2012 dataset. arXiv:1607.03785
52. Song X, Jiang S, Herranz L (2015) Joint multi-feature spatial context for scene recognition on the semantic manifold. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1312–1320
53. Szuster P (2019) Blob extraction algorithm in detection of convective cells for data fusion. J Telecommun Inf Technol (4), pp 65–73
54. Uçar A, Demir Y, Güzeliş C (2016) Moving towards in object recognition with deep learning for autonomous driving applications. In: 2016 International Symposium on INnovations in Intelligent SysTems and Applications (INISTA) (pp 1–5). IEEE
55. Ulhaq A, Born J, Khan A, Gomes DPS, Chakraborty S, Paul M (2020) Covid-19 control by computer vision approaches: a survey. IEEE Access 8:37-179456
56. Vasan D, Alazab M, Wassan S, Naeem H, Safaei B, Zheng Q (2020) IMCFN: Image-Based malware classification using fine-tuned convolutional neural network architecture. Comput Netw 171:107138
57. Veta M, Van Diest PJ, Jiwa M, Al-Janabi S, Pluim JP (2016) Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method. PloS one 11(8):e0161286
58. Xia S, Zeng J, Leng L, Fu X (2019) WS-AM: Weakly Supervised attention map for scene recognition. Electronics 8(10):1072
59. Xu Y, Wu T, Gao F, Charlton JR, Bennett KM (2020) Improved small blob detection in 3D images using jointly constrained deep learning and Hessian analysis. Sci Rep 10(1):1–12
60. Zamani F, Jamzad M (2017) A feature fusion based localized multiple kernel learning system for real world image classification. EURASIP J Image Vid Process 2017(1):78
61. Zhang Y, Jin R, Zhou ZH (2010) Understanding bag-of-words model: a statistical framework. Int J Mach Learn Cybern 1(1-4):43–52
62. Zhang L, Zhen X, Shao L (2014) Learning object-to-class kernels for scene classification. IEEE Trans Image Process 23(8):3241–3253
63. Zhao W, Fu Y, Wei X, Wang H (2018) An improved image semantic segmentation method based on superpixels and conditional random fields. Appl Sci 8(5):837
64. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2881–2890

65. Zheng C, Yi Y, Qi M, Liu F, Bi C, Wang J, Kong J (2017) Multicriteria-based active discriminative dictionary learning for scene recognition. IEEE Access 6:4416–4426
66. Zhu H, Zhuang Z, Zhou J, Wang X, Xu W (2018) Improved graph-cut segmentation for ultrasound liver cyst image. Multimed Tools Appl 77(21):28905–28923
67. Zia S, Yuksel B, Yuret D, Yemez Y (2017) RGB-D object recognition using deep convolutional neural networks. In: Proceedings of the IEEE International conference on computer vision workshops, pp 896–903

Affiliations

Adnan Ahmed Rafique[1,2] · Munkhjargal Gochoo[3] · Ahmad Jalal[1] · Kibum Kim[4]

Adnan Ahmed Rafique
adnanrafique@upr.edu.pk

Munkhjargal Gochoo
mgochoo@uaeu.ac.ae

Ahmad Jalal
ahmadjalal@mail.au.edu.pk

1 Air University, E-9, Islamabad, Pakistan

2 University of Poonch, Rawalakot, Rawalakot, Poonch, AJK, Pakistan

3 Department of Computer Science and Software Engineering, United Arab Emirates University, Al Ain 15551, UAE

4 Department of Human-Computer Interaction, Hanyang University, Seoul, South Korea