IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 22, 2021, accepted February 9, 2021, date of publication February 12, 2021, date of current version February 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3058986

Scene Semantic Recognition Based on Modified Fuzzy C-Mean and Maximum Entropy Using Object-to-Object Relations

AHMAD JALAL[1], ABRAR AHMED[1], ADNAN AHMED RAFIQUE[1], AND KIBUM KIM[2], (Member, IEEE)

[1]Department of Computer Science, Air University, Islamabad 44000, Pakistan
[2]Department of Human-Computer Interaction, Hanyang University, Ansan 15588, South Korea

Corresponding author: Kibum Kim (kibum@hanyang.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) under Grant 2018R1D1A1A02085645.

ABSTRACT With advances in machine vision systems (e.g., artificial eye, unmanned aerial vehicles, surveillance monitoring) scene semantic recognition (SSR) technology has attracted much attention due to its related applications such as autonomous driving, tourist navigation, intelligent traffic and remote aerial sensing. Although tremendous progress has been made in visual interpretation, several challenges remain (i.e., dynamic backgrounds, occlusion, lack of labeled data, changes in illumination, direction, and size). Therefore, we have proposed a novel SSR framework that intelligently segments the locations of objects, generates a novel Bag of Features, and recognizes scenes via Maximum Entropy. First, denoising and smoothing are applied on scene data. Second, modified Fuzzy C-Means integrates with super-pixels and Random Forest for the segmentation of objects. Third, these segmented objects are used to extract a novel Bag of Features that concatenate different blobs, multiple orientations, Fourier transform and geometrical points over the objects. An Artificial Neural Network recognizes the multiple objects using the different patterns of objects. Finally, labels are estimated via Maximum Entropy model. During experimental evaluation, our proposed system illustrated a remarkable mean accuracy rate of 90.07% over the MSRC dataset and 89.26% over the Caltech 101 for object recognition, and 93.53% over the Pascal-VOC12 dataset for scene recognition, respectively. The proposed system should be applicable to various emerging technologies, such as augmented reality, to represent the real-world environment for military training and engineering design, as well as for entertainment, artificial eyes for visually impaired people and traffic monitoring to avoid congestion or road accidents.

INDEX TERMS Scene recognition, object segmentation, recognition, bag of features, artificial neural network, maximum entropy, object pattern.

I. INTRODUCTION

Visual sensor [1] technology is one of the most significant human sensing tools that attains content awareness from the surroundings by exploiting the statistical dependencies of detected objects [2]. Using visual sensor technology, people can also identify multiple objects [3], find and describe the relationships between objects [4], [5] and interpret situations to perceive scene types. To make machines capable of sensing the world as humanoid abilities, researchers have paid major attention to scene semantic recognition (SSR) [6], [7], automatic analysis of object positions [8] and structural correlation between multiple objects in scenery images. However, massive challenges remain (i.e., variation in illumination, size of objects, view orientations, multi-object occlusion and object tracking) to improve the reliability of SSR in practical applications such as security navigation to automatically identify suspicious/violent scenes, recognition of social interaction type in public areas, distinguishing between various sports scenes [9], remote sensing [10] and aerial scene classification [11], [12]. The SSR system proposed in this paper is comprised of the steps illustrated in Fig. 1.SSR systems

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

usually consist of four basic modules: pre-processing, object segmentation, feature extraction and recognition. The pre-processing module is used to target specific areas of interest and also to remove unnecessary information such as noise contents [13]. The object segmentation [14] module deals with the transformation of an image into a set of pixel regions represented by a mask or labels in an image. The feature extraction [15] module is used to identify and label each object. Finally, in the recognition module, a semantic relationship amongst detected objects is observed and a classifier is applied to marked objects with their predefined labels for recognizing scene classes. Many well-known methods such as multiscale contrast, region clustering, histogram spanning, color/texture features, plane-based point clouds, graph-cut and kernel collaborative classification are considered to strengthen all the above modules during SSR.

In this paper, we propose a novel SSR model that integrates modified Fuzzy C-Means [16] and Random Forest to segment single/multiple objects [17]. Then, different features such as discrete Fourier transform, blob extraction, multiple orientation and geometrical shape are merged to develop a Bag of Features. After feature extraction, Artificial Neural Network [18] recognizes single/multiple objects based on extracted features. Finally, these objects estimate the posterior of the class label by employing the Maximum Entropy [19] method for scene classification. In this frameworks, we have resolved several problems that can directly affect the recognition accuracy of multiple objects and scenes, namely, variations in the size of the objects, variations in illumination, multiple view orientations of the object, occlusion caused by multiple objects in the scene, shadows, high intensity edges and changes in the direction of the objects in the images. We obtained remarkable improvement in recognition rates over other state-of-the-art (SOTA) methods in three datasets. The major contributions of this work can be summarized as follows.

- We propose a robust method for multiple objects and scene recognition based on Artificial Neural Network and Maximum Entropy model, which can successfully predict the semantic labels of objects in different scenes.
- Improved segmentation for the estimation of multiple regions of images and a novel Bag of Features is the main contribution of this work. Our precise segmentation improved the overall accuracy of scene recognition.
- A novel Bag of Features for multiple object recognition has improved recognition accuracy by Artificial Neural Network.
- The efficiency and the efficacy of the proposed work are validated in the experimental results over three publicly available datasets demonstrating that the proposed work outperforms other SOTA methods.

The rest of the paper is structured as follows: Section II discusses the related the work. Section III presents the flow architecture of the proposed SSR methodology which includes Fuzzy C-Means and Random Forest algorithms, feature extractions using multiple techniques, multiple objects recognition using ANN and scene recognition via Maximum Entropy. Section IV describes the analysis and comparison of the recognition rate of our work with other SOTA SSR systems using MSRC, Caltech 101 and Pascal-VOC12 datasets. Finally, Section V provides the conclusion with some future directions.

II. RELATED WORK

Visual scenes are examined by humans in a simple manner, however, machines face various challenges, specifically, object location, size variation, view orientation and the impact of these challenges on scenes to makes the extraction and manipulation of useful information for SSR problematic. Furthermore, the visual interpretation of data from these scenes remains a critical task for researchers. We divide the related work into object segmentation and scene recognition.

A. SINGEL/MULTIPLE OBJECT SEGMENTATION AND RECOGNITION

During object segmentation, an image is partitioned into a set of pixel regions represented by a mask or labels. Such segmentation of single/multiple objects with complex backgrounds has been the subject of extensive studies and still has massive gaps which need to be filled to remove the imperfections from the studies. Liu *et al.* [20] proposed a framework to recognize a salient object based on their novel set of features including multiscale contrast, a histogram spanning from the center to the surroundings and spatial distribution with respect to color. They extended their approach by imparting Random Geometric prior Forest for the best segmentation results from sequential images. Li *et al.* [21] proposed a salient segmentation method, by using a saliency mask and distinct object labels. This segmentation method includes three steps: estimation of saliency maps, detection of salient object contours and identification of salient object instances.

Xu *et al.* [22] formulated a technique by integrating Fuzzy C-Means (FCM) and Graph Cut to segment images and split the colors into super-pixels by implementing the Turbo-pixel algorithm. They extracted color histogram features from these super-pixels to make clusters and employed Graph Cut to extract the segmented objects. Khan *et al.* [23] proposed a model in which they fused the pairwise relative spatial information of images with distance and angles between visual information in the images while most approaches considered the distance only. Using the extra information, they have reduced the computational cost and achieve good classification accuracy of 83.50% over the MSRC dataset but they lack in the classification accuracy over the Caltech 101 and Pascal-VOC07 datasets which produced only 68.40% and 54.97% accuracy respectively. Li *et al.* [24] introduced a method for object recognition based on the optimal Bag of Words model and Region of Interest (ROI). They extracted ROI by the combination of a saliency map and Shi-Thomasi corner. They considered SIFT feature descriptor, visual codebook through

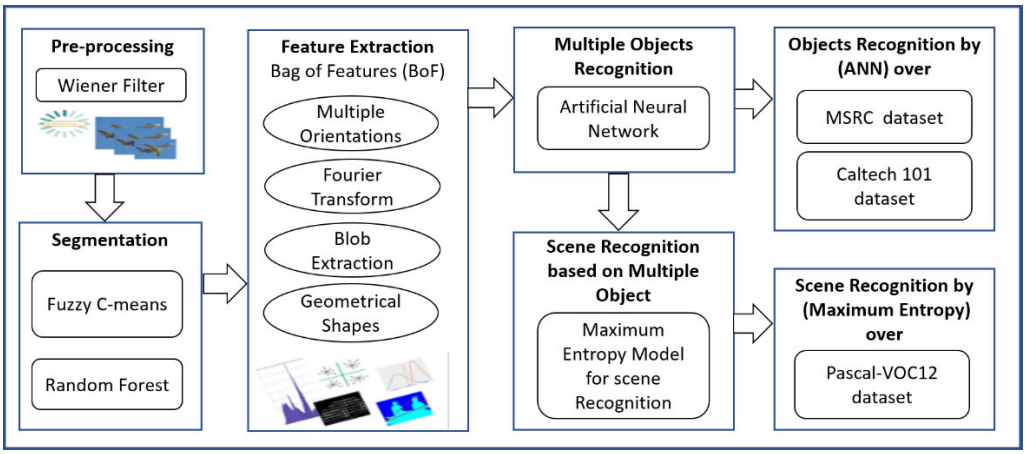


FIGURE 1. Flow architecture of proposed SSR model using Bag of Features and Maximum Entropy model.

Gaussian Mixture model and Support Vector machine to perform object recognition and classification.

B. SCENE RECEOGNITION

Scene recognition is currently the most widely studied topic that explores new directions such as the structure recognition of urban areas based on contents in the scenes, satellite or aerial remote sensing recognition, similar object properties among different scenes and orientation recognition of different objects (i.e., robots and smart machines). In [25], P. Espinace *et al.* proposed a generative probabilistic hierarchal model that uses low-level features for object detection and then these objects are contextually analyzed for scene recognition. They used a probability-based object classifier that worked on pre-searched objects to devise the probability distribution of scenes. In [26], J. Shotton *et al.* introduced a discriminative model which incorporates texture, layout, context information and well using conditional random fields. Their model provided automatic scene recognition based on visual information and semantic segmentation.

Chen *et al.* [27] developed a system, in which a Prototype-Agnostic scene layout construction approach is used to represent the spatial structure of scene images. Their system has flexibility to capture the diverse characteristics of the scene images and has generalization capability. In [28], R. Kachouri *et al.* used unsupervised image segmentation to efficiently segment the image into meaningful regions. They applied the merging mechanism on the color and texture features to explore various objects to sketch scene information. Li and Fei [29] proposed the Generative Graphical model, variation message passing and scene level likelihoods to achieve integrative and holistic scene recognition. Their approach needs proper semantic labels for the scenes, and therefore complex variations among object definition causes effects during recognition. Xie *et al.* [30] designed a framework which employed a Spatial Partitioning Scheme and Spatial Pyramid Matching for scene recognition. They used a low-level visual descriptor and examined various autoencoders to encode low level features into mid-level features. They then performed scene recognition using high level image signatures via modified Spatial Pyramid and the Normalization of mid-level features. Zhang *et al.* [31] proposed object distance and a kernel-based framework that builds an object bank representation to discriminate multiple objects and traverse all pixels of an object to understand its properties. They then used object-to-class kernels to find class distances which properly classify different scenes.

III. OUR APPROACH

Firstly, preprocessing includes noise removal, smoothing and normalization of object sizes in all images of the datasets. Secondly, segmentation of the image is performed by Fuzzy C-Means and Random Forest to efficiently partition the image into segments. At the third stage, a Bag of Features approach integrates multiple orientations, blobs, Fourier transform and geometric features for extraction, and computation is applied. The last and fourth stage comprises a single/multiple object recognition incorporated Artificial Neural Network and achieves scene recognition results by Maximum Entropy.

A. PRE-PROCESSING AND NORMALIZATION

During pre-processing, the raw images in datasets are collected under different circumstances such as illumination changes [32] and contrast distribution which cause extra artifacts, high intensity values and varying scales of objects in the images (see Fig. 2(a)). To clear this unwanted information, initially, we have adjusted the dimension to 320×213 using fixed window resizing. Then, smoothing is applied using wiener filter to remove the extraneous noise (see Fig. 2(b)). Wiener filtering [33] is a linear estimation technique which minimizes the overall mean square error in the process of inverse filtering and noise smoothing which are based on a stochastic framework. Here, the Wiener Filter W (x, y) is used to find the best estimate of the original image $O(i, j)$ from the degraded image $d(i, j)$ caused by extraneous noise

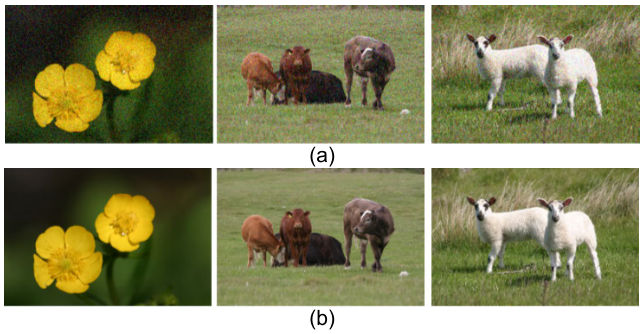


FIGURE 2. Examples of image filtering process including (a) noisy images and (b) filtered images using wiener filter over MSRC objects dataset.

$n(i, j)$. The wiener filter achieves an estimation of the original image as follows;

$$W(x, y) = \frac{p_d(x, y) - p_n(x, y)}{p_d(x, y)} \tag{1}$$

where $p_d(x, y)$ and $p_n(x, y)$ are the power spectra of the degraded image and the extraneous noise as shown in Fig. 3.

B. MULTIPLE OBJECT SEGMENTATION

After the pre-processing phase, multiple objects segmentation [34] algorithms are discussed in which we divide the whole image contents into different parts, i.e., regions or clusters based on various high level and low-level features (i.e., color, pixel orientation and pixel intensity) using Thresholding and Clustering techniques. Here, regions and locations of objects are defined through segmentation and these extracted regions are used for further classification tasks. In this paper, two novel techniques, namely, Fuzzy C-Means using Super-Pixels and Random Forest for multiple objects segmentation, are described as follows.

1) MODIFIED FUZZY C-MEAN ALGORITHM

In this section, we proposed a Modified Fuzzy C-Means based on Super-Pixels along with Mahalanobis distance [35]. Conventional Fuzzy C-Means is a clustering algorithm [36], [37] which uses unsupervised learning and cluster data via minimization criteria and cluster centers. We have modified it by using the super-pixels as input as the MFCM used less computational time compared to Fuzzy C-Means. Super-pixels are achieved applying Mahalanobis distance over the input images. Computational time is shown in Table 2. In the proposed MFCM, we performed segmentation using the objective function J_{MFCM} which minimizes the weighted distance of the total data points X that contain Super-Pixels (as shown in Fig. 3), number of clusters c, cluster centers p_i and membership matrix U which are defined as;

$$X = \{x_1, x_2, \ldots, x_N, \quad P = \{p_1, p, \ldots, p_c \tag{2}$$

$$U = \left[u_{ij}\right] \in [0, 1]^{c \times N} \tag{3}$$

$$J_{MFCM} = \sum_{j}^{N} \sum_{i}^{P} u_{ij}^{m} d_M^2 \tag{4}$$

FIGURE 3. Few examples of super-pixel extraction from the images.

FIGURE 4. Multiple objects segmentation using MFCM based on Super-pixel and Mahalanobis distance.

using

$$\forall j \in \{1, \ldots, N\}, i \in \{1, \ldots, c\} : 1 \geq u_{ij} \geq 0 \tag{5}$$

$$\forall j \in \{1, \ldots, N\} : \sum_{i=1}^{c} u_{ij} = 1, \quad 1 < m < \infty \tag{6}$$

where N is the data points, P is the total number of classes, u_{ij} is the membership degree of point x_i in the j^{th} cluster, m is the weight that represents the degree of fuzziness and d_M is the Mahalanobis distance between given datapoint x_i which is represented as;

$$d_M = (x_i - V)^T \sum^{-1} (x_i - V) \tag{7}$$

$$\sum = \frac{1}{N} \sum_{j=1}^{N} (x_i - V)(x_i - V)^T \tag{8}$$

where V shows the mean vector for all samples. Fig. 4 indicates multiple clusters of objects with different colors using proposed MFSM object segmentation. Iterations of the minimization function of MFCM are carried out according to the Algorithm 1.

2) RANDOM FOREST FOR MULTIPLE OBJECT SEGMENTATION

For pixel-based object segmentation, we also used a Random Forest (RF) algorithm [38], [39] that deals with the multi-class object segmentation problem based on similarity measure [40] between the image patches. Fig. 5 demonstrates the flow diagram for RF. A forest is a combination of decision trees T where each T consists of root nodes, splitting nodes and leaves. Each node contains a set of features (i.e., pixel intensity difference, HOG and SIFT) and a threshold value to be classified. Thus, the set of binary decision trees T, entire training set X_{train}, features X and total number of classes Y are shown as;

$$T = \{t_1, \ldots, t_T \tag{11}$$

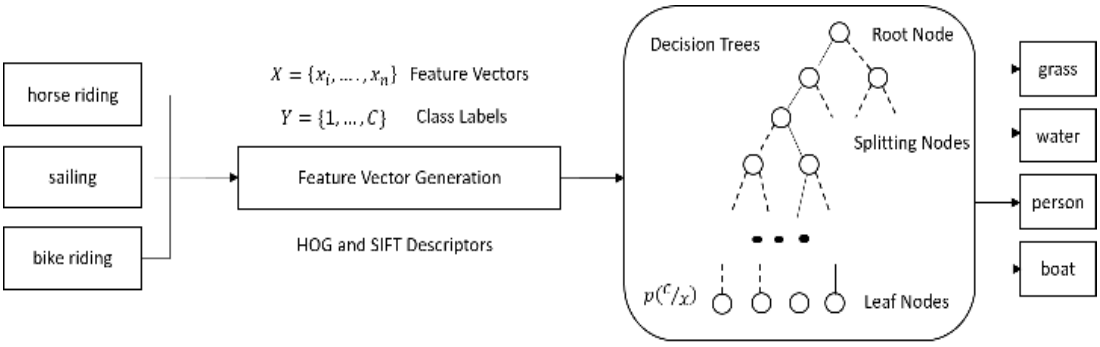


FIGURE 5. General flow of the Random Forest algorithm for object segmentation.

Algorithm 1 Modified Fuzzy C-Mean

Input: All super-pixels X, number of clusters c and fuzziness m $(1 < m < \infty)$

Output: Multiple classes based on clusters and distance d_M

1. Randomly initialize the center of clusters p_i^0 and initialize the termination threshold $\varepsilon > 0$ and fix iteration limit K.
2. $k = 1$
3. Compute the membership U^k using the centers p_i^{k-1}

$$u_{ij}^k = \frac{1}{\sum_{l=1}^{c} \left(\frac{\|x_j - p_i^k\|}{\|x_j - p_l^k\|}\right)^{\frac{2}{m-1}}} \quad 1 \le i \le c, 1 \le j \le N \tag{9}$$

4. Update the cluster center p_i^k using membership matrix U^k

$$p_i^k = \frac{\sum_{j=1}^{N} \left(u_{ij}^{k-1}\right)^m x_j}{\sum_{j=1}^{N} \left(u_{ij}^{k-1}\right)^m}, 1 \le i \le c \tag{10}$$

5. If $\left\|p_i^k - p_i^{k-1}\right\| < \varepsilon$ or $k > K$ then stop

Else $k = k + 1$ and go to step 3.

$$X = \{x_i, \ldots, x_n, \quad Y = \{1, \ldots, C\} \tag{12}$$

$$X_{train} \subseteq X \times Y \tag{13}$$

To train the RF [41], two working nodes (i.e., splitting nodes and leaf nodes) are used for the binary decision function which assigns the present sample to the left or the right node. We have evaluated a set of random decision functions over each sample in order to find a suitable decision based on information gain $H(I)$ which is given as;

$$H(I) = \sum_{c=1}^{C} p_c(1 - p_c) \tag{14}$$

where p_c is the probability of class c and $H(I)$ has chosen the multiple threshold values θ and weight hypothesis W randomly to properly train RFs as $x^T W \le \theta$ to select the best hypothesis for each labeled object class (See Fig. 6). For

FIGURE 6. Multiple objects segmentation using Random Forest over a Pascal-VOC12 scene dataset.

all input vectors, each decision tree t uses the appropriate decision function and assigns a class probability $p_t(^C/_x)$ at the leaf node which is represented as;

$$p\left(^C/_x\right) = \frac{1}{T} p_t\left(^C/_x\right) \tag{15}$$

C. BAG OF FEATURES EXTRACTION

To extract the valuable Bag of Features (BoF), we computed different low-level and geometrical shape features over MFCM segmented objects for better recognition of multiple objects. These features incorporate various characteristics such as extreme concatenate points, invariant properties, multiple orientations, and geometrical displacement values. Each of these features is described as follows.

1) GEOMETRIC SHAPE EXTRACTION

In geometric shape features, we calculated four different extreme points (upper left, lower left, upper right and lower right) and extracted their location values. Then, Euclidian distance [42] measures the distance between the left point $(x1_l, y1_l)$ and the right point $(x2_r, y2_r)$ of the segmented object and is formulated as;

$$\|d_E\| = \sqrt{(x1_l - x2_r)^2 + (y1_l - y2_r)^2} \tag{16}$$

where d_E represents the Euclidian distance and $(x1_l, y1_l)$ and $(x2_r, y2_r)$ are the x, y coordinates of first and second points, respectively. Fig. 7 shows different shapes with the junction of points of multiple objects in the images.

2) DISCRETE FOURIER TRANSFORM

Discrete Fourier Transform (DFT) is used to convert images (spatial domain) into Frequency domains. In frequency

FIGURE 7. Geometrical shape features extraction using Euclidian distance.

Algorithm 2 Bag of Feature Extraction

Input: Segmented images
Output: Feature Vector containing Bag of Features

1) F1 = ExtractGeometricalShapePoints
2) F2 = CalculateFrequenciesByFourierTransform
3) F3 = GenerateMultipleOrientationByGaborFilter
4) F4 = ExtractBlobsOverImages
5) V_i = CreateFeatureVectors f_i)
6) For each f_i in features do
   {
   Concatenate = (V_i, V_{i+1})
   }

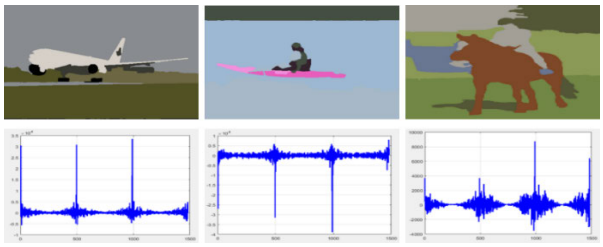


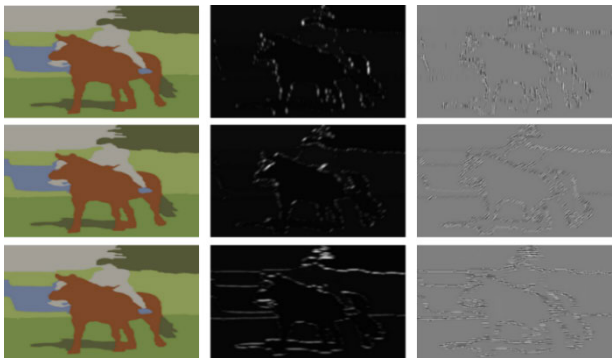
FIGURE 8. Plots of discrete Fourier transform feature having distinct variations among Pascal-VOC12 images.

domains [43], images are decamped to their cosine and sine components and each point (pixel) of the images represents a frequency value with respect to a particular time. Fig. 8 establishes distinct feature vectors of each class by effectively differentiating the frequency values of the shape of each labeled object. Thus, for an image of N pixels separated at sample time t, the DFT for the signal (in the frequency domain) $f(t)$ of the image is given as;

$$\mathrm{DFT}(kl) = \int_{-\infty}^{\infty} f(t)e^{-klt}\,dt \tag{17}$$

3) MULTIPLE ORIENTATIONS USING GABOR FILTER

Among multiple local orientation features, we applied a set of Gabor filters [44] to extract different orientations where Gabor angle $\theta = [00, 450, 900, 1350]$. It measures different frequencies of a particular phase just like a band-pass filter. We have considered 40 Gabor filters [45] having 4 orientations where angles $\theta = [0°, 45°, 90°, 135°]$ and 10 different



(a) (b) (c)

FIGURE 9. Examples of features extraction over an image using Gabor filter (a) segmented images, (b) different phases and (c) magnitude of orientations of images.

FIGURE 10. Multiple blob extraction over objects.

scales to extract the optimal features. It is represented as;

$$G(i,j) = \alpha e^{-\frac{(i^2+j^2)}{2s^2}} \sin\left(2\pi f_\circ\left(i\cos\theta + j\cos\theta\right)\right) \tag{18}$$

where α is a normalizing factor, s is for scale, $f_\circ$ represents the frequency and θ represents orientation. Fig. 9 demonstrates multiple orientations for 45,90 and 135 degree for the horse-riding scene of a Pascal-VOC12 dataset.

4) BLOB EXTRACTION

The basic purpose of blob extraction [46] techniques is to isolate and extract the homogenous regions that have distinct feature. Blobs are extracted over the images in the shape of ellipses with different sizes and they represent the pixels of images that belong to one of many discrete regions. The size of the ellipses can be adjusted by scale parameter. These blobs can be filtered, counted, and tracked. Fig. 10 represents multiple blobs of different scales over the objects.

D. MULTIPLE OBJECTS RECOGNITION BASED ON ARTIFICIAL NEURAL NETWORK

Artificial Neural Network [47] is a computational model, which is used for statistical data modeling over non-linear data. It is a machine learning tool which is inspired by the human brain system and performs learning by replicating the learning system of the human brain. ANN finds different patterns of data or relationships between input and output using artificial inter-connected neurons. ANN consists of input, output and one or more hidden layers. The input layer is transformed into output layer through the hidden layers. ANN can be implemented using different algorithms, and we used a feed forward form multi-layer perceptron (MLP) [48] algorithm to accomplish the recognition of multiple objects

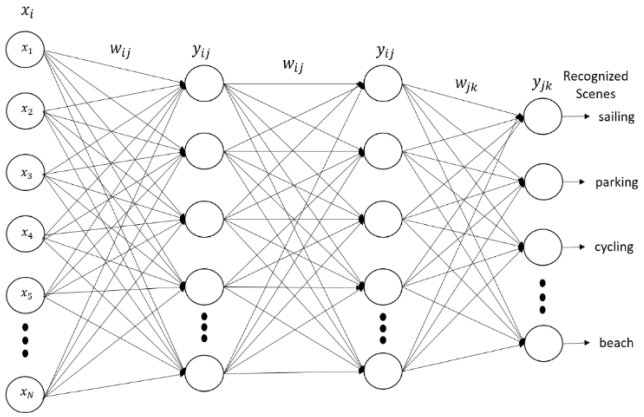


FIGURE 11. Block diagram of the Artificial Neural Network for multiple objects recognition.

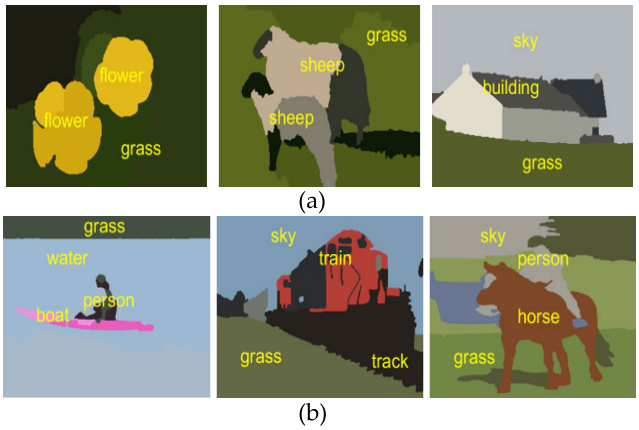


FIGURE 12. Some results of multiple object recognition with class label over the objects using ANN over (a) MSRC and (b) Pascal-VOC12 scene dataset.

patterns. Our MLP model consists of input layer, two hidden layers and an output layer. Fig. 11 shows the overall framework of the ANN algorithm applied for object recognition.

The back-propagation algorithm (BPA) is used for the training of MLP by minimizing the Mean Square Error (MSE) [49] between the actual output and predicted output for an input to the network, based on gradient iterations. The Batch training method (BTM) [50] is used for training.

BTM accelerates the speed of training and the convergence of mean square error to the desired value. The training of MLP is carried out through two broad passes, a feed forward pass and a backward calculation and updating of weights with MSE determination. The iterative updating of weights continues until the desired value or performance is achieved. The mathematical derivation of the of error e_{kn} and MSE is given as;

$$e_{kN} = o_{jN} - y_{kN} \tag{19}$$

$$\text{MSE} = \frac{1}{2M} \sum_{k=1}^{M_k} \sum_{n=1}^{N_n} e_{kN}^2 \tag{20}$$

where $o_j = [o_{j1}, o_{j2}, \ldots . o_{jN}]$ is the desired output for j^{th} input pattern, y_{kN} shows the output as the final recognized objects using k^{th} layer, N is the total input units, M_k is the total output unit and N_n is the total input patterns. It repeats until it meets the performance criteria. Fig. 12 shows a few examples of object recognition using ANN over MSRC and the Pascal-VOC12 scene dataset.

E. SCENE RECOGNITION VIA MAXIMUM ENTROPY (ME) METHOD

A maximum entropy [51], [53] method is used for the estimation of the posterior distribution [53] of class labels given to the objects in the images using a Bag of Features x_i in section 5. In the training, ME regulated the statistics of the Bag of Features set and remained as uniform as possible in the testing. ME estimates the posterior distribution (statistics) of all objects in the images based on Bag of Features x_i, then makes a decision and gives the scene label [54] using the object with maximum posterior distribution [55]. We used $x_i(I, c_k)$ a set of feature functions, where I shows the image and c_k object class labels in the image. To achieve the posterior distribution of a given Bag of Features of an object, the expected distribution [56] estimate $P(c|I)$ is matched to those observed in the training set T. The average value of Bag of Features x_i in the training set T is given as;

$$x_i = \frac{1}{|T|} \sum_{I \in T} x_i(I, c_k(I)) \tag{21}$$

Expected value of x_i over the training set c_k given as;

$$E\,[x_i] = \frac{1}{|T|} \sum_{I \in T} \sum_{c} P(c_k|I) x_i(I, c_k) \tag{22}$$

On the other hand, we have achieved the P(c|I) having maximum conditional entropy as;

$$H = -\frac{1}{|T|} \sum_{I \in T} \sum_{c} P(c_k|I) \log P(I, c_k) \tag{23}$$

where H shows the constraints $E[x_i] = x_i$. Then, desired distribution in exponential form is given as;

$$P(c_k|I) = \frac{1}{Nr} exp(\sum_{k} \lambda_i x_i(I, c_k)) \tag{24}$$

where Nr is a normalizing factor and λ_i are parameters, which are achieved under the exponential model after maximizing the likely values of the training data. After achieving posterior distribution of all objects of a scene, object-to-object relations (OOR) [57], [58] are determined using contextual information of objects to increase the scene recognition accuracy. OOR is important for complex scene, which contains co-occurrent visual patterns such as a bike is likely to be on road while not fly in sky and boat is likely to be over water while not over a road. OOR is achieved by using visual cues (i.e., area, size) and relative object position. Firstly, we applied the dot product operation to find the appearance weight between target object j^{th} for $j \in \{1, 2, \ldots, n$ to another object i^{th} for $i \in \{1, 2, \ldots, n$ as,

$$w_t(j, i) = \frac{f_j . f_i}{d(j, i)} \tag{25}$$

where f_j and f_i are the visual features of j^{th} and i^{th} object respectively. $d(j, i)$ is the distance between the j^{th} and i^{th} object. Then, relation of j^{th} object to other objects is achieved as,

$$R_j = \sum_i w_t\,(j, i)\,.f_i^n \tag{26}$$

where f_i^n are the visual features of i^{th} object. After determining relation between the object, we have achieved scene types such as sailing (OOR as grass, water, person, boat), train track (OOR as sky, grass, train, track) and air plane runway (OOR as air plane, grass, runway road, sky). Fig. 13 shows scene recognition using ME and OOR.

IV. EXPERIMENTAL SETUP AND RESULTS

Experiments are carried-out over a hardware system equipped with Intel Core i7 at 5 (GHz) CPU, and 16 GB RAM having a 64-bit Windows-10 operating system. All experiments are performed using different methods and libraries of image processing in Matlab. For a thorough evaluation of the proposed framework, we performed different sets of experiments, i.e., classification accuracy, precision, recall and F1 score by considering the Leave One Subject Out (LOSO) cross-validation scheme. For training and testing in LOSO, we divided the whole dataset into N subsets. Each subset contains k number of images. The proposed model was trained on all subsets except for one subset which was used as the test set and predictions were made for that set. We then repeated the process N times leaving out a different subset as the single test set each time. During the experimental evaluation, Precision was achieved as the ratio of accurately predicted positive instances to the total predicted positive instances. Sensitivity was achieved as the ratio of accurately predicted positive instances to the total number of instances in the actual class such as true positive and false negative. Specificity was achieved as the ratio of accurately predicted negative instances to all wrong predicted instances. The F1 score was achieved as the weighted average of precision and recall as;

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{27}$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{28}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \tag{29}$$

$$F1\ score = \frac{2(Precision \times Recall)}{(Precision + Recall)} \tag{30}$$

To evaluate the performance of the proposed architecture, we considered three challenging object recognition datasets, i.e., MSRC [59] and Caltech 101 [60] and one scene recognition dataset, i.e., Pascal-VOC12 [61]. In the following subsections, we describe dataset details, multiple experiments of performance results over these datasets and the comparison of the proposed SSR method with other SOTA SSR methods.

FIGURE 13. Scene recognition based on multiple objects using the maximum entropy method expresses scene recognition examples via the maximum entropy method.

FIGURE 14. A set of sample images from the MSRC dataset with corresponding ground-truth.

A. DATASETS DESCRIPTION

We have considered two object datasets (i.e., MSRC and Caltech 101) and one scene dataset (i.e., Pascal-VOC12). The details of these datasets are given as follows:

1) MSRC OBJECTS DATASET

The MSRC dataset includes 591 different object images from real-world environments such as a hilly scene, sea, road sign and houses. The dataset has fifteen object classes such as cow, dog, grass, bench, duck, flower, water, sky, cat, sign, tree, bird, boat, car and building. The MSRC contains 213 × 320 pixel resolution convoluted images with multiple objects. Fig. 14 shows some examples of images with their corresponding ground-truth annotations of the MSCR dataset.

2) CALTECH 101 OBJECTS DATASET

The Caltech 101 dataset consists of images with multiple categories, which are split into object and background categories. Each image has a resolution of about 300×200 pixels. The Caltech 101 dataset includes multiple object classes like camera, cell phone, barrel, bass, fish, cup, elephant, bike, panda, rhino, strawberry, airplane, water, tree, and watch. Fig. 15 shows a few examples of images with their corresponding ground-truth annotations of the Caltech 101 dataset.

3) PASCAL-VOC12 SCENE DATASET

The Pascal-VOC12 dataset is used to recognize visual scene classes in realistic environments based on different objects. The experimental evaluations include fifteen different scene classes such as sailing, horse riding, air plane runway, bike riding, cycling, train track, beach, road traffic, forest, sea, sports, plants, flying plane, city and parking with different image resolutions. The specific goal of experiments is to identify the overall scene type of image based on multiple objects. Fig. 16 presents some scene images with their corresponding ground-truth annotations from the Pascal-VOC12 scenes dataset.

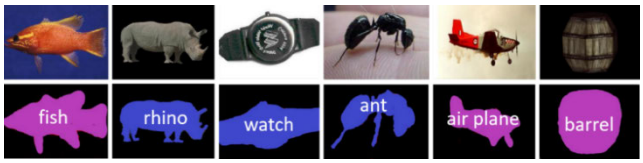


FIGURE 15. A set of sample images from the Caltech 101 dataset with corresponding ground-truth.

FIGURE 16. A set of sample images from the Pascal-VOC12 dataset with corresponding ground-truth annotations.

TABLE 1. Comparison of objects segmentation average accuracies to competitive methods over three benchmark datasets.

Datasets	CSFCM Method [62]	PNC Model [63]	Bag of Contour [64]	Our Method	
				MFCM	RF
MSRC	-	78.00	83.00	84.14	77.50
Caltech 101	76.41	-	-	82.74	74.90
Pascal-VOC12	-	-	43.00	84.51	76.10

B. PARAMETER SETTINGS AND EVALUATION

The proposed system is evaluated against various parameters, i.e., object segmentation accuracies, Computational time of segmentation algorithms, classification accuracy, precision, recall and F1 score to observe maximum distinguishable patterns in the performance of all the datasets.

1) ABLATION EXPERIMENTS

In this section, we compared our method with other SOTA methods. We conducted experiments over three datasets to compare segmentation, object and scene recognition accuracy.

a: SEGMENTATION ACCURACY

We evaluated our method, CSFCM method, PNC Model, Multiple Kernel Learning and Bag of Contour method based on object segmentation and the experimental results are summarized in Tables 1 and 2. As demonstrated in Tables 1 and 2, our method yields higher accuracy of segmentation and lessens computational time for segmentation compared to other methods over similar datasets.

b: OBJECT RECOGNITION ACCURACY

We evaluated the GA based Joint Classifier, Hierarchical abstract semantic model, Image Classification with ELM and CSIFT, Context-aware Network, SPM+SVM model, Adoptive Discriminant Analysis method and our method for object recognition accuracy over the MSRC and Caltech 101 datasets. Table 3 summarizes the objects recognition results for each individual class of MSRC in the form of a confusion matrix; it achieves a mean accuracy of 90.07%. From the experimental results, it can be observed that our proposed bag of features can fairly distinguish different object classes of the MSRC dataset. Some confusion is perceived between closer pairs of objects like cow, dog, and cat; boat and car; and water and sky, but the overall results are quite remarkable. The proposed method is used over 15 different objects which obtained the best classification accuracy of 89.26% over the Caltech101 dataset as shown in Table 4. However, it can be observed that due to significant feature behaviors, a few objects classes, i.e., barrel, bike, strawberry, and tree achieved the highest accuracy rates which are positively reflected in their recognition performance. We have also evaluated the precision, recall and F1 score parameters for the MSRC and Caltech 101 datasets. Tables 5 and 6 shows the precision, recall and F1 scores of all classes used in the MSRC and Caltech 101 datasets respectively. Table 7 shows the comparison of the proposed method with other SOTA methods. The significant performance improvement verifies the efficiency of proposed model.

TABLE 2. Comparison of average segmentation computation time to competitive methods over three benchmark datasets.

Bag of Contour [64]	MKL Method [65]	Our Method	
		MFCM	RF
4.00m per image	4.80m per image	59.65s per class	83.36s per class

c: SCENE RECOGNITION ACCURACY

We evaluated our method, Spatial Sampling Network, HCP-Alex, PSP-Net, PSA-Net, MSC-GPA and Fisher-Net method over the Pascal-VOC 12 dataset for scene recognition accuracy. Table 8 presents the comparison results between the proposed method and the SOTA method over the Pascal-VOC12 scenes dataset while Table 9 depicts scene recognition performance over 15 different scenes with a mean accuracy of 93.53% using a Maximum Entropy classifier. Here, a few scenes boost overall performance due to efficient (OOR) with less uncertainty and maximum conditional entropy such as sailing (i.e., OOR as grass, water, person, boat), train track (i.e., OOR as sky, grass, train, track) and air plane runway (i.e., OOR as air plane, grass, runway road, sky). Table 10 shows the precision, recall and F1 score for all classes used in the Pascal-VOC12 dataset. The significant performance improvement verifies the efficiency of the proposed model over other SOTA methods.

TABLE 3. Confusion matrix of objects recognition accuracies over the msrc dataset using ANN.

Objects	co	do	gr	be	du	fl	wt	sk	ct	si	tr	bo	ca	bi	bu
co	**0.91**	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00
do	0.04	**0.89**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00
gr	0.06	0.00	**0.87**	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.03	0.00	0.01	0.00	0.00
be	0.00	0.00	0.01	**0.92**	0.00	0.01	0.00	0.01	0.00	0.02	0.01	0.02	0.00	0.00	0.00
du	0.00	0.03	0.02	0.00	**0.88**	0.00	0.05	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.00
fl	0.00	0.00	0.05	0.00	0.00	**0.91**	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.00	0.00
wt	0.00	0.00	0.01	0.00	0.00	0.00	**0.92**	0.05	0.00	0.00	0.01	0.01	0.00	0.00	0.00
sk	0.00	0.00	0.03	0.00	0.00	0.00	0.03	**0.89**	0.00	0.00	0.02	0.00	0.00	0.03	0.00
ct	0.01	0.04	0.00	0.00	0.02	0.00	0.00	0.00	**0.90**	0.00	0.00	0.00	0.00	0.01	0.00
si	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	**0.91**	0.00	0.00	0.00	0.00	0.02
tr	0.00	0.00	0.06	0.00	0.00	0.02	0.00	0.02	0.00	0.00	**0.90**	0.00	0.00	0.00	0.00
bo	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	**0.88**	0.06	0.00	0.03
ca	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.02	0.00	0.02	**0.90**	0.00	0.02
bi	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	**0.92**	0.00
bu	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.02	0.01	0.00	**0.91**

Mean objects recognition accuracy = 90.07%

*co =cow; do =dog; gr =grass; be = bench; du =duck; fl =flower; wt =water; sk =sky; ct =cat; si =sign; tr =tree; bo =boat; ca =car; bi =bird; bu =building.

TABLE 4. Confusion matrix of individual object class accuracies over the caltech 101 dataset using ANN.

Objects	ca	ce	ba	bs	ch	cu	el	bi	pa	rh	st	ar	wa	tr	wt
ca	**0.87**	0.07	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
ce	0.06	**0.88**	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.02
ba	0.00	0.00	**0.93**	0.06	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
bs	0.00	0.00	0.10	**0.85**	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
ch	0.03	0.03	0.00	0.00	**0.88**	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02
cu	0.03	0.02	0.02	0.00	0.00	**0.90**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
el	0.00	0.00	0.00	0.00	0.00	0.00	**0.89**	0.00	0.04	0.07	0.00	0.00	0.00	0.00	0.00
bi	0.00	0.00	0.01	0.00	0.02	0.00	0.00	**0.90**	0.00	0.00	0.00	0.07	0.00	0.00	0.00
pa	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	**0.92**	0.05	0.00	0.00	0.00	0.00	0.00
rh	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.05	**0.89**	0.00	0.00	0.00	0.00	0.00
st	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	**0.90**	0.00	0.00	0.03	0.04
ar	0.00	0.00	0.00	0.01	0.02	0.00	0.02	0.04	0.00	0.00	0.00	**0.89**	0.01	0.01	0.00
wa	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.89**	0.11	0.00
tr	0.00	0.00	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	**0.92**	0.00
wt	0.03	0.03	0.01	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	**0.88**

Mean objects recognition accuracy = 89.26%

*ca = camera; ce = cell phone; ba = barrel; bs = bass; ch = chair; cu = cup; el = elephant; bi = bike; pa = panda; rh = rhino; st = strawberry, ar = airplane; wa = water; tr = tree; wt = watch

C. IMPACT OF THE NUMBER OF IMAGE SAMPLES PER CLASS

Experiments of the proposed model performed on k number of image samples per class shows that the distribution of image samples must be adequate, as inadequate distribution of image samples could affect the performance of the system. Increasing the number of image samples could increase the accuracy of the system. In this section, we evaluate the impact of the number of image samples by varying the value of k from 5 to 25, as demonstrated in Fig. 17.

TABLE 5. Measurements of precision, recall and F1 score of proposed method over the msrc dataset.

Class	Precision	Sensitivity	Specificity	F1 score	Class	Precision	Sensitivity	Specificity	F1 score
co	0.843	0.812	0.893	0.827	ct	0.807	0.759	0.823	0.782
do	0.828	0.799	0.876	0.813	si	0.843	0.812	0.880	0.827
gr	0.816	0.770	0.875	0.792	tr	0.816	0.770	0.864	0.792
be	0.763	0.735	0.862	0.748	bo	0.816	0.770	0.870	0.792
du	0.785	0.749	0.810	0.767	ca	0.799	0.757	0.810	0.777
fl	0.810	0.778	0.892	0.794	bi	0.855	0.817	0.902	0.836
wt	0.807	0.759	0.853	0.782	bu	0.770	0.738	0.794	0.754
sk	0.763	0.755	0.873	0.748					
Mean Precision = 0.808		**Mean Recall = 0.772**			**Mean Specificity = 0.858**			**Mean F1 score = 0.789**	

TABLE 6. Measurements of precision, recall and F1 score of proposed method over the Caltech 101 dataset.

Class	Precision	Sensitivity	Specificity	F1 score	Class	Precision	Sensitivity	Specificity	F1 score
ca	0.783	0.728	0.774	0.754	pa	0.754	0.709	0.805	0.731
ce	0.796	0.752	0.818	0.773	rh	0.775	0.722	0.775	0.748
ba	0.828	0.779	0.810	0.803	st	0.812	0.761	0.829	0.761
bs	0.768	0.718	0.764	0.752	ar	0.754	0.709	0.794	0.709
ch	0.731	0.685	0.770	0.707	wa	0.801	0.759	0.820	0.759
cu	0.796	0.752	0.821	0.773	tr	0.820	0.767	0.819	0.793
el	0.801	0.759	0.819	0.779	wt	0.768	0.718	0.801	0.752
bi	0.812	0.761	0.793	0.786					
Mean Precision = 0.787		**Mean Sensitivity = 0.739**			**Mean Specificity = 0.800**			**Mean F1 score = 0.767**	

TABLE 7. Comparison of object recognition accuracy of proposed method with other SOTA methods over the MSRC and Caltech 101 datasets.

Methods	MSRC	Caltech 101
GA based Joint Classifier [66]	75.00	.
Hierarchical abstract semantic model [67]	77.00	76.00
Image Classification with ELM and CSIFT [68]	.	78.00
Context-aware Network [69]	.	80.00
SPM+SVM [70]	90.3	77.93
Adoptive Discriminant Analysis [71]	92.59	87.24
Proposed Method	**90.07**	**89.26**

TABLE 8. Comparison of scene recognition accuracy of the proposed method with other sota methods over the PASCAL-VOC12 dataset.

Methods	Pascal-VOC12
Spatial Sampling Network [72]	63.54
HCP-Alex [73]	81.80
PSP-Net [74]	82.60
PSA-Net [75]	85.70
MSC-GPA [76]	82.40
Fisher-Net [77]	92.90
Proposed Method	**93.53**

As shown in Fig. 17, the accuracy of the proposed system increased significantly as the value of k varies from 5 to 15 but it became slower when value is greater than 15. Thus, we conclude that changes in the number of image samples can affect the accuracy of the system.

V. DISCUSSIONS

Our adaptation of the Scene Semantic Recognition framework is designed to obtain high F1 scores and recognition accuracy by considering all regions/objects in the images. Initially, datasets are denoised using the Wiener filter. We then passed the input images through the segmentation section. First, we applied the Fuzzy C-Mean for segmentation but it cost high computational time because it works at the basic pixel level. We modified the Fuzzy C-Mean to decrease the computational time by using super-pixels as input. These super-pixels are achieve applying the Mahalanobis distance

TABLE 9. Confusion matrix of individual scene class of accuracies over the PASCAL-VOC12 dataset using maximum entropy.

Scenes	sa	ho	ai	bi	cy	tr	be	ro	fo	se	sp	pl	fl	ci	pa
sa	**0.95**	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00
ho	0.00	**0.90**	0.00	0.01	0.00	0.00	0.00	0.02	0.03	0.00	0.04	0.00	0.00	0.00	0.00
ai	0.00	0.00	**0.92**	0.03	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.03	0.01	0.00
bi	0.00	0.02	0.00	**0.89**	0.02	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.01	0.00
cy	0.00	0.02	0.00	0.03	**0.90**	0.00	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.02	0.00
tr	0.00	0.00	0.01	0.00	0.00	**0.93**	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.03	0.00
be	0.03	0.00	0.00	0.00	0.00	0.00	**0.90**	0.00	0.02	0.03	0.01	0.00	0.00	0.00	0.01
ro	0.00	0.00	0.00	0.04	0.02	0.00	0.00	**0.88**	0.00	0.00	0.02	0.00	0.00	0.02	0.02
fo	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.01	**0.92**	0.00	0.01	0.01	0.00	0.01	0.01
se	0.01	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	**0.93**	0.00	0.01	0.01	0.00	0.02
sp	0.00	0.03	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	**0.91**	0.00	0.00	0.02	0.00
pl	0.02	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.00	**0.89**	0.00	0.01	0.02
fl	0.00	0.00	0.09	0.00	0.00	0.03	000	0.00	0.00	0.00	0.00	0.00	**0.88**	0.00	0.00
ci	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.03	0.01	0.00	0.02	0.00	0.00	**0.87**	0.04
pa	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.00	0.01	0.00	0.02	**0.91**

Mean scene recognition accuracy = 93.53%

*sa = sailing; ho = horse riding; ai = air plane runway; bi = bike riding; cy = cycling; tr = train track; be = beach; ro = road traffic; fo = forest; se = sea; sp = sports; pl = plants; fl = flaying plane; ci = city; pa = parking.

TABLE 10. Measurements of precision, recall and F1 score of the proposed method over the PASCAL-VOC12 dataset.

Class	Precision	Sensitivity	Specificity	F1 score	Class	Precision	Sensitivity	Specificity	F1 score
Sa	0.887	0.853	0.923	0.870	**fo**	0.830	0.801	0.799	0.815
Ho	0.830	0.801	0.905	0.815	**ce**	0.851	0.820	0.876	0.835
Ai	0.872	0.836	0.911	0.853	**sp**	0.824	0.795	0.855	0.809
Bi	0.830	0.798	0.901	0.815	**pl**	0.801	0.774	0.810	0.787
cy	0.790	0.764	0.856	0.777	**fl**	0.824	0.795	0.904	0.809
tr	0.872	0.836	0.891	0.854	**ci**	0.814	0.781	0.885	0.797
be	0.842	0.811	0.843	0.826	**pa**	0.824	0.795	0.819	0.809
ro	0.814	0.781	0.850	0.797					
Mean Precision = 0.834		**Mean Recall = 0.803**			**Mean Specificity = 0.868**			**Mean F1 score = 0.818**	

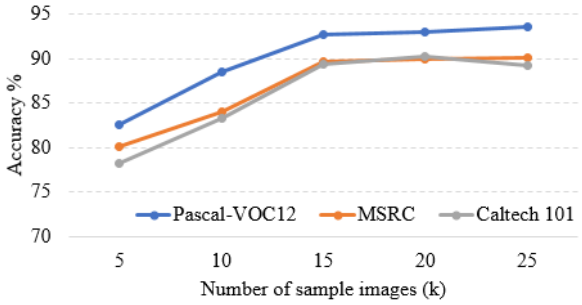


FIGURE 17. Effect of image samples on accuracy.

over the images. We also applied the RF to compare segmentation results. The results of MFCM were improved, so we used these results for further implementation. Furthermore, by using the novel bag of features, the proposed system shows robustness regarding changes in illumination, shadows, occlusion of multiple objects, direction, and changes in the size of objects. Then, the Patterns of segmented objects are considered to assign class labeling to all objects using ANN. Finally, scenes are recognized through the ME model and object-to-object relations. Thus, our approach has achieved remarkable accuracy with less computational time.

VI. CONCLUSION

In this paper, we proposed a novel and effective framework that robustly segments the location of objects, generates a new Bag of Features, and recognizes complex scene scenarios using five steps. Firstly, preprocessing is carried out over the images to remove extra artifacts and noise. We have used a linear estimation technique to minimize the mean square error and stochastic framework for smoothing the images through Wiener Filtering. Secondly, we explored Fuzzy C-Means encapsulated with random forest to split the objects into different regions during the object's segmentation problem. After segmentation various are passed through a feature extraction section of the proposed model and these segmented objects are analyzed by the dynamic geometrical shapes, discrete Fourier transform, dominant orientations and blobs of objects features to examine the directional properties, invariant characteristics and object size normalization among these Bag of Features. Fourthly, we employed an ANN model for objects recognition through different patterns of data and relationships between input and output using artificial inter-connected neurons. Finally, Maximum Entropy estimates the class labels of all scenes. We adopted three benchmark datasets MSRC, Caltech 101 and Pascal-VOC12 for training and testing of the proposed framework. Our proposed model demonstrated remarkable performance in terms of computation, segmentation and accuracy compared to statistically well-known SOTA systems.

In future work, we plan to investigate new features such as entropy-based features, depth, and energy features of multiple objects to improve region extraction and object recognition accuracy. The effectiveness of Maximum Entropy as a predictor and estimator may be improved using deep learning methods for overall scene recognition based on object-to-object and object-to-scene relations over real-world scenarios such as crowd detection and event surveillance. We also plan to apply the proposed framework over Depth and RGB-D datasets for scene recognition.

REFERENCES

[1] M. Rusci, D. Rossi, M. Lecca, M. Gottardi, E. Farella, and L. Benini, "An event-driven Ultra-Low-Power smart visual sensor," *IEEE Sensors J.*, vol. 16, no. 13, pp. 5344–5353, Jul. 2016.

[2] A. Ahmed, A. Jalal, and A. A. Rafique, "Salient segmentation based object detection and recognition using hybrid genetic transform," in *Proc. Int. Conf. Appl. Eng. Math. (ICAEM)*, Aug. 2019, pp. 203–208.

[3] S. Song, C. Hu, and M. Q.-H. Meng, "Multiple objects positioning and identification method based on magnetic localization system," *IEEE Trans. Magn.*, vol. 52, no. 10, pp. 1–4, Oct. 2016.

[4] R. E. O'Donnell, A. Clement, and J. R. Brockmole, "Semantic and functional relationships among objects increase the capacity of visual working memory," *J. Exp. Psychol., Learn., Memory, Cognition*, vol. 44, no. 7, pp. 1151–1158, Jul. 2018.

[5] V. Ramanathan, C. Li, J. Deng, W. Han, Z. Li, K. Gu, Y. Song, S. Bengio, C. Rossenberg, and L. Fei-Fei, "Learning semantic relationships for better action retrieval in images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1100–1109.

[6] X. Song, S. Jiang, and L. Herranz, "Multi-scale multi-feature context modeling for scene recognition in the semantic manifold," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2721–2735, Jun. 2017.

[7] A. Ahmed, A. Jalal, and K. Kim, "A novel statistical method for scene classification based on multi-object categorization and logistic regression," *Sensors*, vol. 20, no. 14, p. 3871, Jul. 2020.

[8] G. Liu, S. Liu, K. Muhammad, A. K. Sangaiah, and F. Doctor, "Object tracking in vary lighting conditions for fog based intelligent surveillance of public spaces," *IEEE Access*, vol. 6, pp. 29283–29296, 2018.

[9] Q. Mi and D. Xue, "A sound-based video clipping framework toward sports scenes," *J. Vis. Commun. Image Represent.*, vol. 55, pp. 648–653, Aug. 2018.

[10] Z. Yang, T. Dan, and Y. Yang, "Multi-temporal remote sensing image registration using deep convolutional features," *IEEE Access*, vol. 6, pp. 38544–38555, 2018.

[11] R. Dube, D. Dugas, E. Stumm, J. Nieto, R. Siegwart, and C. Cadena, "SegMatch: Segment based place recognition in 3D point clouds," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 5266–5272.

[12] Y. Guo, J. Ji, X. Lu, H. Huo, T. Fang, and D. Li, "Global-local attention network for aerial scene classification," *IEEE Access*, vol. 7, pp. 67200–67212, 2019.

[13] U. Erkan, D. N. H. Thanh, L. M. Hieu, and S. Enginoglu, "An iterative mean filter for image denoising," *IEEE Access*, vol. 7, pp. 167847–167859, 2019.

[14] Y. Niu, C. Su, and W. Guo, "Salient object segmentation based on superpixel and background connectivity prior," *IEEE Access*, vol. 6, pp. 56170–56183, 2018.

[15] M. Li, Z. Fang, and S. Lu, "An accurate object detector with effective feature extraction by intrinsic prior knowledge," *IEEE Access*, vol. 8, pp. 130607–130615, 2020.

[16] B. Liu, S. He, D. He, Y. Zhang, and M. Guizani, "A spark-based parallel fuzzy C-means segmentation algorithm for agricultural image big data," *IEEE Access*, vol. 7, pp. 42169–42180, 2019.

[17] F. Zhao, P. Gao, H. Hu, X. He, Y. Hou, and X. He, "Efficient kidney segmentation in micro-CT based on multi-atlas registration and random forests," *IEEE Access*, vol. 6, pp. 43712–43723, 2018.

[18] A. Sagheer, M. Zidan, and M. M. Abdelsamea, "A novel autonomous perceptron model for pattern classification applications," *Entropy*, vol. 21, no. 8, p. 763, Aug. 2019.

[19] S. B. U. D. Tahir, A. Jalal, and K. Kim, "Wearable inertial sensors for daily activity analysis based on adam optimization and the maximum entropy Markov model," *Entropy*, vol. 22, no. 5, p. 579, May 2020.

[20] T. Liu, Z. Yuan, J. Sun, J. Wang, and N. Zheng, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[21] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *Proc. CVPR*, Jul. 2017, pp. 2386–2395.

[22] M. Xu, M. Guo, L. Shang, and X. Jia, "Multi-value image segmentation based on FCM algorithm and graph cut theory," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2016, pp. 1333–1340.

[23] R. Khan, C. Barat, D. Muselet, and C. Ducottet, "Spatial histograms of soft pairwise similar patches to improve the bag-of-visual-words model," *Comput. Vis. Image Understand.*, vol. 132, pp. 102–112, Mar. 2015.

[24] W. Li, P. Dong, B. Xiao, and L. Zhou, "Object recognition based on the region of interest and optimal bag of words model," *Neurocomputing*, vol. 172, pp. 271–280, Jan. 2016.

[25] P. Espinace, T. Kollar, A. Soto, and N. Roy, "Indoor scene recognition through object detection," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 1406–1413.

[26] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *Int. J. Comput. Vis.*, vol. 81, no. 1, pp. 2–23, Dec. 2007.

[27] G. Chen, X. Song, H. Zeng, and S. Jiang, "Scene recognition with prototype-agnostic scene layout," *IEEE Trans. Image Process.*, vol. 29, pp. 5877–5888, 2020.

[28] R. Kachouri, M. Soua, and M. Akil, "Unsupervised image segmentation based on local pixel clustering and low-level region merging," in *Proc. 2nd Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, Mar. 2016, pp. 177–182.

[29] L.-J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. Oct. 2007, pp. 1–8.

[30] L. Xie, F. Lee, L. Liu, Z. Yin, Y. Yan, W. Wang, J. Zhao, and Q. Chen, "Improved spatial pyramid matching for scene recognition," *Pattern Recognit.*, vol. 82, pp. 118–129, Oct. 2018.

[31] L. Zhang, X. Zhen, and L. Shao, "Learning object-to-class kernels for scene classification," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3241–3253, Aug. 2014.

[32] W. Zhou, A. Zyner, S. Worrall, and E. Nebot, "Adapting semantic segmentation models for changes in illumination and camera perspective," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 461–468, Apr. 2019.

[33] N. J. Nyunt, Y. Sugiura, and T. Shimamura, "Parametric Wiener filter based on image power spectrum sparsity," *J. Signal Process.*, vol. 22, no. 6, pp. 287–297, Nov. 2018.

[34] A. Ahmed, A. Jalal, and K. Kim, "Region and decision tree-based segmentations for multi-objects detection and classification in outdoor scenes," in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, Dec. 2019, pp. 209–2095.

[35] S.-D. Wu, C.-W. Wu, T.-Y. Wu, and C.-C. Wang, "Multi-scale analysis based ball bearing defect diagnostics using mahalanobis distance and support vector machine," *Entropy*, vol. 15, no. 2, pp. 416–433, Jan. 2013.

[36] Y. Zhang, Z. Li, J. Cai, and J. Wang, "Image segmentation based on FCM with mahalanobis distance," in *Proc. Int. Conf. Inf. Comput. Appl.*, 2010, pp. 205–212.

[37] A. N. Benaichouche, H. Oulhadj, and P. Siarry, "Improved spatial fuzzy *C*-means clustering for image segmentation using PSO initialization, mahalanobis distance and post-segmentation correction," *Digit. Signal Process.*, vol. 23, no. 5, pp. 1390–1400, Sep. 2013.

[38] J. Santner, M. Unger, T. Pock, C. Leistner, A. Saffari, and H. Bischof, "Interactive texture segmentation using random forests and total variation," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 1–12.

[39] L. Lefkovits, S. Lefkovits, S. Emerich, and M. F. Vaida, "Random forest feature selection approach for image segmentation," in *Proc. 9th Int. Conf. Mach. Vis. (ICMV)*, Mar. 2017, Art. no. 1034117.

[40] N. Shnain, Z. Hussain, and S. Lu, "A feature-based structural measure: An image similarity measure for face recognition," *Appl. Sci.*, vol. 7, no. 8, p. 786, Aug. 2017.

[41] X. Liu, M. Song, D. Tao, J. Bu, and C. Chen, "Random geometric prior forest for multiclass object segmentation," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3060–3070, Oct. 2015.

[42] A. A. Rafique, A. Jalal, and A. Ahmed, "Scene understanding and recognition: Statistical segmented model using geometrical features and Gaussian Naïve bayes," in *Proc. Int. Conf. Appl. Eng. Math. (ICAEM)*, Aug. 2019, pp. 225–230.

[43] S. S. Khan, M. Khan, and Q. Ran, "Multi-focus color image fusion using Laplacian filter and discrete Fourier transformation with qualitative error image metrics," in *Proc. 2nd Int. Conf. Control Comput. Vis. (ICCCV)*, 2019, pp. 41–45.

[44] V. John, J. Toyota Technological InstituteNagoya, A. Boyali, and S. Mita, "Gabor filter and gershgorin disk-based convolutional filter constraining for image classification," *Int. J. Mach. Learn. Comput.*, vol. 7, no. 4, pp. 55–60, Oct. 2017.

[45] A. C. Turlapaty, H. K. Goru, and B. Gokaraju, "Gabor filter based entropy and energy features for basic scene recognition," in *Proc. AIPRW*, Oct. 2016, pp. 1–4.

[46] P. Szuster, "Blob extraction algorithm in detection of convective cells for data fusion," *J. Telecommun. Inf. Technol.*, vol. 4, no. 2019, pp. 65–73, Dec. 2019.

[47] M. Barthakur, T. Thakuria, and K. K. Sarma, "Artificial Neural Network (ANN) based object recognition using multiple feature sets," in *Proc. CTVS*, 2012, pp. 127–135.

[48] L. Jiann-Der, "Object recognition using a neural network with optimal feature extraction," *Math. Comput. Model.*, vol. 25, no. 12, pp. 105–117, Jun. 1997.

[49] M. Nuutinen, T. Virtanen, and J. Häkkinen, "Performance measure of image and video quality assessment algorithms: Subjective root-mean-square error," *J. Electron. Imag.*, vol. 25, no. 2, Mar. 2016, Art. no. 023012.

[50] C. W. Lee, "Training feedforward neural networks: An algorithm giving improved generalization," *Neural Netw.*, vol. 10, no. 1, pp. 61–68, Jan. 1997.

[51] S. Lazebnik, C. Schmid, and J. Ponce, "A maximum entropy framework for part-based texture and object recognition," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Oct. 2005, pp. 832–838.

[52] F. A. N. Palmieri and D. Ciuonzo, "Objective priors from maximum entropy in data classification," *Inf. Fusion*, vol. 14, no. 2, pp. 186–198, Apr. 2013.

[53] W.-H. Chu, Y.-J. Li, J.-C. Chang, and Y.-C.-F. Wang, "Spot and learn: A maximum-entropy patch sampler for few-shot image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6251–6260.

[54] M. Brown and S. Susstrunk, "Multi-spectral SIFT for scene category recognition," in *Proc. CVPR*, Jun. 2011, pp. 177–184.

[55] A. Jalal, N. Khalid, and K. Kim, "Automatic recognition of human interaction via hybrid descriptors and maximum entropy Markov model using depth sensors," *Entropy*, vol. 22, no. 8, p. 817, Jul. 2020.

[56] A. Rajan, Y. Chow Kuang, M. Po-Leen Ooi, and S. N. Demidenko, "Moments and maximum entropy method for expanded uncertainty estimation in measurements," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, May 2017, pp. 1–6.

[57] Y. Zhang, J. Kong, M. Qi, Y. Liu, J. Wang, and Y. Lu, "Object detection based on multiple information fusion net," *Appl. Sci.*, vol. 10, no. 1, p. 418, Jan. 2020.

[58] Y. Liu, R. Wang, S. Shan, and X. Chen, "Structure inference net: Object detection using scene-level context and instance-level relationships," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6985–6994.

[59] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "*TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Proc. ECCV*, 2006, pp. 1–15.

[60] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, 2004, p. 178.

[61] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jun. 2014.

[62] H. Abdellahoum, N. Mokhtari, A. Brahimi, and A. Boukra, "CSFCM: An improved fuzzy C-Means image segmentation algorithm using a cooperative approach," *Expert Syst. Appl.*, vol. 166, Mar. 2021, Art. no. 114063.

[63] P. Aliniya and P. Razzaghi, "Parametric and nonparametric context models: A unified approach to scene parsing," *Pattern Recognit.*, vol. 84, pp. 165–181, Dec. 2018.

[64] Q. Yu, C. Yang, H. Fan, H. Zhu, F. Ye, and H. Wei, "Bag of contour fragments for improvement of object segmentation," *Int. J. Speech Technol.*, vol. 50, no. 1, pp. 203–221, Jan. 2020.

[65] A. Jain, S. V. N. Vishwanathan, and M. Varma, "SPF-GMKL: Generalized multiple kernel learning with a million kernels," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 750–758.

[66] A. V. Luong, T. T. Nguyen, X. C. Pham, T. T. T. Nguyen, A. W. C. Liew, and B. Stantic, "Automatic image region annotation by genetic algorithm-based joint classifier and feature selection in ensemble system," in *Proc. IIDS*, 2018, pp. 599–609.

[67] Z. Ye, P. Liu, W. Zhao, and X. Tang, "Hierarchical abstract semantic model for image classification," *J. Electron. Imag.*, vol. 24, no. 5, Oct. 2015, Art. no. 053022.

[68] Q. Li, Q. Peng, J. Chen, and C. Yan, "Improving image classification accuracy with ELM and CSIFT," *Comput. Sci. Eng.*, vol. 21, no. 5, pp. 26–34, Sep. 2019.

[69] J. Leng, Y. Liu, and S. Chen, "Context-aware attention network for image recognition," *Neural Comput. Appl.*, vol. 31, no. 12, pp. 9295–9305, Jun. 2019.

[70] C. Wu, Y. Li, Z. Zhao, and B. Liu, "Image classification method rationally utilizing spatial information of the image," *Multimedia Tools Appl.*, vol. 78, no. 14, pp. 19181–19199, Jul. 2019.

[71] D. Xie, Q. Li, W. Xia, S. Pang, H. He, and Q. Gao, "Multi-view classification via adaptive discriminant analysis," *IEEE Access*, vol. 7, pp. 36702–36709, 2019.

[72] D. Mazzini and R. Schettini, "Spatial sampling network for fast scene understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1286–1296.

[73] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "HCP: A flexible CNN framework for multi-label image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1901–1907, Sep. 2016.

[74] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[75] H. Zhao, Y. Zhang, S. Liu, and J. Shi, "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. ECCV*. Cham, Switzerland: Springer, 2018, pp. 270–286.

[76] Q. Zhou, B. Zheng, W. Zhu, and L. Jan Latecki, "Multi-scale context for scene labeling via flexible segmentation graph," *Pattern Recognit.*, vol. 59, pp. 312–324, Nov. 2016.

[77] P. Tang, X. Wang, B. Shi, X. Bai, W. Liu, and Z. Tu, "Deep FisherNet for image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 2244–2250, Jul. 2019.

AHMAD JALAL received the Ph.D. degree from the Department of Biomedical Engineering, Kyung Hee University, South Korea. He was working as a Postdoctoral Research Fellowship with POSTECH. He is currently an Associate Professor with the Department of Computer Science and Engineering, Air University, Pakistan. His research interests include multimedia contents and artificial intelligence.

ABRAR AHMED received the M.S. degree in computer science from Air University, Islamabad, Pakistan. He is currently a Research Assistant with Air University. His research interests include multimedia contents, artificial intelligence, machine learning, and computer vision.

ADNAN AHMED RAFIQUE is currently pursuing the Ph.D. degree with the Department of Computer Science, Air University, Pakistan. His research interests include artificial intelligence, machine learning, and computer vision.

KIBUM KIM (Member, IEEE) received the bachelor's degree in computer science from Korea University, Seoul, the master's degree in computer science from the University of Illinois at Urbana-Champaign, and the Ph.D. degree in computer science from the Virginia Tech. He is currently an Associate Professor with the Department of Human-Computer Interaction, Hanyang University.