Integrated Feature Boosting Algorithm for Efficient Classification of Outdoor Natural Scene Images

C. A. Laulkar [1]

Computer Science and Engineering, Walchand College of Engineering, Sangli, India

P. J. Kulkarni

Computer Science and Engineering, Walchand College of Engineering, Sangli, India

Abstract: Classification of outdoor natural scene images that share common visual objects and comprise the objects having common characteristics, is one of the many challenging tasks in computer vision. Proposed research work intends to provide solutions to these challenges to improve the performance of classification of the scene images. Outdoor natural scene images contain the objects such as sky, green land, house, rocky land, water waves, and plain water to describe the scene categories of habitation, water body, verdant and craggy land. The sharing of the sky, green land, and rocky land objects by multiple scene categories escalate the misclassification of images. A weighted neural network is proposed in this paper that updates the weights of a network according to the importance of the objects in the description of the images of the scene. An object may be recognized under multiple categories if it possesses similar characteristics to other objects. The multiple label detection for a single object leads to misclassification of the scene image. A novel feature-boosting algorithm is introduced in this paper that analyses the features of detected objects and eliminates the redundant objects from an image of the scene to boost the features of essential objects of the scene. The integration of a feature-boosting algorithm with a neural network has improved the performance of scene classification by 7.71%. The research work has also introduced a more adaptive Long Short Term Memory based method that uses the natural language representation of the scene image for the classification and it has shown equivalent performance.

Keywords: Object Detection and Recognition; YOLO; Feature Boosting; Neural Network; Semantic Rules; Scene Classification

1. Introduction

Real-life scene images are an integral part of the present human life. With the rapid technological advancements, the image data of real scene images is increasing exponentially. These images are used in applications that facilitate the day-to-day tasks of human life. Multiple scenes are captured, analyzed, and interpreted in many applications of computer vision like data retrieval, robotics, and video surveillance. Real-life natural scenes are mainly categorized into two types: indoor and outdoor scene images. Every natural scene is described by the collective presence of various objects and their relationships with each other. Over the past decades, several approaches have been proposed to extract the visual concepts with their relationships for scene classification.

The problem of image classification can be formulated in two phases: (a) selection of appropriate image features and (b) usage of selected features for the labeling of scene images. Earlier efforts show the use of low-level features like color, texture, shape, etc. in the process of scene classification. These low-level features have today, evolved into high-level features like Local Binary Pattern (LBP) [40], Scale Invariant Feature Transform (SIFT) [39], Global Image Descriptor (Gist) [30], Histogram of Oriented Gradient (HOG) [37, 38], Bag of Visual Words (BoVW) [36, 43], Centrist [41], Object Bank (OB)[32, 33], Codebook [45], Fisher Vector [42] etc. that improved the performance of scene classification. Recently, various models of CNN architecture have been proposed to classify the scene images which includes the layers for feature extraction.

Present research work intends to classify the outdoor natural-scene images comprising the objects such as sky, green land, rocky land, house, plain water, and water wave into the scene categories like habitation, craggy land, verdant, and water body using semantic relationships of the objects. A scene-image is classified into a specific category due to the presence of a mandatory object

Corresponding author: C.A.Laulkar[1]

in an image, like as for habitation: house, verdant: green land, water body: plain water or water wave, and for craggy land: rocky land. The outdoor scene categories may share multiple visual concepts like green land, sky, and rocky land as shown in Figure 1.

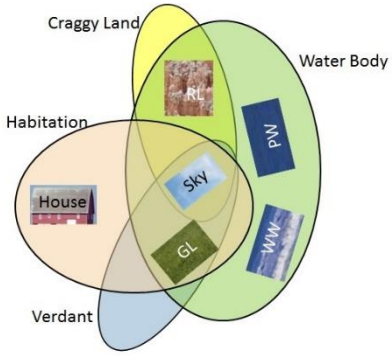

Fig. 1 Contribution of visual concepts in scene image classification

(GL-Green Land, PW- Plain Water, WW-Water Wave, RL- Rocky Land)

The proposed method of outdoor natural-scene image classification begins with the extraction of objects using CNN based model: You Only Look Once (YOLO) [44]. It extracts the objects present in a scene image along with their attributes: type, size, confidence score, and location. The attributes of the objects are applied to a feed-forward neural network (NN) for the classification of scene images. It has been observed that some of the objects are shared among multiple scene categories as shown in Figure 1, which leads to misclassification of scene images. To improve the performance of scene classification, a weighted neural network (WNN) is proposed that updates the weights of NN according to the importance of that object in a scene category. Further, it is identified that some of the objects are recognized into two or more object categories due to the similarity in their characteristics. A novel feature-boosting algorithm is designed that analyses the type, occurrence, and confidence score of the recognized objects to boost the features of the most appropriate objects. The boosted features are applied to NN and WNN for the classification of scene images.

The present work has also proposed a novel natural language-based approach that is more adaptive to the inclusion of additional objects and scene categories into the system. In this approach, an image is represented with the set of natural language sentences which are further processed through the Long Short-Term Memory (LSTM) network for the classification of scene images.

The rest of this paper is organized as follows; Section 2 reviews the work related to the methods used for the classification of scene images. Section 3 discusses the detailed architecture of the proposed method that includes feature extraction, feature boosting, and classification of scene images. Section 4 comprises the experimental results and detailed analysis of methods applied in the paper. Finally, Section 5 elicits the conclusion based on the performance of proposed methods for the classification of the scene images.

2. Related Work

The problem of scene classification is attempted by many researchers by applying holistic, semantic, and hybrid strategies, for the last three decades. The scene is described by its global features in a holistic approach, whereas in the semantic approach various regions of the scene image along with their relationships are considered. The mid-level features like histogram, orientation, and density are extracted from low-level features like color, shape, and texture, were considered for the interpretation and classification of scene images [1-3]. Image retrieval is the application that uses image classification implicitly to extract similar images for a given query image. The image contains heterogeneous features like color, texture, shape that are used together to give an efficient performance for image retrieval. The heterogeneous features extracted for a given input query image are compared with the features of the image database that retrieves the images based on individual features which are merged to generate a ranked set of similar images [24, 47]. A configural recognition scheme uses inter as well as intra qualitative and photometric relationships of the regions for the classification of scene images [25]. An image is represented as a bag of multiple instances, which is labeled as negative if all its instances are negative and it is labeled as positive if the bag contains at least one positive instance [26]. Three types of features computed for the sub-blocks of an image are as color features in Ohta color space, texture features using Multi-Resolution Autoregressive Model (MSAR), and frequency features obtained by applying 2D DFT and 2D DCT. These features are applied individually and in combination for the classification of sub-blocks. The best performance of image classification is achieved by combining the results of sub-block classification for color and frequency features [27].

The objects comprised in an image are used as high-level features by subsequent researchers for the scene classification. These object features are extracted by partitioning the image into fixed-size grids [4] or variable size regions which are generated using various methods of segmentation [5, 6]. A template of the image is a set of regions, which capture the key semantic meaning of a scene under dissimilar pose, lighting, and landscape change. Templates and their Relationship Extraction and Estimation (TREE) algorithm comprises of two algorithms: Template Extraction and Analysis (TEA) for template extraction; and Spatial Template Relation Extraction and Measuring (STREAM) for extracting template relations. Integrated Template and Relation Indexing (ITRI) automatically analyses visual and relation similarity and provides appropriate weights to combine for efficient image retrieval [28].

A hierarchical generative model is proposed [29] to combine patch level, object level, and scene level information to classify an image by recognizing, annotating, and segmenting the objects within an image.

Classification of indoor scene images is one of the challenging tasks due to the cluttered, occluded arrangement of man-made objects. It is efficiently handled by combining global (Gist descriptor) and local (spatial pyramid of visual words) image descriptors [30]. A framework is proposed to bridge the semantic gaps between the tags and images. A unified graph has been built by combining similarity graphs, that are constructed with the visual features and the image tag bipartite graph. These fused parameters are further used by the Markov random walk model to balance the influences between the image content and tags. Four types of global features: Grid color moment, Local Binary Pattern, Gabor wavelet texture, and edge are combined to form a 297-D vector, which is normalized to zero mean, and unit variance is used to build an image similarity graph. This framework can be used in applications that include image to image retrieval, image to tag retrieval, tag to image retrieval, tag to tag retrieval [31].

SIFT is used to transform the local patches formed by segmentation into 128-dimensional vectors. Distinct code-words are assigned to similar types of vectors, which are clustered together to generate a codebook called Bag of Visual Word (BoVW). The spatial information of the visual words ignored by BoVW is considered in Spatial Pyramid Matching (SPM). SPM is an order-less image representation in which an image is repeatedly subdivided and a histogram of features is calculated for each partition of the image, at each level [7]. An object bank (OB) is a high-level description of an image with high dimension features. A logistic regression method is proposed to explore feature and object scarcity in object bank representation to learn and classify complex scene images. The OB along with linear regression classifier and linear SVM classifier has achieved superior performance over GIST, BoVW, and SPM methods for LabelMe, UIUC-sports, and MIT Indoor image dataset [32, 33]. The holistic approach is good for simple scene representation. As it does not consider the relationships of internal objects, it is unable to characterize scenes with multiple objects. The object-based approach considers the relationship between the objects which can characterize complex scenes but not simple scenes. As both the approaches complement each other, their combination performs better for scene classification. Both approaches along with the SVM classifier, classify the input image separately. If the decision of both the classifiers is the same, then the result is considered as a final result, but if both the methods disagree with each other, then the final decision of image classification is taken by the voting system [34]. A reconfigurable Tangram model is presented to represent a scene layout in which a scene image lattice is tiled in a hierarchical and compositional way by using three types of shapes: triangle, rectangle, and trapezoid. The dictionary of tan instances which are seen in image lattice to explore a large number of scene layouts is organized with a directed acyclic And-Or graph (AOG). Multiple tangram templates for all scene categories presented explicitly are learned by combining exemplar-based clustering method and dynamic programming algorithm. All learned tangram templates are combined to form a tangram bank. The linear classifiers: SVM and logistic regression are employed on tangram bank for scene classification. The tangram bank representation has shown outperformance over the spatial pyramid model with OB and spatial pyramid BoVW scene representation [35].

Semantic meanings generated by local patches of the image are not sufficient for high-level applications of computer vision. Li-jia-li has applied latent Support Vector Machine (SVM) and texture-based object detector that detects the objects to form an object bank. The object bank representation of an image encodes spatial and semantic information of the objects. This image representation shows superior performance for classification of the scene on Scene-15, UIUC-sports, and MIT indoor image datasets, when combined with linear regression and linear SVM classifier [8, 9].

Traditional methods of scene classification are executed in two stages: a) feature extraction and b) classification of the scene with a machine learning algorithm. Extraction of handcrafted features requires the majority of the time and gets more importance than classifiers used for scene classification. These methods can be effectively applied to small size datasets. Automatic scene classification on large-scale dataset handling is a longstanding challenge in computer vision. Convolution Neural Network (CNN) is a type of deep neural network, which extracts global and local features and has shown good performance for large-scale datasets like ImageNet, CIFAR, Places, etc. Due to the ability of transfer learning, CNN model like AlexNet has received much attention after winning the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012[10]. Subsequently, various models of CNN like ZFNet (2013), GoogleNet (2014) [12], VGG-16 (2014 Runner), ResNet (2015) [13], ResNeXt-10 (2016), SENet (2017), PNASNet-5 (2018) have shown their best performance in ILSVRC [11]. Guo-Sen Xia (2015) has proposed a hybrid method that associates the dictionary with CNN models for scene classification and has shown enhanced performance when combined with GoogleNet or VGG-11 on MIT Indoor-67, SUN-397 datasets [14]. Many research works have attempted to boost the performance of the scene classification by combining the features extracted through CNN models with machine learning techniques. Jing sun (2016), has proposed the method for scene classification in which features are extracted using AlexNet, are classified with a Lib-SVM classifier [15]. Object as a high-level feature extracted through CNN models like AlexNet and YOLO are analyzed for scene classification, by applying semantic rules based on, object attributes like type and spatial location of object [5, 17]. The Bag of Semantic (BoS) features is derived by scoring image patches with a pre-trained CNN model that are used to transfer the object into a scene. It is further extended with a fisher vector and classified using a traditional classifier [18].

CNN architecture extracts the features which are unable to relate the visual concepts within a scene for classification. Hence, it is challenging to classify the scene images having inter-class similarity with only the CNN model. Combining context information along with the CNN model through an attention module has shown improved performance for the classification of scenes having interclass-similarity [16]. Only the features of objects are not sufficient for the classification of scene images comprising similar visual concepts; Haitao Zeng (2019) has used the scene attributes to describe the similar scene from various aspects. These attributes are modeled to represent local patches and used as complementary features to the object as well as scene features in scene recognition. The globally and locally derived scene attributes, object, and scene features are aggregated into image-level representation through the MRF encoding method and classified with linear SVM [19].

The previous research work presented various low-level, mid-level, and high-level features along with machine learning algorithms, and several techniques based on deep learning for the classification of scene images using holistic and semantic approaches. Most of the semantic approach applied for image classification uses spatial and category features of visual concepts. Also, recently proposed deep learning methods, calculates the features implicitly and perform the classification in a single iteration.

The research work presented in this paper has first time attempted the use of four attributes: type, size, confidence score, and location for the classification of scene images. In the proposed research work, the multiple visual objects present in a scene image are detected and recognized with the state-of-the-art CNN model: YOLO, along with the attributes: type, size, and confidence score. The location of the object in a scene image is calculated by logically partitioning the image into grids of 6x6. The proposed research work has also identified the causes that intensify the misclassification of scene images are: a) the sharing of the same visual object among multiple scene categories and b) the classification of visual objects into multiple object categories due to possession of common characteristics. The paper has discussed the solutions provided through current research work that improves the performance of classification of scene images are: 1) First, a newly presented *weighted neural network* method is applied to minimize the misclassification of scene images that arises due to the sharing of similar objects among multiple scene categories. During the training phase of a weighted neural network, the weights are updated according to the importance of the objects in a scene image. Thus, even though similar objects are shared by multiple categories of scene images, their weightage assists the network to classify the scene image precisely. 2) Second, a novel *feature boosting algorithm* is presented in this paper that boosts the features of precise objects by analyzing the type, occurrence, and confidence score of the objects present in a scene image. It reduces the misclassification of scene images that occurs due to the classification of the same object into multiple categories. The classification of scene images using a weighted neural network requires the estimation of weights of objects for each scene category. 3) Third, a novel *natural language-based approach* of scene classification is presented in this paper, in which the scene images are described with natural language sentences by relating the objects with each other concerning their type and position. These sentences are further processed through the LSTM network for the classification of scene images. The natural language-based approach eliminates the requirement of weight estimation for the objects of each scene category.

3. Classification of Outdoor Natural Scene Images using Semantic Approach

Natural scene image is described by the set of objects along with their organization. Proposed research work intends to classify the outdoor scene images which share the common objects as well the objects possessing similar shape and texture properties. The input color images selected for this research from the SUN-397 dataset are categorized into four classes: habitation, verdant, water body, and craggy land that comprises six types of objects: sky, green land, house, rocky land, water wave, and plain water. The multiple objects describing the scene image are extracted with YOLO along with their attributes like type, size, confidence score, and location, which are analyzed and boosted to improve the performance of scene classification. The task of scene classification is performed with two novel approaches: a) weighted neural network and b) LSTM network. For the first approach, the boosted features are applied to a weighted neural network, and the weights of the neural network are updated according to the values of weight matrix. In the second approach, the boosted features are transformed into natural language sentences using wordbook. A codebook is generated with a set of unique sentences and used to replace each sentence describing an image with corresponding code. The LSTM network is trained and tested with the codes of images to classify the scene images.

3.1. High-level feature extraction using YOLO

YOLO is a single shot multiple object detector, which identifies the objects with bounding boxes as shown in Figure 2. YOLO is extremely fast in object detection and composed of two sub-networks: a) feature extraction and b) object detection [44]. In the present research work, AlexNet is used as a pre-trained CNN for feature extraction. The object detection sub-network is a small CNN compared to the feature extraction network which replaces the layers of AlexNet from the ReLU5 layer. The detection network consists of groups of serially connected convolution, ReLU, and batch normalization layers, which are followed by a yolov2TransformLayer and a yolov2OutputLayer. The YOLO network detects multiple

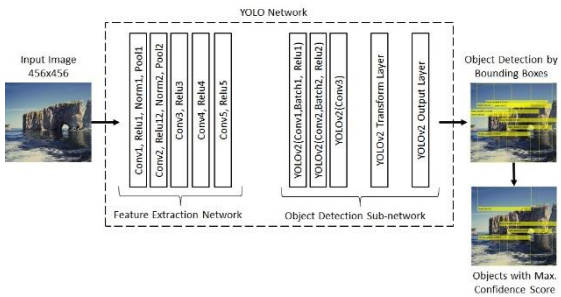

Fig. 2 Object detection and recognition using YOLO network

bounding boxes for each type of object present in the color input image of size 456x456. The bounded box with the highest confidence score is selected for each type of object. An image is logically partitioned into 36 (6x6) regions. A region enclosing the center of a bounding box is selected as the *location* of an object as shown in Figure 3. The area of an object is calculated as the difference between the sum of the area of intersection of detected bounding boxes with each other

Fig. 3 Location of objects in a scene image.

$$A(Ob) = \left(\sum_{i=1}^{n} A(Ob)_i\right) - \sum_{i=1}^{n}\sum_{j=i+1}^{n} A\left((Ob)_i \cap (Ob)_j\right) \quad (1)$$

and the sum of the area of all detected objects as given in Eq. (1). Here, *n* is the total number of the bounding boxes detected for the specific object. The four features: type, size, confidence score, and location of each detected object are used further for the classification of outdoor scene images.

3.2. Weighted Neural Network

As explained in the previous section, the objects present in a scene image are extracted with YOLO along with their attributes. The attributes are applied to a feed-forward neural network for the classification of the scene image. It has been observed that the sharing of objects among multiple scene categories affects adversely the performance of a neural network. To overcome this challenge, a weight matrix is designed that assigns the weights to each object according to its importance in a scene category as displayed in Table 1. A mandatory object of a scene image

Table 1 Weights of objects for NN

Sr. No.	Object→ / Scene Class ↓	Sky	Green land	House	Rocky land	Plain water	Water wave
1	Habitation	0.8	0.9	1	0	0	0
2	Water body	0.9	0	0	0.5	1	1
3	Verdant	0.9	1	0	0	0	0
4	Craggy land	1	0.1	0	1	0	0

is assigned with the highest value of 1 as its presence in a scene image plays a significant role in the classification of the scene image. All other objects are assigned with the values in the range of 0 to 1 according to their importance in describing a scene image e.g., the object *sky* is present in most of the outdoor scene images; so, it is assigned with next highest value whereas the object *green land* occurs very rarely in *craggy land* scene images; so, it is assigned with value 0.1 for that scene category. A weighted neural network updates the weights of the neural network during the process of training according to the values in the weight matrix as shown in Figure 4. The test images are applied to the trained weighted neural network that performs the classification of scene images more precisely.

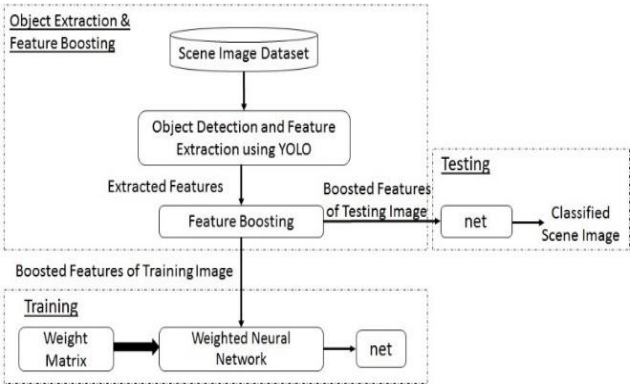

Fig. 4 Classification of scene images using weighted neural network.

3.3. Feature boosting algorithm

The semantic rules extracted for the classification of scene images are based on features of detected objects. Some images of habitation and craggy land are misclassified due to the detection of multiple labels for the same visual object that having similarity in their shape and texture with another object e.g. The object 'house' in habitation class is built up of stone detected with two labels: rocky land and house. It leads to the misclassification of an image in either *habitation* or *craggy land* class. The features of the objects are analyzed to improve the classification performance of misclassified images by boosting the features of the most probable class as shown in Figure 5 and Figure 6.

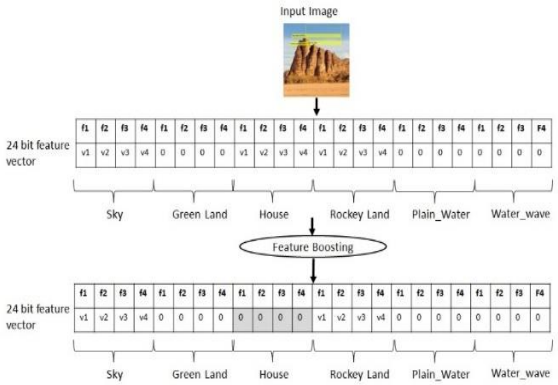

Fig. 5 Feature boosting

Feature boosting algorithm performs the analysis of features: type, confidence score, and occurrence of the objects in a scene image. The process of the feature boosting algorithm starts with the initialization of weight buckets with the value '0' for each scene category. For the occurrence of an object, the weights of the weight buckets are increased for those scene categories of which the object is a part. The feature boosting algorithm boosts the features of precise objects by eliminating the features of redundant objects, in a scene image as displayed in Figure 6.

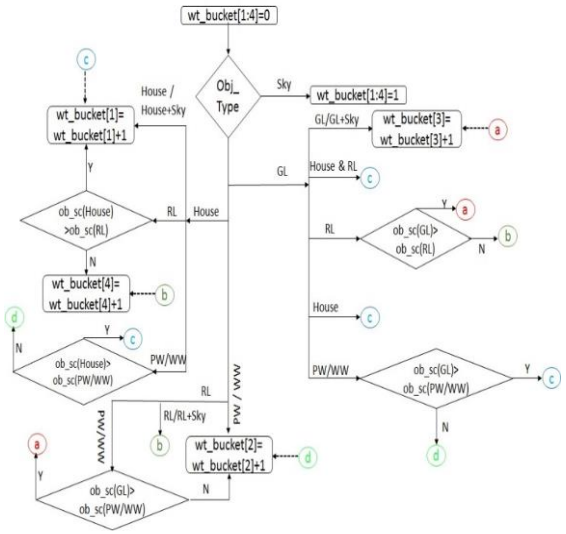

Fig. 6 Feature boosting algorithm

(GL-Green land, RL-Rocky land, PW-Plain water, WW-Water wave, ob_sc-Confidence score of object, Scene Categories [1-Habitation, 2- Water body, 3- Verdant, 4- Craggy Land])

In the process of feature boosting, if the detected object is *the sky*, then the weights of buckets of all scene categories are assigned with value '1', as the *sky* object is present mostly in all outdoor scene categories. If the *green land* object is detected and if it is the only object detected in a scene image or if it is detected with a sky object, then the weight of *the verdant* scene category is increased by '1'. For the occurrence of a *green land* object, the algorithm checks for the presence of the house, rocky land, plain water, and water wave objects. If the *house* and *rocky land* both the objects are detected, then the weight of *verdant* class is increased as the probability of occurrence of *the green land* object with the object *house* is greater than with the *rocky land* object. If either of *the house* or *rocky land* objects is present in a scene image, then the confidence score of present objects is compared with the confidence score of *the green land* object. The weight of *verdant* class is increased for the greater confidence score of *the green land* object, the weight of *habitation* class is increased for the greater confidence score of *house* object and the weight of *craggy land* class is increased for the greater confidence score of *rocky land*. In case of *plain water* and *water wave* objects, if either or both the objects are present, then with the same logic described above, their confidence scores are compared with the confidence score of *green land* object and weights of either *verdant* or *water body* scene class are updated for greater confidence score value of *green land* or *plain water/ water wave* objects respectively. For the occurrence of a house object, the presence of rocky land, plain water, and water

wave objects are checked and with similar logic as explained above, the weights of habitation, craggy land, and water body scene categories are updated. The same logic is applied for the occurrence of rocky land, plain water, and water wave objects.

After processing all the objects detected in a scene image, the weight bucket with maximum weightage is selected to identify the class of scene images. If there is more than one bucket present with identical maximum weightage values, then the confidence scores of the mandatory objects present in the scene categories are compared. The bucket of the scene category with the highest confidence score is predicted as the final image class. The features of the redundant object present in a scene image are eliminated from the feature set to boost the features of precise objects.

3.4. Natural language-based representation of scene image

Visual semantics is the study of relationships between visual concepts to understand the meaning of the image in terms of context. Abhinav Gupta (2008), has presented an EM-based approach to learn the visual classifier for nouns, prepositions, and comparative adjectives [22]. Benjamin Z. Yao (2010), has presented image parsing to text description framework to describe the scene with AoG graphical representation [23]. In the proposed research work, a pixel-level scene image is represented with a set of objects. The semantics between objects are derived from object features in the form of natural language sentences. A predefined wordbook of 13 unique words describing the type, position, and score of the object used to design the natural language sentences. A codebook of 127 codes generated for the unique sentences is used to convert the natural language representation of an image into code representations as illustrated in Figure 7 a) and b).

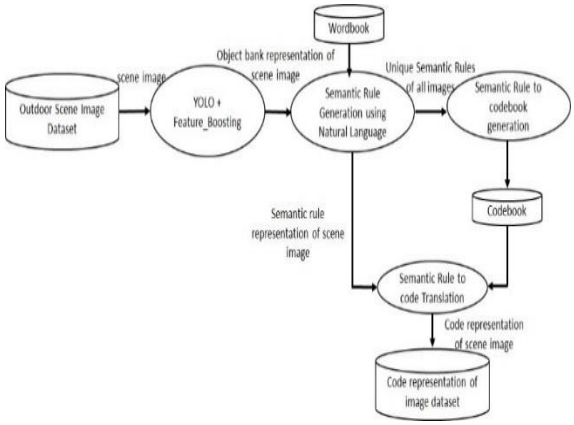

Fig. 7 a) Conversion of scene image to code image

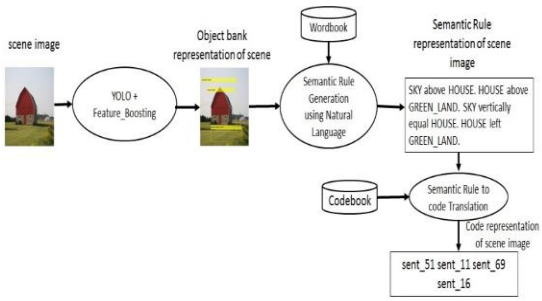

Fig. 7 b) Example of conversion of scene image to code image using codebook

3.5. Scene image classification using LSTM

A recurrent Neural Network (RNN) is a feed-forward neural network and recurring in nature that possesses internal memory to process sequential or time-series data. The problem of vanishing gradient that makes RNN incapable of learning long data sequence is resolved by LSTM which is a modified version of RNN. LSTM can learn the long-term dependencies by remembering information for long periods [20, 21, 46].

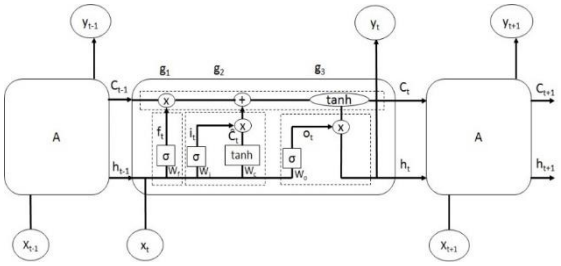

Fig. 8 Architecture of LSTM

The LSTM is built with three gates (forget gate, input gate, and output gate) that control the cell state as shown in Figure 8. The information to be omitted in from the cell at a particular timestamp is decided by the *sigmoid* function of *forget gate*. It refers to previous state h_{t-1} and current state x_t to compute f_t as given in Eq. (2) where W_f is the weight from the previous state.

$$f_t = \sigma\left(W_f . [h_{t-1}, x_t] + b_f\right) \quad (2)$$

The *sigmoid* function of *the input gate* decides which value to accept and weightage is given by *tanh* function to the values which are passed to decide their level of importance as given in Eq. (3) and Eq.(4).

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i) \quad (3)$$

$$\hat{C}_t = tanh(W_C . [h_{t-1}, x_t] + b_C) \quad (4)$$

The *output gate* decides, what will be the output of the current state. The *sigmoid* function decides which part of the cell state creates as an output o_t and the *tanh* function pushes the cell state to decide the values in between range -1 to 1 as given in Eq. (5) and Eq.(6).

$$o_t = \sigma(W_o . [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * tanh(C_t) \quad (6)$$

There are three types of LSTM model used for various applications as: vector to sequence model (e.g. image classification), sequence to vector model (e.g. text classification, sentiment classification), and sequence to sequence model (e.g. prediction of next word in sequence, language translation).

In the pre-processing phases of research work, a complete scene image containing objects is described using a set of natural language sentences. Each sentence represents the relation between two objects. The scene classification with LSTM applied on sentence description of image; relates the words of sentences. The whole scene image is described by relating all the sentences with each other. This is achieved by transforming sentence representation of the image into code representation by replacing each sentence with corresponding code in the codebook. The LSTM model processes the codes of scene image to relate the objects within the scene image for the classification of outdoor scene image into one of the four categories as shown in Figure 9.

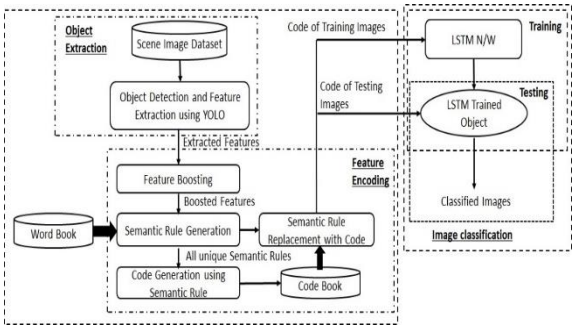

Fig. 9 Architecture of natural language-based classification of scene image with LSTM.

The architecture of the present study uses a sequence to vector model of LSTM as depicted in Figure 10. For the purpose of experimentation, the code representation image dataset is partitioned into 80% training and 20% validation dataset along with its label. As a part of word pre-processing, codes of the scene image are tokenized and converted into lower case. The word encoding method maps the vocabulary code-words into numeric indices. A sequence vector of indices is generated for each code-word image with the length of S.

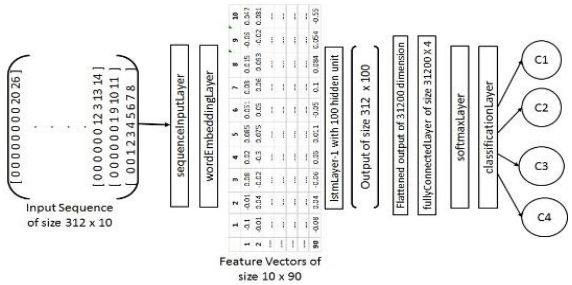

Fig. 10 Sequence to vector model of LSTM for classification of the scene image.

The input sequence of data is given to the LSTM network through *SequenceInputLayer*. A word embedding layer represents the features of words by mapping the sequence of word indices to embedding vectors of size D X S, where D is the length of features for each word. For the proposed study, the length of the feature dimension is 90 and the length of the sequence is 10. Hence, the embedding layer generates the output of size 90 X 10 for each sequence input. These vectors are given as input to the LSTM network having 100 hidden units. The hidden state can hold information from all previous time steps, irrespective of the sequence length. The LSTM layer processes the input sequence vectors for all images (312 training images) and generates the output of size 312 X 100. This output is flattened to a vector of size 31200 which is multiplied by input weight in a fully connected layer to give probabilities of each class label. The *Softmax* layer turns a vector of real values of the fully connected layer into a vector of real

values within the range of 0 to 1. The input values to the *softmax* layer can be positive, negative, zero, or greater than one, which gets transformed into the values between 0 and 1 to interpret it, as a probability. If one of the inputs is small or negative, the *softmax* turns it into a small probability, and if the input is large, then it turns it into a large probability, but it will always remain between 0 and 1. The classification layer computes the cross-entropy i.e. a measure of the difference between two probability distributions, to classify the input into one of the given class labels.

4. Experimental Results

Proposed research work is intended to classify the scene images into a habitation, verdant, water body, and craggy land. The habitation class describes the places for accommodation or storage which include the house-like structure built up of material like stone (abbey), wood (barn, chalet, shed), tree trunk (log cabin), and bricks (house, manufactured home, schoolhouse, mansion). The verdant class describes lush green land. It includes images of the athletic field, corn field, cultivated and wild field, golf course, hay field, hill, park, pasture, putting green, rice paddy, valley, and yard. It comprises the objects like the sky and green land. The water body class images shall contain plain water or water wave objects (wave and waterfall). The images are selected for this class are from the beach, canal_natural, coast, ocean, sandbar, sea cliff, waterfall_block, waterfall_fan, waterfall_plunge, and wave subclasses. The craggy land defines the area of mountains, hills, and slopes with a rock-like rough texture. The images are selected for this class are from bad_land, butte, canyon, cliff, and rock_arch.

4.1. Object detection and recognition

The performance of the YOLO network is tested for various threshold values, for various connecting layers of AlexNet (relu2, relu3, and relu5) with detection sub-network, for multiple numbers of filters (96,128,256) and different filter sizes ([3 3], [5 5]). The architecture of YOLO has shown the highest F-measure value of 78.75% as illustrated in Table 2 with the threshold value of 0.16, filter size of [3, 3], and 256 numbers of filters.

Table 2 Performance of various models of YOLO for object detection and recognition

Network Layers; Filter_cnt, Filter_size	Count & Size of Anchor Boxes	F-Measure in %						
		House	GL	Sky	RL	WW	PW	Average
relu2+3_Yolo_Layers; 256, [3 3]	5 Anchor boxes with size: [200 200;100 200;150 200;50 200;150 50]	62.74	77.67	68.47	14.91	24.82	34.47	47.18
relu3+3_Yolo_Layers; 256, [3 3]		87.15	64.15	94.67	35.05	66.34	70.22	69.60
relu5+3_Yolo_Layers; 256, [3 3]		92.77	68.66	92.71	36.97	68.00	74.78	72.32
relu5+3_Yolo_Layers;128, [3 3]		78.58	59.73	93.69	18.52	66.84	78.19	65.92
relu5+3_Yolo_Layers;128, [5 5]		84.10	62.32	91.79	31.68	71.81	76.23	69.66
relu5+5_Yolo_Layers;128, [3 3]		86.88	61.05	92.31	40.13	56.62	80.41	69.57
relu5+5_Yolo_Layers;96, [3 3]		91.77	58.38	91.65	40.29	50.09	74.88	67.84
relu5+5_Yolo_Layers;128, [5 5]		89.69	57.56	88.89	37.22	72.73	79.66	70.96
relu5+5_Yolo_Layers;96, [5 5]		91.90	65.17	89.76	21.17	53.80	79.90	66.95
relu5+3_Yolo_Layers; 256, [3 3]	Size of 5 anchor boxes is estimated on size of bounding boxes in training labeled images	87.38	83.21	**91.30**	51.29	73.25	86.05	**78.75**

4.2. Feature selection

The objects are extracted and recognized along with four attributes: type, location, size, and confidence score. For the purpose of feature selection, all these four attributes are tested individually and in combination with a neural network for the classification scene images. It is illustrated from Table 3 that the performance of scene classification is maximum with the F-measure value of 82.79% when all the four features are applied together.

4.3. Scene Classification

For the purpose of the experiment of proposed research work, 37 categories of SUN-397 dataset are merged to form the dataset of four scene categories: Habitation, Water body, Verdant and Craggy land. The image dataset is partitioned into 5660 training and 5442 testing images for the classification of outdoor natural scene images. The size of the image is considered as 456x456. The YOLO model is trained with the total number of 3757 objects of six scene categories and tested for 10772 objects. It has achieved the F-measure of 64.61% as illustrated in Table 4.

As illustrated in Table 5, the performance of scene classification is initially tested with the state-of-the-art CNN model: AlexNet having 8 layers (1 Input + 5 Convolution + 1 Fully Connected +1 Output) with a learning rate of 0.0001 and 15 epochs. It has shown a performance of 65.01%.For the classification of scene images with semantic approach, the features of extracted objects are applied to unsupervised and supervised machine learning techniques: k-means clustering algorithm, Probabilistic Neural Network (PNN), Generalized Regression Neural Networks (GRNN), Neural Network (NN), and Weighted Neural Network (WNN). The object features: type and confidence score are applied to an unsupervised K-means clustering algorithm to form 4 clusters of four scene categories. It has shown a performance of 44.79%, which is not satisfactory. Further, all the four features of objects are applied together to PNN, GRNN, and NN have shown the performance of 24.35%, 62.79%, and 66.91% respectively. The performance of NN has shown a slight improvement in F-measure value compared to the performance of AlexNet. The sharing of visual objects among multiple scene categories leads to the misclassification of scene images. A weighted neural network proposed in research work has shown an improved performance of 69.9%. As already mentioned earlier, some of the objects are classified into more than one category due to the similarities in their characteristics. A feature boosting algorithm is proposed that performs the analysis of type, confidence score, and occurrence of the objects in a scene image for the boosting of features of an appropriate object by eliminating features of redundant objects. The integration of the feature boosting algorithm along with NN and WNN has shown the improved performance of 72.72% and 72.04% respectively. The inclusion of additional objects and scene categories into the system requires the modification into the matrix of weighted neural network. A natural language-based image representation applied to LSTM is a more adaptive approach to such changes and it has shown a performance of 71.51%.

Table 3 Feature selection for scene classification

Class →	Habitation	Water_body	Verdant	Craggy_land	Average
Features ↓	F-Measure in %				
1	86.83	86.82	86.60	47.06	76.83
2	84.72	86.82	86.60	47.06	76.30
3	91.91	91.57	86.69	53.15	80.83
4	90.36	90.45	91.97	45.16	79.49
123	91.93	90.44	86.21	44.04	78.15
124	86.07	90.51	83.56	36.36	74.13
134	90.29	87.06	89.12	32.73	74.80
1234	93.54	90.82	91.30	55.48	**82.79**
* 1-Obj_type, 2- Obj_size, 3- obj_score, 4-Obj_location					

Table 4 Performance of YOLO for object detection and recognition

Object Class	Precision	Recall	F-Measure in %
House	71.07	89.54	79.24
Green_land	85.51	80.88	83.13
Sky	77.01	71.89	74.36
Rock_land	25.58	61.18	36.07
Water_wave	54.11	67.12	59.92
Plain_water	42.40	77.93	54.92
Average			64.61

Table 5 Performance of various methods used for the classification of scene images

Sr. No.	Methods / Scene Class	F-Measure in % for classification of all scene images								
		AlexNet	K-Means	PNN	GRNN	NN	WNN	FB+NN	FB+WNN	FB+LSTM
1	Craggy_land	45.12	49.37	13.73	40.80	44.91	52.38	57.37	54.93	55.57
2	Habitation	74.62	32.46	49.86	68.13	71.70	78.78	80.37	80.28	79.88
3	Verdant	74.30	52.44	22.60	74.39	77.43	76.39	78.81	78.87	77.85
4	Water_body	65.99	44.90	11.21	67.83	73.59	72.04	74.34	74.06	72.72
	Average	65.01	44.79	24.35	62.79	66.91	69.90	**72.72**	72.04	71.51

Table 6 Performance of various methods used for the classification of scene images with mandatory objects

Sr. No.	Methods / Scene Class	F-Measure in % for classification of scene images in which mandatory object detected								
		AlexNet	K-Means	PNN	GRNN	NN	WNN	FB+NN	FB+WNN	FB+LSTM
1	Craggy_land	49.62	79.77	17.86	50.77	57.22	68.43	71.29	68.37	70.11
2	Habitation	77.03	55.37	51.96	74.71	79.21	86.23	88.36	88.27	88.40
3	Verdant	80.88	82.13	26.58	85.54	89.21	88.13	90.01	90.08	90.04
4	Water_body	68.64	80.07	12.78	77.60	84.39	82.53	84.61	84.41	84.66
	Average	69.04	74.33	27.30	72.15	77.51	81.33	**83.57**	82.78	83.30

There are some images of the dataset that are: a) poor in quality, b) having very small objects, and c) captured from a very near position of an object such that the whole object is not covered. The objects present in such images are not detected with YOLO. So, to test the performance of the proposed research work for the images in which mandatory objects are detected, a modified dataset is generated. The proposed system has shown improvement in performance as illustrated in Table 6.

Compared to the results calculated on the original dataset, the modified dataset has shown improvement in F-measure value for all methods used for scene classification. Again, Table 6 illustrated that the integration of feature boosting algorithm along with neural network has shown the maximum performance of 83.57%. Overall, the FB+NN has shown the improved performance of 7.72% for the original dataset and 14.53% for the modified dataset when compared with the performance of state of the art AlexNet model.

5. Conclusion

The existing scene classification methods are mostly based on high-level features that are classified with traditional machine learning algorithms. This paper has presented the weighted neural network to diminish the misclassification of scene images that occurs due to the sharing of common objects among multiple scene categories. A novel feature boosting algorithm is introduced to boost the features of appropriate objects that improve the performance of the system for the scene classification. A new approach of LSTM based classification of scene images is proposed in this paper that uses the scene images represented with natural language sentences.

In the present work, the given image dataset is tested with standard techniques: a) unsupervised k-means clustering algorithm, b) supervised methods: PNN, GRNN, NN, and c) AlexNet CNN model. The performance of proposed methods WNN, FB+NN, FB+WNN, and FB+LSTM have shown improved performance over the standard techniques used for the classification of scene images. The proposed system has shown improvement of 7.71% and 14.53% over the performance of state-of-the-art CNN model: AlexNet with the method FB+NN for the classification of scene images when applied to original and modified image datasets respectively. The LSTM approach is more adaptive to the inclusion of additional objects and scene categories into the system. It has shown nearly equivalent performance to FB+NN.

6. Future Scope

The present work has focused on the classification of outdoor natural scene images containing objects of similar types and characteristics. In this study, the attributes of each object like type, location, size, and confidence score are considered for the classification of outdoor scene images. In the future scope of the research, additional attributes of the objects like a) total count of similar type of the object in a scene image, b) depth of the objects from each other, c) neighborhood of the objects with each other

and d) overlapping area of bounding boxes of distinct types of objects with each other can be studied for detail understanding of the scene image which will be suitable for the images containing a greater number of distinct objects. The proposed approach of research work can be extended for the classification of indoor scene images.

Declaration of Competing Interest:

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments:

The authors would like to thank Dr. M.B. Kokare and Dr. Amita Pradhan for their comments and linguistic assistance that greatly improved the manuscript.

Funding:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contributions:

For the present research work, C.A. Laulkar has presented the concept and carried out the implementation of the proposed methods. She has also carried out the experiments and generated the results that are reviewed by both the authors. She has wrote the manuscript, and prepared the figures and tables for the present work. P.J. Kulkarni has reviewed the manuscript, corrected the grammatical mistakes, and suggested the improvements in the manuscript.

References

1. Hild, M., & Shirai, Y. (1993, May). Interpretation of natural scenes using multi-parameter default models and qualitative constraints. In 1993 (4th) International Conference on Computer Vision (pp. 497-501). IEEE.
2. Gorkani, M. M., & Picard, R. W. (1994, October). Texture orientation for sorting photos" at a glance". In Proceedings of 12th International Conference on Pattern Recognition (Vol. 1, pp. 459-464). IEEE.
3. De La Escalera, A., Moreno, L. E., Salichs, M. A., & Armingol, J. M. (1997). Road traffic sign --detection and classification. IEEE transactions on industrial electronics, 44(6), 848-859.
4. Vogel, J., & Schiele, B. (2007). Semantic modeling of natural scenes for content-based image retrieval. International Journal of Computer Vision, 72(2), 133-157.
5. Laulkar, C. A., & Kulkarni, P. J. (2020). Semantic rules-based Classification of outdoor natural scene images. In Computing in Engineering and Technology (pp. 543-555). Springer, Singapore.
6. Raja, R., Roomi, S. M. M., & Dharmalakshmi, D. (2013, July). Classification and retrieval of natural scenes. In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) (pp. 1-8). IEEE.
7. Lazebnik, S., Schmid, C., & Ponce, J. (2009). Spatial pyramid matching. Object Categorization: Computer and Human Vision Perspectives, 3(4).
8. Li, L. J., Su, H., Li, F. F., & P Xing, E. (2010). Object bank: A high-level image representation for scene classification & semantic feature sparsification.
9. Li, L. J., Su, H., Lim, Y., & Fei-Fei, L. (2014). Object bank: An object-level image representation for high-level visual recognition. International journal of computer vision, 107(1), 20-39.
10. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.
11. https://machinelearningknowledge.ai/popular-image-classification-models-in-imagenet-challenge-ilsvrc-competition-history/
12. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
13. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
14. Xie, G. S., Zhang, X. Y., Yan, S., & Liu, C. L. (2015). Hybrid CNN and dictionary-based models for scene recognition and domain adaptation. IEEE Transactions on Circuits and Systems for Video Technology, 27(6), 1263-1274.
15. Sun, J., Cai, X., Sun, F., & Zhang, J. (2016, August). Scene image classification method based on Alex-Net model. In 2016 3rd International Conference on Informative and Cybernetics for Computational Social Systems (ICCSS) (pp. 363-367). IEEE.
16. López-Cifuentes, A., Escudero-Viñolo, M., Bescós, J., & García-Martín, Á. (2020). Semantic-aware scene recognition. Pattern Recognition, 102, 107256.

17. Laulkar, C. A., & Kulkarni, P. J. (2020). Integrated YOLO Based Object Detection for Semantic Outdoor Natural Scene Classification. In Applied Computer Vision and Image Processing (pp. 398-408). Springer, Singapore.
18. Dixit, M., Li, Y., & Vasconcelos, N. (2019). Semantic Fisher scores for task transfer: Using objects to classify scenes. IEEE transactions on pattern analysis and machine intelligence, 42(12), 3102-3118.
19. Zeng, H., Song, X., Chen, G., & Jiang, S. (2019). Learning Scene Attribute for Scene Recognition. IEEE Transactions on Multimedia, 22(6), 1519-1530.
20. Dong, Y., & Zhang, Q. (2019). A Combined Deep Learning Model for the Scene Classification of High-Resolution Remote Sensing Image. IEEE Geoscience and Remote Sensing Letters, 16(10), 1540-1544.
21. Chen, P. J., Ding, J. J., Hsu, H. W., Wang, C. Y., & Wang, J. C. (2017, July). Improved convolutional neural network-based scene classification using long short-term memory and label relations. In 2017 IEEE international conference on multimedia & expo workshops (ICMEW) (pp. 429-434). IEEE.
22. Gupta, A., & Davis, L. S. (2008, October). Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In European conference on computer vision (pp. 16-29). Springer, Berlin, Heidelberg.
23. Yao, B. Z., Yang, X., Lin, L., Lee, M. W., & Zhu, S. C. (2010). I2t: Image parsing to text description. Proceedings of the IEEE, 98(8), 1485-1508.
24. Sheikholeslami, G., Chang, W., & Zhang, A. (2002). Semquery: Semantic clustering and querying on heterogeneous features for visual data. IEEE Transactions on Knowledge and Data Engineering, 14(5), 988-1002.
25. Lipson, P., Grimson, E., & Sinha, P. (1997, June). Configuration-based scene classification and image indexing. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (pp. 1007-1013). IEEE.
26. Maron, O., & Ratan, A. L. (1998, July). Multiple-instance learning for natural scene classification. In ICML (Vol. 98, pp. 341-349).
27. Szummer, M., & Picard, R. W. (1998, January). Indoor-outdoor image classification. In Proceedings 1998 IEEE International Workshop on Content-Based Access of Image and Video Database (pp. 42-51). IEEE.
28. Hsieh, J. W., & Grimson, W. E. L. (2003). Spatial template extraction for image retrieval by region matching. IEEE Transactions on Image Processing, 12(11), 1404-1415.
29. Li, L. J., Socher, R., & Fei-Fei, L. (2009, June). Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 2036-2043). IEEE.
30. Quattoni, A., & Torralba, A. (2009, June). Recognizing indoor scenes. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 413-420). IEEE.
31. Ma, H., Zhu, J., Lyu, M. R. T., & King, I. (2010). Bridging the semantic gap between image contents and tags. IEEE Transactions on Multimedia, 12(5), 462-473.
32. Li, L. J., Su, H., Li, F. F., & P Xing, E. (2010). Object bank: A high-level image representation for scene classification & semantic feature sparsification.
33. Li, L. J., Su, H., Lim, Y., & Fei-Fei, L. (2010, September). Objects as attributes for scene classification. In European conference on computer vision (pp. 57-69). Springer, Berlin, Heidelberg.
34. Chen, Z., Chi, Z., Fu, H., & Feng, D. (2012, October). Combining holistic and object-based approaches for scene classification. In 2012 Fifth International Symposium on Computational Intelligence and Design (Vol. 1, pp. 65-68). IEEE.
35. Zhu, J., Wu, T., Zhu, S. C., Yang, X., & Zhang, W. (2015). A reconfigurable tangram model for scene representation and categorization. IEEE Transactions on Image Processing, 25(1), 150-166.
36. Zhu, L., Jin, H., Zheng, R., & Feng, X. (2014). Weighting scheme for image retrieval based on bag-of-visual-words. IET Image Processing, 8(9), 509-518.
37. Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886-893). IEEE.
38. Déniz, O., Bueno, G., Salido, J., & De la Torre, F. (2011). Face recognition using histograms of oriented gradients. Pattern recognition letters, 32(12), 1598-1603.
39. Xu, P., Zhang, L., Yang, K., & Yao, H. (2013). Nested-SIFT for efficient image matching and retrieval. IEEE MultiMedia, 20(3), 34-46.
40. Raja, R., Roomi, S. M. M., & Kalaiyarasi, D. (2012, December). Semantic modeling of natural scenes by local binary pattern. In 2012 International Conference on Machine Vision and Image Processing (MVIP) (pp. 169-172). IEEE.
41. Wu, J., & Rehg, J. M. (2010). Centrist: A visual descriptor for scene categorization. IEEE transactions on pattern analysis and machine intelligence, 33(8), 1489-1501.

42. Perronnin, F., & Dance, C. (2007, June). Fisher kernels on visual vocabularies for image categorization. In 2007 IEEE conference on computer vision and pattern recognition (pp. 1-8). IEEE.

43. Li, T., Mei, T., Kweon, I. S., & Hua, X. S. (2010). Contextual bag-of-words for visual categorization. IEEE Transactions on Circuits and Systems for Video Technology, 21(4), 381-392

44. Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271).

45. Jurie, F., & Triggs, B. (2005, October). Creating efficient codebooks for visual recognition. In Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 (Vol. 1, pp. 604-610). IEEE.

46. Jain, G., Sharma, M., & Agarwal, B. (2019). Optimizing semantic LSTM for spam detection. International Journal of Information Technology, 11(2), 239-250.

47. Ahmad, K., Sahu, M., Shrivastava, M., Rizvi, M. A., & Jain, V. (2020). An efficient image retrieval tool: query based image management system. International Journal of Information Technology, 12(1), 103-111.