Automatic Indoor Scene Recognition Based on Mandatory and Desirable Objects and a Simple Coding Scheme

Kathirvel N (kathir.nagaraj@gmail.com)
Anna University - BIT Campus Tiruchirappalli: Anna University Chennai - Regional Office Tiruchirappalli
Thanabal M.S
PSNA College of Engineering and Technology

Research Article

Keywords: BSR, Detection, Desirable Objects, Indoor Scene, Mandatory Objects, Recognition and Scene-number

Posted Date: June 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-474393/v1

Automatic Indoor Scene Recognition based on Mandatory and Desirable objects and a simple coding scheme

N Kathirvel[1], M.S Thanabal[2]

1 Department of Information Technology, Bharathidasan Institute of Technology, Anna University, Tiruchirappalli 620024, India,
mail_id: kathir.nagaraj@gmail.com (corresponding author)

2 Department of Computer Science and Engineering, PSNA College of Engineering Technology, Dindigul 624622, India. mail_id: msthanabal@psnacet.edu.in

Abstract:

In this paper, a simple at the same time effective recognition system for indoor scenes is presented. The proposed system has two phases, namely, creation of mandatory and desirable objects and an indoor scene recognition system. In the first phase a list of probable objects and their classification, such as mandatory and desirable objects, for any generic scene is created based on real time indoor environment clubbed with human knowledge on standard datasets. In the second phase, the proposed system contains four stages. In the first stage, the proposed indoor scene recognition system identifies and recognizes the objects of the given key frame based on simplified version of CNN architecture of YOLO v3. In the second stage, the identified objects are divided into two sets of mandatory and desirable objects with a simple dictionary look-up. In the third stage, the objects are identified to belong to a probable scene and this technique is called scene-object identification. Simple algorithms have been proposed to effect the above three stages. In the final stage, a novel Binary Scene Representation (BSR) is proposed for each of the probable scenes and the final scene recognition is obtained with a new scene-number, obtained after converting the binary BSR into decimal number system. The effect of proposed indoor scene recognition system has been experimented with standard input datasets and measured in terms of standard measures, besides comparison with existing schemes. The results are encouraging.

Keywords: BSR, Detection, Desirable Objects, Indoor Scene, Mandatory Objects, Recognition and Scene-number.

1. Introduction

Numerous approaches for indoor scene classification have been introduced over the past decennial and these approaches are facing abundant challenges. The major challenge is the accuracy rate. Indoor scene classification is a greatly challenging task due to the presence of a large number of objects in indoor scenes, background clutter, partial occlusion, viewpoint direction, and illumination and it is more complicated as compared with outdoor scenes. However, scene classification plays a vital significant role in few applications, especially for the blind people and service robots. Scene recognition provides a solution to assist object recognition problems intended for people who have very low vision or blindness. The idea behind the proposed work is to resolve the rehabilitation of blind people by classifying the indoor scenes through object detection, and hence they can do their day-to-day activities such as cocking, walking, and reading, etc. in a normal way like other people. Indoor scene classification is the process of identifying indoor scenes based on detected objects. Object detection is the process of drawing a rectangular bounding box over located indoor objects. Nowadays, mobile robots use conventional methods and CNNs in identifying different objects in their surroundings and it has partial capabilities to understand the objects. To overcome these problems, a deep learning technique has been introduced for object detection as intermediate representation and the system recognizes the scenes based on the detection of different indoor objects. So, the result is more accurate than the existing machine-learning techniques.

For the purpose of indoor scene recognition, most of the existing systems use the machine learning-techniques for object detection and to recognize the scene with classifiers such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Auto Associative Neural Network (AANN), etc. But, it produces lesser accuracy. In this work, for the purpose of improving the accuracy of scene recognition, the CNN architecture of YOLO v3 model is utilized to detect and recognize the objects such as chair, bed, TV, monitor, etc., The main contribution of indoor scene recognition system proposed in this work is the design of three novel algorithms to recognize the indoor scenes, viz., identification of mandatory and desirable objects, scene-object identification and scene recognition with a binary scene representation (BSR) coding technique. The proposed BSR coding technique is efficient and accurate, and replaces the performance of popular classifiers such as SVM and ANN and AANN and it produces the comparable results. This technique requires the set of mandatory (Mobj) and desirable objects (Dobj) to recognize the indoor scenes. Lots of efforts are carried out in our

laboratory to create these sets of Mobj and Dobj based on real time environment clubbed with human knowledge on standard datasets.

The rest of the paper is organized as follows. Section 2 describes the related literature on scene recognition. Section 3 describes the architecture of proposed indoor scene recognition system. The details of proposed phases of indoor scene recognition system are presented in Section 4. The performance measures are reviewed in Section 5. The experimentation parts carried out to evaluate the performance of the proposed ISRS with obtained results are presented in Section 6. The conclusions are drawn in Section 7.

2. Literature Survey

In this section, the literature survey related to object detection, recognition and scene classification is presented. A good number of techniques are reported in the literature for object detection. In general, object detection (Viola et al. 2001; Begum & Askarunisa, 2020) deals with feature extraction and classification.

(Viola et al. 2001) developed a rapid object detection framework with three main contributions, viz, integral image for image representation, constructing a classifier with ada boost by selecting a small number of Haar-like features with sliding window and successively combine the more complex classifiers in a single cascade structure to increase the speed of the detector by focusing attention on promising parts of the image. However, this method is not effective for detecting the tilted or turned faces. Due to lighting conditions and sub windows overlapping, there could possibly be different detections of the exact face. Hence (Dalal et al. 2005) introduced a locally normalized histogram of gradient orientation features for human detection. This approach depends on fine-scale gradient, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization in overlapping descriptor blocks learned with linear Support Vector Machine for good performance in representing the object categories. The Deformable Part Model (DPM) by (Felzenszwalb et al. 2010) described an object detection method based on mixtures of multiscale deformable part models that represent highly variable object classes. This method heavily relies on discriminative training of latent SVM classifiers for the purpose of latent information and hence matching deformable models to images. (Wang et al. 2013) explored the Regionlet method for generic object detection, learned with cascaded boosting classifier which integrates the histogram of oriented gradients (HOG) and Local Binary

Pattern (LPB) descriptors. The integration of the different features forms a regionlet and is organized as small units to delineate fine-grained spatial layouts inside objects. But the speed of this method is limited for object proposal generation and repeated CNN evaluation.

Deep learning method is one the most important and widely used methods in object detection and scene classification. (Girshick et al. 2015) introduced a Fast R-CNN based on deep convolutional network that could classify the objects in two-stages. In the first stage, it generates region proposals with selective search algorithm. But it consumes more time for object detection. In the second stage a neural network is used to extract features and to decide categories of proposal regions. However, two-stage approaches work well when objects are relatively small in images and but fail for larger images. The Faster R-CNN (Ren et al. 2017) is introduced with Regional Proposal Network (RPN) to generate the high quality region proposal to hypothesis the object location. RPN and Fast R-CNN are concatenated into single network by sharing their convolution features. (Eghbali et al. 2019) investigated the accuracy of depth of convolutional network in large scale image recognition setting. A very small filter say of size (3x3) is used to increase the depth of the architecture. Another approach called the saliency-inspired network (Erhan et al. 2014) is also reported for scalable object detection. But it is dependent on D-CNN. (Ciregan et al. 2012) predicted multiple bounding boxes along with single confidence score for each object category. (Ouyang et al. 2015) framed an architecture for generic object detection, based on deformable deep convolutional neural networks (Deepid-net). It contains deformation constrained pooling (def-pooling) layer so as to model the deformation of object parts with geometric constraint. The Deepid-net designed with a pre-training strategy to learn feature representations aims to increase the capability of detection rate. Though Deepid-net outperforms well on still images, it could not specifically be designed for object detection for videos. The T-CNN reported in (Kang et al. 2019), introduces a deep learning method based on temporal and contextual information in Tubelets obtained from videos. Temporal information is a key element for the videos especially in finding the locations and appearances of objects in videos. Contextual information is used to reduce the detection of objects with low-confidence score in the overall detection result.

Several methods have been proposed in the literature for indoor and outdoor scene classification (Quattoni et al. 2009). Some of them utilizes coordinate CNNs and LSTMs (Tong et al. 2017), Bayesian network classifier (Szummer et al. 1998), object bank based network

(Madokoro et al. 2012) and context features based on SIFT and Gist (Madokoro et al. 2012). For the purpose of classifying any scene image, the literature suggests a few techniques based on statistical properties of the scene image, without considering its constituent object and local features (Lazebnik et al. 2006; Lowe 2004; Moosmann et al. 2007; Kövesi et al. 2001). (Fu et al. 2018; You et al. 2016) used the Spatial Pyramid Matching (SPM) method for scene classification based on the popular bag-of-feature (BoF) model introduced in (Lazebnik et al. 2006). In this method the Scale Invariant Feature Transforms (SIFT) feature (Lowe 2004) is employed for feature description.

(Moosmann et al. 2007) introduced the image path quantization algorithm which is derived from K-means vector quantization. This method contains the process of selecting the patches, characterization of local visual descriptor and quantizing the feature vector learned with visual dictionary depending on extremely randomized clustering forest. (Bosch et al. 2007) addressed the issues in classifying the scene categories. In this, the shape and appearance representation are based on spatial pyramid matching over ROI (Brett et al. 2002) with added background cluttering to objects and SVM as classifier to rectify the issues. (Van Gemert et al. 2009) introduced the kernel code books with the concept of bag of discrete visual code words. This is based on the frequency distribution of visual code words used for categorizing the image. But, it has limitation in assigning the visual features to single codeword and ambiguity in assigning a codeword. (Juneja et al. 2013) developed a challenging automatic discovery of distinctive part for scene class by learning with exemplar SVM to identify part based occurrences in an image. (Liu et al. 2011) explored the region conditional random field model based on spatial interaction between the homogenous region to learn using bag-of-visual-words approach (Yang et al. 2007). (Lin et al. 2014) constructed spatial pooling regions as mid-level representation to find discriminative parts, such as sofa in hall, stove in kitchen and other distinct objects that are significant to recognize scene and classifier as SVM. The random forest (Xu et al. 2021) and SVM (Liu et al. 2019) are used as classifier for recognizing the indoor-outdoor scenes. (Espinace et al. 2013) used the common object as intermediate representation to identify the indoor scenes. For this, a method has been explored based on generative probabilistic hierarchical model, where object category classifier and contextual relations were represented to map identified objects into scenes. From the literature, it can be observed that the classical scene recognition systems could not perform well for all scenes, but these consume high computational

effort and time. On the other hand, systems that make use of machine learning techniques need more training and testing. Hence a simple scene recognition system is designed in this work, that utilizes deep learning concept initially, for the purpose of identification and recognition of objects and later recognizes the scene with look-up dictionary and a simple coding technique that entirely relies upon the content of the scene.

3 Architecture of Proposed Indoor Scene Recognition System (ISRS)

The proposed ISRS consists of 2 major phases: (i) Creating of look-up tables for the purpose of identifying mandatory and desirable objects and (ii) Recognition of indoor scenes. In the first phase, three tables are created that contain possible mandatory and desirable objects for scenes based on human knowledge. In the second stage, the key frame under consideration is subjected to proposed scene recognition system. It consists of 4 stages. They are: (i) Detection and recognition of objects, (ii) Generic identification of mandatory objects and desirable objects for any key frame, (iii) Extraction of scene mandatory objects and scene desirable objects and (iv) Recognition of indoor scenes with a newly proposed coding technique. The generic architecture of the proposed indoor scene recognition system is presented in Fig.1.

The first stage of the proposed recognition system aims to detect and recognize the objects present in the given key frame. Even though many techniques have been reported in the literature for the purpose of detection and recognition of objects, the work reported in this paper makes use of existing CNN architecture of YOLO v3 (Redmon et al. 2018). This architecture is popular due to its less complexity but higher accuracy in detecting and recognizing the objects present in the key frame under analysis. The second phase, viz., identification phase of proposed ISRS aims to identify and separate the mandatory and desirable objects from the set of all detected objects, obtained as a result of applying CNN architecture of YOLO v3. For the purpose of such an identification of objects, separate tables consisting of generic objects, set of mandatory and desirable objects are established and utilized. Based on the objects present in the Tables 3 and 4, the proposed work identifies the mandatory and desirable objects present in the key frame.

It can be observed that even though mandatory and desirable objects are identified in this stage, it is still insufficient to finalize the scene to be recognized. Hence a new algorithm is proposed in the third stage of the proposed indoor scene recognition system. The main aim of

introduction of this new algorithm is to extract the related mandatory and desirable objects so as to specify or facilitate in bringing a solution for the scene recognition. The mandatory and desirable objects of a key frame are thus identified to belong to a scene in this intermediate stage of scene recognition system. The process involved in this proposed work that aims to identify the scene for each of mandatory and desirable objects of the given key frame is termed as scene-object identification. Even though the algorithm proposed in this stage brings mandatory and desirable objects, to particularize the scene, the proposed work introduces another stage to finalize/optimize the result of scene recognition. Hence in the final stage of proposed ISRS, a new at the same time simple coding technique called Binary Scene Representation (BSR) is proposed. It utilizes the probability of a number of scene mandatory and scene desirable objects obtained in the previous stage. Thus, the fourth stage of the proposed recognition system introduces a new algorithm for the purpose of finalizing the process of indoor scene recognition besides a coding technique. A detailed description of each of these phases of proposed indoor scene recognition system is explained in the next sections.

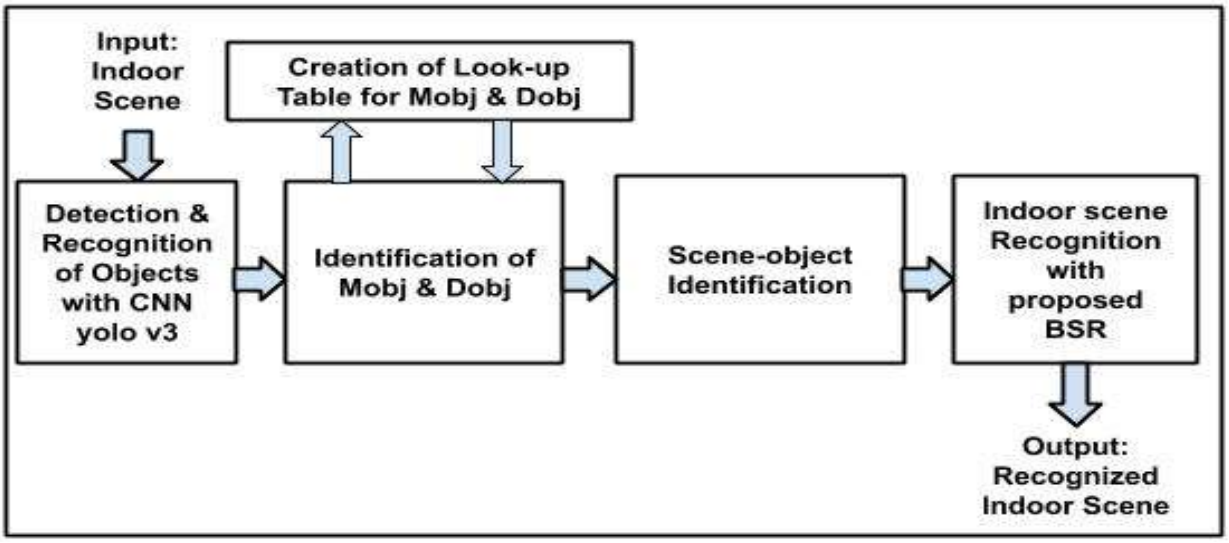

Fig.1. Proposed Architecture of Indoor Scene Recognition System.

4. Proposed Indoor Scene Recognition System

In this section a detailed description of each of the functional parts of the proposed indoor scene recognition system is presented. The ultimate aim of the indoor scene recognition system is to recognize the indoor scene present in the given key frame. The proposed indoor scene recognition system has two major phases: (i) Creation of mandatory objects and Desirable

objects and (ii) Recognition of indoor scenes. Initially, the proposed system utilizes the standard and universally accepted YOLO v3 algorithm for the purpose of detecting and recognizing the set of all objects that are present in the given key frame. The result of utilization of YOLO v3 namely bounded rectangular boxes that represent the detected objects is input to the proposed indoor scene recognition system. The first phase of the indoor scene recognition system aims to create the mandatory objects and desirable objects from the input of detected objects. In the second phase of ISRS, three novel algorithms and a simple coding technique are proposed for the purpose of recognizing the indoor scene of the given input key frame. More details of the proposed two phase indoor scene recognition system is presented in the following sections.

4.1 Creation of Mandatory and Desirable objects

The objective of creation of mandatory and desirable objects of the proposed ISRS is to identify the objects that are highly essential to point out and hence recognize the given scene. These objects are termed as mandatory objects. The remaining objects are considered as either desirable objects or objects that are of no or less importance for the purpose of recognizing the scene under analysis. The creation of such mandatory and desirable objects involves huge volume of work, especially in a generic scene scenario. However, based on real time indoor environment clubbed with human knowledge, on standard indoor datasets, a set of tables are initially designed in this work. Besides the proposed indoor scene recognition system, the creation of mandatory and desirable objects finds applications in various fields such as computer vision, scene understanding system for visually impaired and service robot. Hence lots of efforts have been carried out in our laboratory for the purpose of creation of mandatory objects and desirable objects for all the scenes. As it caters to all the indoor scenes, the work involved was separately made available as a technical paper. At the same time, the result of such creation of mandatory and desirable objects of all the scenes is very much essential in the work reported in this paper and hence a summarization is presented in tabular forms in Table.1 and Table.2, for the purpose of easy referencing. The summarization takes objects that are present in the generic indoor scenes. For the purpose of creation and summarization of mandatory and desirable objects for all the scenes, we need indoor objects for each of the scene scenario. With an extensive study, we have created a set of indoor objects that may present in a collection of the scene scenarios. A sample list of indoor objects that were identifiable in the scenes is presented in Table 3. It can be observed that the sample collection of indoor objects, reported in Table 3 refers to the scenes of

kitchen, dining hall, library, hall, toilet, laboratory, classroom, games room, etc. Even though a good number of scene scenarios are considered for this purpose, for want of space and time, the proposed work considers initially 25 scene scenarios. The proposed work assumes that every scene must have a set of mandatory objects and/or desirable objects for the purpose of initial fixing of indoor scene recognition. This initial fixing of scene scenario is termed as identification of scene-objects in this proposed work.

Table.1. List of mandatory objects
Oven, Sink, Toilet, Bed, Pillow, TV, Sofa, Table, Projector, Computer Monitor, Network switch, AC, Chair, CPU, Book, Rack, Dining table ,Chair, Tumbles, Chest instrument, Bike, Parking sign, Car, Parking sign, Mirror, Chair, Table, Telephone, Game board, Stick, Bottle, Wine glass, Bowling Ball, Pin Mirror, Heater, Screen, Projector, Syrup, Tablet, Screen, Speaker, Oven, Cake, Water area, Washing machine, Iron box, Round table, Projector, Apple, Orange.

Table.2. List of desirable objects
AC, Chair, Refrigerator, Computer Monitor Fire extinguisher, Entry/Exit board, Rack, snacks, Exit/Entry step, Screen, AC Banana, Guava, Papaya, Pomegranate, Strawberry, TV stand, Clock, Blackboard, Chair, AC, Chair, CPU, Computer Monitor, chair, Vessel, Bottle, Mug, fruits, Abdomen instrument, Rod, Left Arrow sign, Right Arrow sign, Left Arrow sign, Right Arrow sign Table, TV, Computer, Chair, Refrigerator, Mixy, Shower, Tap, night lamp, Coin, Chair, Chair, table, Scissor, Chair,

Table 3. Set of Indoor Objects.
Oven, Sink, Refrigerator, Mixy, Toilet, Sink, Shower, Tap, Bed, Pillow, lamp, TV, Sofa TV, stand, Clock, Table, Projector, Blackboard, Chair, Game board, Stick Coin, Chair Bottle, Wine glass, Chair, table, Bowling Ball, Pin, Scissor, Mirror, Heater, Screen, Projector, AC, Chair, Computer Monitor, Network switch, AC, Chair, CPU, Book, Rack, Computer Monitor, Dining table ,Chair, Vessel, Bottle, Mug, fruits, Tumbles, Chest instrument, Abdomen instrument, Rod, Bike, Parking sign, Left Arrow sign, Right Arrow sign Car, Parking sign, Left Arrow sign, Right Arrow sign, Mirror, Chair, Table, TV, Table, Telephone, Computer, Chair, Screen, Speaker, Fire extinguisher, Entry/Exit board, Oven, Cake Rack, snacks, Water area, Exit/Entry step, Washing machine, Iron box, Round table, Projector, Screen, Apple, Orange, Banana, Guava, Papaya, Pomegranate, Strawberry.

4.2 Proposed Recognition System for Indoor Scenes

The recognition system proposed in this work consists of four stages. They are: (i) Detection and recognition of objects, (ii) Generic identification of mandatory objects and desirable objects for any key frame, (iii) Extraction of scene mandatory objects and scene desirable objects and (iv) Recognition of indoor scenes with BSR, a newly proposed coding technique. The details of functional components of these stages are described in the following subsections.

4.2.1 Object Detection and Recognition

Given a key frame, this object detection and recognition stage of the proposed indoor scene recognition system aims to detect the individual objects. For the detection and recognition purposes, the proposed ISRS utilizes the CNN architecture of YOLO v3 (Redmon et al. 2018). In this stage the proposed system not only detects the objects present in the key frame under analysis, but also recognizes the objects with the same YOLO v3 algorithm. This results in producing bounding boxes to represent the objects, besides recognizing these objects, based upon the probability of certainty of such recognition of the objects. The rationale behind the utilization of YOLO v3 is its simplicity, fastness and accuracy in detecting and recognizing the objects present in the key frame. The proposed object detection and recognition stage performs training and testing procedures for this purpose. While approximately 9000 objects were trained in the original YOLO v3 algorithm, the work proposed in this stage makes use of a subset of objects that are highly relevant for a specific scenario for the purpose of training and testing procedures.

Alg 1. YOLO V3 Object Detection Algorithm.

INPUT: Indoor object image, (I_{obj}).
BEGIN:
1. Each indoor object image is given as input (75 different categories) to Darknet-53 CNN model for training (feature extractor and training).
2. Testing phase for object detection,
    i. Anchor box coordinates values $[t_x, t_y, t_w, t_h]$ are predicted.
    ii. Objectness score value is predicted using logistic regression.
3. Testing phase for object recognition,
    i. Features are extracted from 3 different scales of indoor object images.
    ii. Several convolution layers are concatenated with the base feature extractor (Daeknet-53).
4. Final layer of Darknet-53 predicts the anchor box objectness and its class.
5. K-means clustering is used here as well to find better bounding box prior.

OUTPUT: Detected indoor object with coordinates $[C_s, t_x, t_y, t_w, t_h, C_{1,} C_{2 \ldots\ldots} C_{75,}]$

This subset of objects utilized for detecting and recognizing the objects is computationally effective besides consumption of much lesser time. Thus, a simplified version of steps involved in utilizing the YOLO v3 algorithm is introduced in this stage. This simplified and at the same time cost effective version of YOLO v3 introduced for the purpose of detecting and recognizing the objects present in the key frame is presented as an algorithm in Alg. 1. The result

of the proposed object detection and recognition phase, namely set of all detected and recognized objects of the given key frame, say for example, {DRobj(i), i=1,2, …,n, n∈ I,}, I being the set of integers shall be passed on to the next stage of the proposed indoor scene recognition system for the purpose of identifying mandatory and desirable objects. The steps involved in the proposed identification stage are detailed in the next subsection.

4.2.2 Identification of Mandatory objects and Desirable objects

This stage of proposed indoor scene recognition system aims to divide the set of all {DRobj(i), i=1,2,…,n, n∈I}, I being the set of integers, obtained in the previous phase, into mandatory and desirable objects, for the input key frame under analysis. This proposed stage defines an object DRobj(i) as mandatory object, denoted as DRMobj(i), if the DRobj(i) is highly essential for the purpose of recognizing the given key frame as any of the scene. The remaining objects, DRobj(i), are defined as either desirable objects, denoted as DRDobj(j), j=1,2,…,m and m<=n or objects that are of less or no significance for the purpose of proposed scene recognition system. Under normal circumstances, this set of objects which are neither mandatory nor desirable shall be empty or it is a very small collection of DRobj(i)s. This set of objects is denoted in this proposed work as DRUobj(*l*). The proposed work ignores such a set of objects. The set of mandatory and desirable objects that are identified at this stage shall be the input to the next phase of the scene recognition system. Thus, the proposed identification of mandatory and desirable objects divides the input {DRobj(i), i=1,2,…,n} as DRMobj(j), j=1,2,….,x, DRDobj(k), k=1,2,….,y and DRUobj(*l*), *l*=1,2,….,z where x,y,z∈I and x+y+z=n. A new algorithm is designed at this stage to identify each of the given DRobj(i) as DRMobj(j) or DRDobj(k) or DRUobj(*l*). For this purpose the proposed work makes use of a table consisting of objects that classify the given DRobj(i) as mandatory object DRMobj(i).

The creation of such a mandatory objects was explained in Section 4.1 and the same is utilized here. Similarly, for the purpose of identifying the given DRobj(i) as desirable objects DRDobj(j), the proposed system makes use of Table 2, described in Section 4.1. Any object DRobj(i) that is not present in either of mandatory or desirable objects, is identified as DRUobj(*l*), set of all objects that are of less importance in the proposed system. It can be observed that the proposed system assumes the presence of some of the DRobj as neither mandatory nor desirable, unlike other related works reported in the literature. The very purpose

of separating such objects which are neither mandatory nor desirable in this work is to treat them as insignificant and hence such objects are ignored. This ignoring of object(s) is found to help in achieving reduction in space and time complexity. The steps involved in the identification of mandatory and desirable objects of the given key frame are presented as an algorithm, IMDKF, hereunder. Having separated the individual DRobj(i)s as either mandatory or desirable, the proposed system aims to identify the given key frame as an approximation to a particular indoor scene. Subsequently, in the next stage this approximate scene shall be optimized so as to recognize the indoor scene.

Algorithm- IMDKF

INPUT: Input objects DRobj(i), i=1,2,….n , obtained from stage 1.

BEGIN:

For each DRobj(i), i = 1,2,…,n

If DRobj(i) ∈ Table 3 then

DRMobj(i) ← DRobj(i)

Else If DRobj(i) ∈ Table 4 then

DRDobj(i) ← DRobj(i)

Else

Ignore the object, as it is neither mandatory nor desirable.

OUTPUT: Mandatory objects (DRMobj) and Desirable objects (DRDobj) for given key frame.

4.2.3 Scene-object Identification

The aim of scene-object identification technique of the proposed indoor scene recognition system is to identify each of DRMobj and DRDobj obtained as a result of the previous stage, as to particularize the scene that each of these objects belonging to. For example, given a mandatory object or a desirable object of the given key frame, the proposed scene-object identification technique aims to identify a scene S_i, i=1,2,…,s, $s \in I$, that the given Mobj belongs to. Similarly, the desirable object DRDobj of the given key frame shall be identified to present on a scene, say S_i, i=1,2,…,s where s is the total number of scenes. For the purpose of identifying the scene that the given DRMobj or DRDobj belongs to, the proposed system makes use of a table consisting of set of all mandatory and desirable objects that a scene has.

This table is created to specify the given mandatory and desirable objects of the input key frame to belong to a scene. As a sample, scenes S_i, i=1,2,…,s, consisting of 25 known indoor scenes, is initially created with human knowledge from the standard indoor datasets and the same

is presented in Table 4. It is evident that Table 4 specifies the set of all mandatory and desirable objects that are required to specify the scene. From the Table 4, the proposed system gives the information about the scene consisting of the required mandatory and desirable objects, so as to particularize the scene. That is, the proposed system identifies a possible scene to which the given object(s) belonging to. The steps involved in identifying the possible scenes for each of DRMobj and DRDobj, are presented as an algorithm, SOIA, here under.

Algorithm- SOIA

INPUT: $DRMob_i$, where i=1,..,x,

$DRDob_j$, where j=1,..,y where x+y<=n, n being total number of objects present in the given key frame, Table 5.

BEGIN:

Step 1: For l = 1,2,…,n

-Verify whether $DRMobj_{(i)}$ belongs to Table 5,

- If so, then mark $DRMobj_{(i)}$ as $S_{(l)}Mobj_{(i)}$, otherwise goto step 2

Step 2:

-Verify whether $DRDobj_{(j)}$ belongs to Table 5,

- If so, then mark $DRDobj_{(j)}$ as $S_{(l)}Dobj_{(j)}$, otherwise goto step 3

Step 3:

- Ignore the object, if it does not belong to Table 5.

OUTPUT: Obtained Scene Mandatory objects (OSMobj) and Desirable objects (OSDobj).

Table 4. List of Mandatory Object and Desirable Objects for Possible Scene.

MO	DO	Possible Scene.	MO	DO	Possible Scene.
Oven, Sink	Refrigerator, Mixy, bowl	Scene 1 (Kitchen)	Game board, Stick	Coin, Chair	Scene 14 (Game room)
Toilet, Sink,	Shower, Tap, Mirror	Scene 2 (Bathroom)	Bottle, Wine glass	Chair, table	Scene 15 (Bar)
Bed, Pillow	night lamp , AC, Mirror	Scene 3 (Bedroom)	Bowling Ball, Pin	-	Scene 16 (bowling court)
TV, Sofa	TV stand, Clock, night lamp	Scene 4 (Living hall)	Mirror, Heater	Scissor, Chair	Scene 17 (hair salon)
Table, Projector	Blackboard, Chair	Scene 5 (Class room)	Screen, Projector	AC, Chair	Scene 18 (Conference hall)

Computer Monitor, Network switch	AC, Chair, CPU	Scene 6 (Computer Lab)
Book, Rack	Computer Monitor, chair	Scene 7 (Library)
Dining table ,Chair,	Vessel, Bottle, Mug, fruits, Refrigerator	Scene 8 (Dining hall)
Tumbles, Chest instrument,	Abdomen instrument, Rod	Scene 9 (Gym)
Bike, Parking sign,	Left Arrow sign, Right Arrow sign	Scene 10 (two wheeler stand)
Car, Parking sign,	Left Arrow sign, Right Arrow sign	Scene 11 (Four wheeler stand)
Mirror, Chair,	Table, TV	Scene 12 (Waiting hall)
Table, Telephone	Computer, Chair	Scene 13 (Office)

Syrup, Tablet	Refrigerator, Computer Monitor	Scene 19 (Medical shop)
Screen, Speaker	Fire extinguisher, Entry/Exit board	Scene 20 (Cinema theatre)
Oven, Cake	Rack, snacks	Scene 21 (Bakery)
Water area	Exit/Entry step	Scene 22 (swimming pool)
Washing machine, Iron box	-	Scene 23 (Laundromat)
Round table, Projector,	Screen, AC	Scene 24 (Meeting room)
Apple, Orange	Banana, Guava, Papaya, Pomegranate, Strawberry	Scene 25 (Fruit shop)

It is clear from the work proposed in this stage that each of the mandatory and desirable objects of the original input key frame is identified to represent a possible scene. This identification is termed as scene-object identification in this proposed work. Even though this scene object identification could result in identifying the scenes, still we cannot recognize the given scene as all the mandatory and desirable objects of the given key frame may not particularize a scene. This identification is insufficient to recognize the given key frame even though it gives possible scene(s). For the purpose of finalizing the scene to which the given key frame belongs, a novel coding technique called Binary Scene Representation (BSR) is proposed and the same is presented in the next subsection. Thus, all the m mandatory and d desirable objects, m+d<=n, n being the total number of objects present in the given key frame may have a chance to refer to any number of scenes. Thus there needs a mechanism to devise a scene finalization and hence a scene recognition scheme from the results of objects obtained from scene object recognition stage.

4.2.4 Scene Recognition with Binary Scene Representation

For the purpose of scene recognition the proposed ISRS introduces a simple, at the same time, effective coding technique, BSR for each of the key frame under analysis. This coding technique considers all the mandatory and desirable objects present in the input key frame. In the

earlier stage, all the DRobj(i) were classified as mandatory and desirable objects, that were identified, are made to particularize scene(s). Since the input object can belong to different scenes, there arises difficulty in recognizing the given key frame to a scene, in terms of scene-object identified ones. This situation is elegantly handled by the proposed scene recognition stage of the indoor scene recognition system. The recognized scene-objects are input to this proposed scene recognition stage. That is, for example, if there are n objects in the given key frame, these n objects are identified to belong to a scene in terms of mandatory and desirable nature of objects of the scene. Since these mandatory and desirable objects belong to different scenes a simple at the same time, a novel, and effective a binary scene representation scheme is devised for each possible scene of the given key frame. The steps involved in the BSR and hence the scene recognition is presented in this sub section.

The proposed BSR coding technique introduces a new n bit binary representation, where n represents the total number of objects present in the key frame, with m number of mandatory objects and d number of desirable objects, for each of the scene Si, i=1, 2,…….s. The proposed n bit coding technique shall consist of least d bits for representing desirable objects and the m most significant bits are utilized to represent the mandatory object of the key frame under analysis. The proposed coding scheme assigns the numeral 1 for each of the mandatory object representing the scene and present in the key frame, starting from the most significant bit. Out of m number of bits that represent mandatory objects only few of bits in m would have been assigned the numeral 1 that belong to the scene under analysis as they are few m objects not representing the belongingness of the scene. The other bits, representing mandatory objects shall be coded with the numeral 0. In the case of d bits representing the desirable objects, the proposed encoding scheme assigns numeral 1 for each of the desirable objects belonging to the scene under experimentation from the most significant digits of (n-m) bits. The remaining bits for desirable objects are coded with numeral 0. Thus, the proposed binary scene representation scheme assigns either 1 or 0 for presence of both mandatory and desirable object and absence of mandatory and desirable objects, respectively. These 1s and 0s of n bit binary scene representation are then converted into a decimal number based on 8421 code, considering the presence and absence of mandatory and desirable objects of the scene under experimentation. The converted decimal number is called *scene-number* in this proposed ISRS. This representation of scene-number is computed for all the scenes Si, i=1, 2,…,s. The proposed scene

recognition system arranges these scene-numbers, computes the highest score scene-number and corresponding scene shall be concluded as scene recognition of the given key frame. It can be observed that the proposed BSR is independent of the number of mandatory and desirable objects present in the key frame under analysis. In the case of same scene-number for more than one scene, the proposed system considers either of the scenes as recognized. Having described the details of all the stages of proposed ISRS, the proposed work aims to compute the measure of its performance. The measures of performance, used in this work are presented in the section.

5. Measures of Performance

In order to measure the performance of the proposed indoor scene recognition system, standard measures, viz, accuracy, recall, precision, F1_score and error rate are used as reported in the literature by similar recognition systems. They are reviewed in this section.

5.1. Accuracy

Accuracy is defined as the proportion of the total number of true positive and true negative classes divided by the total number of positive and negative classes. That is,

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (5.1)$$

where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

5.2. Recall

Recall is defined as the proportion of the total number of correctly identified positive classes divided by the total number of positive classes. In other words, out of all the positive classes, how much of the indoor scenes predicted correctly. Recall should be always high. Mathematically recall is defined as,

$$Recall = \frac{TP}{TP + FN} \qquad (5.2)$$

where TP is True positive, FN is False Negative.

5.3. Precision

Precision is the ratio between the total number of correctly classified positive classes and the total number of predicted positive classes. Otherwise, out of all the predictive positive classes, how much we predicted correctly. Precision should be high. It can be mathematically defined as,

$$Precision = \frac{TP}{TP + FP} \qquad (5.3)$$

where TP is True Positive, FP is False Positive.

5.4. F1_ score

It is difficult to compare two models with different Precision and Recall. So to make them comparable, F1-Score is used. It is the Harmonic Mean of Precision and Recall. In general, Harmonic Mean punishes the extreme values more. F1-score should be high and it is defined as,

$$F1\_score = \frac{2 * Recall * Precision}{Recall + Precision} \qquad (5.4)$$

5.5. Error Rate (ER)

Error rate is calculated as the number of all incorrect predictions divided by the total number of the dataset. The best error rate is 0.0, whereas the worst is 1.0.

$$Error\ Rate = \frac{FP + FN}{P + N} \qquad (5.5)$$

where FN is False Negative, P is Positive and N is Negative.

The performance of the proposed ISRS has been evaluated with the standard datasets and the same is presented in the next section.

6. Experiments and Results

The performance of the proposed indoor scene recognition system has been experimented with universal standard scene datasets viz., indoor scene 67 (Quattoni et al. 2009), scene15 (Lazebnik et al. 2006) and Google images. Three sample key frames taken from indoor scene 67 and one sample from Google are presented in Fig.1. The key frames are color images of varying size with pixels values in the range 0 to 255 in each of the color bands R, G and B. These key frames, taken from indoor scene 67 and Google, are of varying sizes. The input key frames are then

subjected to the detection and recognition of objects with YOLO v3 as described in the section 4.2.1. During the detection and recognition stage, the proposed system converts the variable size key frames into a standard fixed size of (416 x 416) for the purpose of uniformity.

Fig.2 Sample key frames considered in the experimentation. (a), (b) and (c) from indoor scene 67 and (d) from Google.

The detected and recognized objects are marked with bounded boxes. The results of detected and recognized objects with YOLO v3 for the key frames shown in the Fig.2 are presented in the Fig.3, correspondingly.

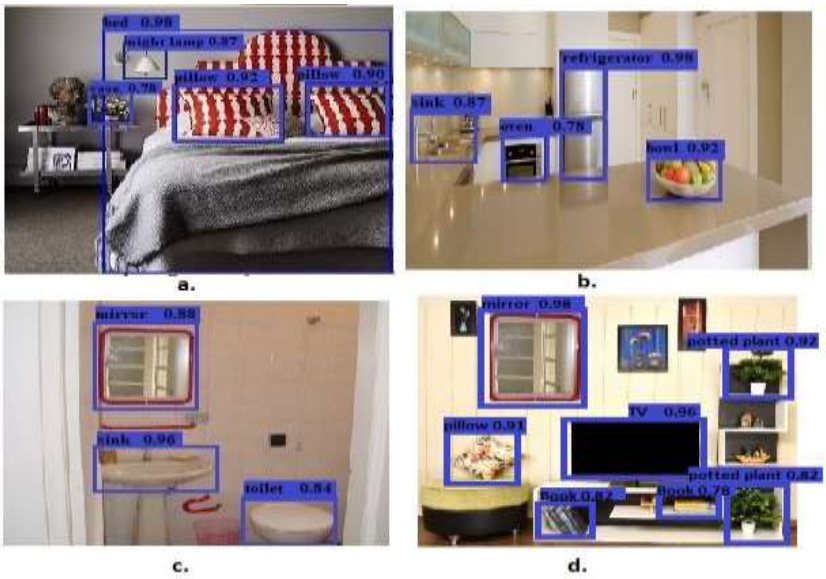

Fig.3. Results of detection and recognition of objects with yolo v3, corresponding to the key frames shown in Fig.2.

It is evident from Fig.3, that the YOLO v3 could correctly detect and recognize indoor objects for the given key frames. The proposed system could detect and recognize the objects (bed, pillow, night lamp, vase) in the case of key frame shown in Fig. 2(a). Similarly, for the key frame shown in Fig. 2(b) the proposed system identifies the following four objects: (sink, oven, refrigerator, bowl). For the key frame shown in Fig. 2(c), the detected and recognized stage of the proposed system could identify the following three objects: (sink, toilet, mirror). For the complicated key frame shown in Fig.2(d) the proposed system recognizes the following five objects: (TV, book pillow, mirror, potted plant). These results of detected and recognized objects corresponding to the key frames shown in Fig.3 are presented in a neat form in Table.5. The set of objects detected from the given key frames are then subjected to the second stage of proposed indoor scene recognition system, viz, identification of mandatory and desirable objects.

Table.5. Results of detected and recognized objects for sample input key frames.

Key frame	Resulting objects
Fig.2(a)	bed, pillow, night lamp, vase
Fig.2(b)	sink, oven, refrigerator, bowl
Fig.2(c)	sink, toilet, mirror
Fig.2(d)	Mirror, pillow, TV, book, potted plant

The identification of mandatory and desirable objects stage of ISRS divides the set of all DRobj(i), obtained in the previous stage, into mandatory and desirable objects, for the input key frame under analysis. This proposed stage defines an object DRobj(i) as mandatory object, denoted as DRMobj(i), if the DRobj(i) is highly essential for the purpose of recognizing the given key frame as any of the scene as described in Section 4.2.2. If any DRobj(i) belongs to the set of mandatory objects given in Table 1, then those DRobj(i) are considered as DRMobj(j). Likewise, if any of DRobj(i) belongs to the set of desirable objects given in Table 2, then those DRobj(i)s are considered as DRDobj(k). The objects which are neither mandatory nor desirable is denoted in this proposed work as DRUobj(*l*). For the input key frame shown in Fig.2(a) the proposed stage identifies {bed, pillow} as mandatory objects and {night lamp} as desirable object. The fourth object {vase} is identified as neither mandatory nor desirable object. In the case of key frame shown in Fig. 2(b) containing {sink, oven, refrigerator, bowl}, the proposed system could identify {sink, oven} as mandatory objects and {refrigerator} as desirable object, besides {bowl} as DRUobj. The proposed stage of ISRS identifies {sink, toilet} as mandatory objects and {mirror} as desirable object for the input key frame shown in Fig.2(c). It can be observed that the proposed stage of ISRS could identify all the three objects in this case, as mandatory and desirable and hence there is no insignificant object. For the complicated input key frame shown in Fig.2(d), the proposed system identifies {pillow, TV, book} as mandatory objects, {mirror} as desirable object and {potted plant} as insignificant object. These results of identification of mandatory and desirable objects of the proposed system corresponding to the key frames shown in Fig.2 are presented in Table.6.

The results obtained in the identification of mandatory and desirable objects are then fed into the scene-object identification stage, for the purpose of probable particularization of each of the objects to a scene as described in Section 4.2.3. For the objects resulted corresponding to the key frames shown in Fig.2(a) the proposed scene objects identification stage could identify the object (bed) as belonging to the scene S3 (bedroom). Similarly, the mandatory object (pillow) of the same key frame is again identified as belonging to same S3 (bedroom). In the case of desirable object (night lamp) the proposed system results in identification of the scene S3 (bedroom). It can be observed that all the two mandatory objects and one desirable object of the key frame shown in Fig.2(a) are pointing to the probable scene S3(bedroom). The results of scene-object identification stage of proposed ISRS, corresponding to the key frames shown in

Fig.2(b), 2(c) and 2(d), are obtained and presented along with the results of key frames shown in Fig.2(a) in Table 7(a) and Table 7(b), respectively for mandatory and desirable objects.

Table 6. Results of Identification of mandatory, desirable and unnecessary objects for different input key frames, shown in Fig.1

Figure No.	DRMobj(i)	DRDobj(j)	DRUobj(k)
Fig.2(a)	bed, pillow	night lamp, vase	---
Fig.2(b)	sink, oven	Refrigerator, bowl	---
Fig.2(c)	sink, toilet	mirror	---
Fig.2(d)	pillow, TV, book	mirror	potted plant

Table 7(a). Results of mandatory Scene-object Identification.

KF	DRMobj(i)	Scene
Fig.2(a)	bed	S3
	pillow	S3
Fig.2(b)	sink	S1, S2
	oven	S1
Fig.2(c)	sink	S1, S2
	toilet	S2
Fig.2(d)	pillow	S3
	TV	S4
	book	S7

Table 7(b). Results of desirable Scene-object Identification.

KF	DRMobj(i)	Scene
Fig.2(a)	night lamp	S3, S4
	vase	S3
Fig.2(b)	refrigerator	S1, S8
	bowl	S1
Fig.2(c)	mirror	S2, S3
Fig.2(d)	mirror	S2, S3

The results of scene-objects identification stage, viz, the list of scenes with their corresponding mandatory and desirable objects are then input to the final stage of the proposed ISRS, for the purpose of finalizing the scene recognition. For each of the scene, the binary representation scheme, proposed in Section 4.2.4 is then carried out. For the input key frame shown in Fig. 2(a), the proposed system codes the binary scene representation as [1 1 1 0] and obtains the scene-number 14 for the possible scene S3 (bedroom). It can be observed that there are four objects in this key frame, out of which two are mandatory, two desirable and empty

unnecessary. Hence the binary scene representation contains a four bit code [1 1 1 0] representing most two significant digits for the presence of mandatory objects, the third digit (left to right) representing the desirable objects and code 0 in least significant digit of proposed binary scene representation, representing unnecessary objects. In a similar way, the proposed system codes the binary scene representation for the key frames shown in Fig.2(b), 2(c) and 2(d) and the results obtained in terms of binary scene representation, and hence the scene-numbers are presented in Table.8. It can be observed from the results presented in Table.8, that the key frames shown in Fig. 2(a) consist of three objects of same scene S3(bedroom) and hence the proposed scene recognition system clearly points to the scene S3(bedroom). But in the case of key frame shown in Fig.2(b) containing two mandatory objects, one desirable object and one unnecessary object, the proposed system identifies three scenes S1(kitchen), S2(bath room) and S8(dining hall). In this case, even though there are different number of scenes being pointed, the proposed system could elegantly identify proper scene recognition, with the help of proposed binary scene representation technique and scene-number. For the scenes S1 (kitchen), S2 (bath room) and S8 (dining hall) that are pointed as the possible scenes, the proposed system codes the binary scene representation [1 1 1 1], [1 1 1 0] and [0 0 1 1], respectively with corresponding scene-number as 15, 14 and 3. In such case of identification of multiple scene scenarios, the proposed system makes use of the highest score of scene-number to decide the result of scene recognition. Since the maximum scene-number score 15 corresponding to the scene S1 (kitchen) is resulted, as scene recognition for the key frame shown in Fig.2(b).

In the case of key frame shown in Fig.2(c) there is no much complication and hence the proposed system could easily recognize the scene, as given in Table.8. For the input complicated key frame shown in Fig.2(d), the proposed system points the possible scene S3 (bed room), S2 (bath room), S4 (living hall) and S7 (library). The corresponding binary scene representation is evaluated for these four scenes and converted to decimal number with 8421 coding, as described in Section 4.2.4 and the following scene-numbers are obtained respectively: 18, 2, 16, and 16. The proposed system is capable of elegantly identifying the scene bedroom in spite of complicated objects.

Table 8. Results of mandatory, desirable objects with BSR to recognize the indoor for different input key frames shown in Fig.2.

Key frame	DRMobj(i)	DRDobj(j)	Belonging Scene	BSR value	Scene-number	Results of scene

						Recognition
Fig.2(a)	bed, pillow	night lamp	S3(Bed room)	[1 1 1 0]	14	Bed room
	---	night lamp, vase	S4(Living hall)	[0 0 1 1]	3	
Fig.2(b)	sink, oven	Refrigerator, bowl	S1(Kitchen)	[1 1 1 1]	15	Kitchen
	---	refrigerator, bowl	S8(Dining hall)	[0 0 1 1]	3	
Fig.2(c)	sink, toilet	mirror	S2(Bath room)	[1 1 1 0]	14	Bath room
	sink	---	S1(Kitchen)	[1 0 0 0]	8	
	---	mirror	S3(Bedroom)	[0 0 1 0]	2	
Fig.2(d)	pillow	mirror	S3(Bedroom)	[1 0 0 1]	9	Living hall (or) Bedroom
	TV	mirror	S4(Living hall)	[1 0 0 1]	9	
	book	---	S7(Library)	[1 0 0 0]	8	

The performance of the proposed indoor scene recognition system is measured with standard measures as introduced in Section 5. For this purpose of computing the performance measure, the experiments are carried out with 200 samples from each of the MIT indoor 67 and scene15 datasets. MIT indoor 67 dataset is a large dataset containing of 15,620 images which are organized into 67 categories of indoor scene. These images are collected from different sources such as LabelMe dataset, online image search engines, and photos. The Scene 15 dataset contains 15 scene categories combined with indoor and outdoor scenes. Each category is composed of 200 to 400 images with the size (300 x 250). In this work, the samples are considered form indoor categories (i.e., kitchen, living room, office, and bedroom) that are specific to the proposed experimental part.

The performance of the ISRS is evaluated and presented as a confusion matrix for 25 indoor scenes and the results are presented in Table 9. It can be observed from this confusion matrix that the proposed system could identify good number of objects present in complicated input key frames.

Table 9. Confusion matrix obtained with proposed ISRS for 25 number of scenes for MIT indoor 67.

Scene	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...	25
1	0.98	0.01	0	0	0	0	0	0.01	0	0	0	0	0	0	0	...	0
2	0.01	0.97	0	0	0	0	0	0	0	0	0	0.01	0	0	0	...	0
3	0.02	0.01	0.97	0	0	0	0	0	0	0	0	0	0	0	0	...	0
4	0.01	0	0.01	0.97	0	0	0	0	0	0	0	0	0	0	0	...	0

5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	…	0
6	0	0	0	0.01	0.01	0.97	0	0	0	0	0	0	0.01	0	0	…	0
7	0	0	0.01	0	0.01	0.01	0.98	0	0	0	0	0	0	0	0	…	0
8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	…	0
9	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	…	0
10	0	0	0	0	0	0	0	0	0	0.99	0.01	0	0	0	0	…	0
11	0	0	0	0	0	0	0	0	0	0.01	0.99	0	0	0	0	…	0
12	0	0.01	0	0	0	0	0	0	0	0	0	0.99	0	0	0	…	0
13	0	0	0	0	0.01	0	0.01	0	0	0	0	0	0.98	0	0	…	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0	…	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	…	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	…	1

25 number of scene, viz, 1. Kitchen2. Bath Room 3.Bed Room 4.Living Hall 5.Class Room 6.Computer Lab 7.Library 8.Dining Hall 9.Gym 10.Two Wheeler Stand 11.Four Wheeler Stand 12.Waiting Hall 13.Office 14.Game Room15.Bar 16.Bowling Court 17.Hair Salon 18.Conference Hall 19.Medical Shop 20.Cinema Theatre 21.Bakery 22.Swimming Pool23.Laundromat 24.Meeting Room 25.Fruit Shop

The proposed system gives an accuracy of 98.27%, a recall value of 98.31%, a precision value of 98.30% and 98.28 as F1_score value besides error rate of 1.82%. The performance of the system for Scene15 dataset is also measured and results are presented. The proposed system achieves an accuracy of 92.30%, a recall value of 92.99%, a precision value of 93% and 92.98% as F1_score value besides error rate of 6.77%. These results of performance measure by the proposed system for the two standard datasets MIT indoor 67 and scene 15 are presented in Table10.

Table 10. Results of performance of proposed indoor scene recognition system and comparison with existing schemes.

Dataset		**Accuracy (%)**	**Recall (%)**	**Precision (%)**	**F1_score (%)**	**ER (%)**
MIT Indoor 67	**Proposed ISRS**	**98.27**	**98.31**	**98.30**	**98.28**	**1.82**
	Cheng et al. 2017	83.98	84.22	83.99	84.12	14.02

	Zrira et al. 2018	76.45	76.67	76.50	76.48	23.21
Scene15	**Proposed ISRS**	**92.30**	**92.99**	**93**	**92.98**	**6.77**
	Cheng et al. 2017	80.24	80.14	80.26	80.30	17.25
	Zrira et al. 2018	69.43	70	68.82	69.10	31.32

The efficiency of the proposed indoor scene recognition system is also measured by comparing with it existing systems namely(Cheng et al. 2017 and Zrira et al. 2018) on the same datasets MIT indoor 67 (Quattoni et al. 2009) and scene15 (Lazebnik et al. 2006).

The method by (Cheng et al. 2017) could produce an accuracy rate of 83.98%, a recall value of 84.22%, a precision value of 83.99% and 84.12% as F1_score value besides error rate of 14.02% for the dataset MIT indoor 67. The same method (Cheng et al. 2017) is also evaluated for Scene15 dataset with four indoor categories (kitchen, bedroom, living hall and office), and could produce an accuracy rate of 83.98%, a recall value of 84.22%, a precision value of 83.99% and 84.12% as F1_score value besides error rate of 14.02%.

The method by (Zrira et al. 2018) could produce an accuracy rate of 76.45%, a recall value of 76.67%, a precision value of 76.50% and 76.48% as F1_score value besides error rate of 23.21% for the dataset MIT indoor 67. The same method is also evaluated with the dataset Scene15 (Zrira et al. 2018) and it could produce an accuracy rate of 69.43%, a recall value of 70%, a precision value of 68.82% and 69.10% as F1_score value besides error rate of 31.32%. These results of existing systems are incorporated in the Table.10 for easy comparison. It is evident from Table.10, that the proposed indoor scene recognition system could give a better performance measure than the existing systems.

7. Conclusion

In this paper, a novel scheme is proposed for automatic indoor scene recognition system. Initially, the proposed system utilizes the YOLO v3 and then the results are divided into mandatory and desirable objects based on simple look-up table, created exclusively for this purpose. These mandatory and desirable objects then point to a possible scene in the newly introduced scene-object identification stage. Even though probable scenes are obtained, for the purpose of proposed scene recognition, a binary coding technique is proposed and converted into a decimal number called *scene-number*. The performance of the proposed indoor scene recognition system has been experimented with bench mark scene datasets and a maximum accuracy rate of 98.27% is achieved. The result of performance measure obtained with the proposed system is also compared with that of the existing techniques. It is evident from the experiments presented in Section.6 that the proposed system could achieve better results when compared with existing systems in terms of accuracy, error rate, precision, recall and F1_score.

Funding: No funding
Conflicts of interest/Competing interests: No conflict and Competing Interest
Availability of data and material: Available in paper
Code availability: Available in paper
Ethics approval: Not applicable
Consent to participate: Not applicable
Consent for publication: Not applicable

References

[1]Viola, P., & Jones, M.: Rapid object detection using a boosted cascade of simple features, in Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. (CVPR 2001) Vol. 1, pp. I-I(2001).

[2] Begum, S. A., & Askarunisa, A. (2020). Performance Analysis of Machine Learning Classification Algorithms in Static Object Detection for Video Surveillance Applications. Wireless Personal Communications, Vol.115(2), pp.1291-1307.

[3]Dalal, Navneet, and Bill Triggs.: 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(2005).

[4]Felzenszwalb, Pedro F., Ross B. Girshick, David McAllester, and Deva Ramanan.: Object detection with discriminatively trained part-based models, in IEEE transactions on pattern analysis and machine intelligence 32, Vol. no. 9, pp. 1627-1645(2009).

[5]Wang, X., Yang, M., Zhu, S., & Lin, Y.: Regionlets for generic object detection, in Proceedings of the IEEE international conference on computer vision, pp. 17-24(2013).

[6] Girshick, R.: Fast r-cn, in Proceedings of the IEEE international conference on computer vision, pp. 1440-1448(2015).

[7]Ren, S., He, K., Girshick, R., & Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks, in IEEE transactions on pattern analysis and machine intelligence, Vol. 39(6), pp1137-1149(2016).

[8]Eghbali, H., & Hajihosseini, N.: Deep Convolutional Neural Network (CNN) for Large-Scale Images Classification. SSRN Electronic Journal. doi: 10.2139/ssrn.3476258(2019).

[9]Erhan, Dumitru, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov.: Scalable object detection using deep neural networks, in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2147-2154(2014).

[10]Ciregan, D., Meier, U., and Schmidhuber, J.: Multi-column deep neural networks for image classification, in 2012 IEEE conference on computer vision and pattern recognition, pp. 3642-3649(2012).

[11]Ouyang, W., Wang, X., Zeng, X., Qiu, S., Luo, P., Tian, Y., and Tang, X.: Deepid-net: Deformable deep convolutional neural networks for object detection, in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2403-2412(2015).

[12]Kang, K., Li, H., Yan, J., Zeng, X., Yang, B., Xiao, T., Ouyang, W.: T-CNN: Tubelets With Convolutional Neural Networks for Object Detection From Videos, in IEEE Transactions on Circuits and Systems for Video Technology, Vol. 28(10), pp. 2896–2907(2018).

[13]Quattoni, A., Torralba, A.: Recognizing indoor scenes, in 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 413-420(2009). IEEE.

[14]Tong, Z., Shi, D., Yan, B., & Wei, J.: A review of indoor-outdoor scene classification, in 2017 2nd International Conference on Control, Automation and Artificial Intelligence (CAAI 2017). Atlantis Press.

[15]Szummer, M., Picard, R.W.: Indoor-Outdoor Image Classification. in Proceedings 1998 IEEE International Workshop on Content-Based Access of Image and Video Database, IEEE Comput. Soc, pp. 42–51(1998).

[16]Madokoro, H,. Utsumi, U,. Sato, K.: Scene classification using unsupervised neural networks for mobile robot vision, in SICE Annual Conference (SICE), 2012 Proceedings of, pp.1568–1573(2012).

[17]Lazebnik, S,. Schmid, C., and Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, pp. 2169-2178(2006).

[18]Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints in International Journal of Computer Vision , Vol. 60, pp. 91–110(2004).

[19]Moosmann, F., Triggs, B., & Jurie, F.: Fast discriminative visual codebooks using randomized clustering forests, in Advances in neural information processing systems, pp. 985-992(2007).

[20]Kövesi, B., Boucher, J. M., & Saoudi, S.: Stochastic K-means algorithm for vector quantization, in Pattern Recognition Letters, Vol. 22(6-7), pp. 603-610(2001).

[21]Fu, K,, Li, J., Jin, J., Zhang, C.: Image-text surgery: Efficient concept learning in image captioning by generating pseudo pairs, IEEE Transactions on Neural Networks and Learning Systems Vol. 29 (12), pp. 5910–5921(2018).

[22]You, Q,. Jin, H., Wang, Z,. Fang, C., Luo, J.: Image captioning with semantic attention, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4651–4659(2016).

[23]Bosch, A,. Andrew, Z,. Xavier, M.: Image classification using random forests and ferns. 2007 IEEE 11th international conference on computer vision, pp. 1-8(2007), IEEE.

[24]Brett, M., Anton, J.L., Valabregue, R. and Poline, J.B.: Region of interest analysis using the MarsBar toolbox for SPM 99. Neuroimage, Vol. 16(2), pp.S497(2002).

[25] Van Gemert, J.C., Veenman, C.J., Smeulders, A.W. and Geusebroek, J.M.: Visual word ambiguity, in IEEE transactions on pattern analysis and machine intelligence, Vol. 32(7), pp.1271-1283(2009).

[26]Juneja, M., Vedaldi, A., Jawahar, C.V. and Zisserman, A.: Blocks that shout: Distinctive parts for scene classification, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 923-930(2013).

[27]Liu, S., Xu, D., & Feng, S.: Region contextual visual words for scene categorization. in Expert Systems with Applications, Vol. 38(9), pp. 11591-11597(2011).

[28]Yang, J., Jiang, Y. G., Hauptmann, A. G., Ngo, C. W.: Evaluating bag-of-visual-words representations in scene classification. In Proceedings of the international workshop on Workshop on multimedia information retrieval, pp. 197-206(2007).

[29]Lin, D., Lu, C., Liao, R., and Jia, J.: Learning important spatial pooling regions for scene classification, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3726-3733(2014).

[30]Espinace, P., Kollar, T., Roy, N., and Soto, A.: Indoor scene recognition by a mobile robot through adaptive object detection. Robotics and Autonomous Systems, Vol. 61(9), pp. 932-947(2013).

[31]Redmon, J., Farhadi, A.: YOLOv3: An incremental improvement, in: IEEE conference on Computer Vision and Pattern Recognition, arXiv:1804.0276,(2018).

[32]Cheng, X,. Lu, J,. Feng, J,. Yuan, B,. Zhou, J.: Scene recognition with objectness, Pattern Recognit. Vol. 74 (2017), pp. 474–487(2017).

[33]Zrira, N., Khan, H. A., & Bouyakhf, E. H.: Discriminative deep belief network for indoor environment classification using global visual features. Cognitive Computation, Vol. 10(3), pp. 437-453(2018).

[34] Xu, X., Chen, Y., Zhang, J., Chen, Y., Anandhan, P., & Manickam, A. (2021). A novel approach for scene classification from remote sensing images using deep learning methods. European Journal of Remote Sensing, 54(sup2), 383-395.

[35] Liu, X., Wang, W., Guo, Z., Wang, C., & Tu, C. (2019). Research on adaptive SVR indoor location based on GA optimization. Wireless Personal Communications, Vol.109(2), pp.1095-1120.

Figures

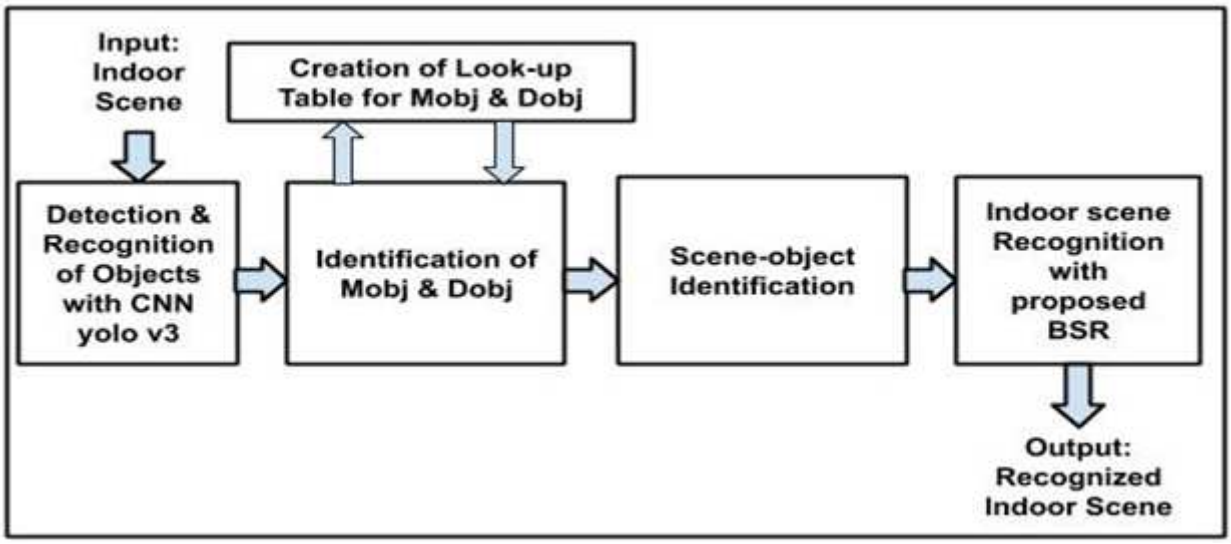

Figure 1

Proposed Architecture of Indoor Scene Recognition System.

Figure 2

Sample key frames considered in the experimentation. (a), (b) and (c) from indoor scene 67 and (d) from Google.

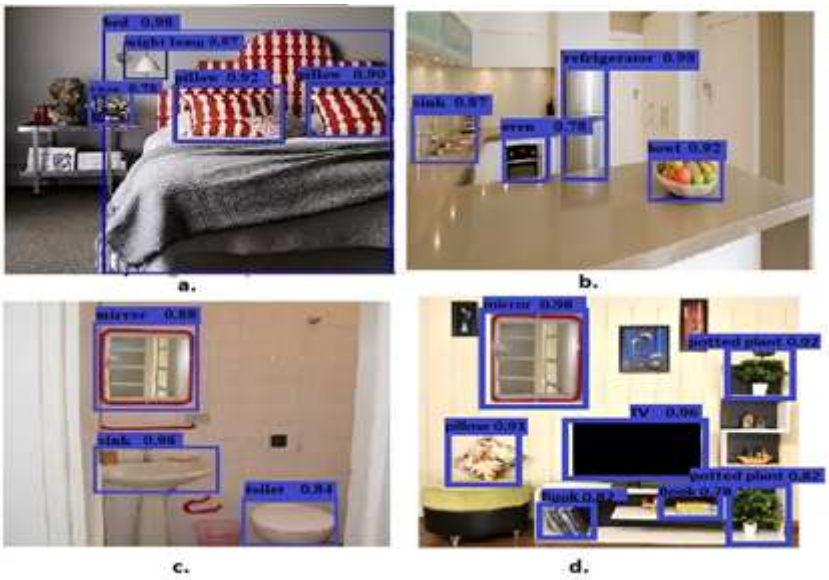

Figure 3

Results of detection and recognition of objects with yolo v3, corresponding to the key frames shown in Fig.2.