Sigma J Eng Nat Sci, Vol. 43, No. 4, pp. 1355–1365, August, 2025

Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2024.00157

Research Article

Indoor-outdoor scene recognition: A multi-feature framework using CNN for complex environment

Pandit NAGRALE[1,*], Sarika KHANDELWAL[1]

[1]*Department of Computer Science & Engineering, G.H. Raisoni University, Maharashtra, India*

ARTICLE INFO

Article history
Received: 02 May 2024
Revised: 24 June 2024
Accepted: 05 August 2024

Keywords:
Classification Accuracy; CNN; Cross-Domain Dataset; Handcrafted Features; Image-Based Features; Object-Based Features; Scene Recognition; Semantic Context

ABSTRACT

Scene Recognition is deeply governed by the semantic context in the scene images. The challenge introduced by diversity in intra-class spatial layouts, and similar object's existence in inter-classes imposes great difficulties in adapting image characteristics. Today existing approaches either incorporate either object-based features, image-based features, and handcrafted features or a combination of two feature extraction strategies. Therefore, existing models are unable to represent the spatial context, and overlook the distinctiveness of coexisting objects across different scenes. These issues have degraded the performance of scene recognition systems even over a single dataset. The work presented in this article uses distinct features obtained using the scene objects (object-based), complete scene images (spatial layout-based), and eight handcrafted features. A fully connected convolutional neural network (CNN) is trained on cross-domain dataset images obtained from three distinct datasets using the combined features. Experimental evaluation of the framework over the test samples showed that the transfer learned-based CNN model was able to obtain a mean classification accuracy of 95.84% (indoor and outdoor scenes) outperforming other better approaches. The sample groups for training, validation, and tests were obtained randomly from the self-generated dataset.

Cite this article as: Nagrale P, Khandelwal S. Indoor-outdoor scene recognition: A multi-feature framework using CNN for complex environment. Sigma J Eng Nat Sci 2025;43(4):1355–1365.

INTRODUCTION

Humans are gifted with the best sense for quick and easy categorization of visual sceneries [1-2]. Within a fraction of a second, a human eye in coordination with his brain can classify indoor and outdoor (I&O) scenes. Even humans are immediately responsive to basic-level categorization more specifically, they can distinguish a hall, a bedroom, a bathroom, shops, entrance, etc. [3-4]. However, the relevant details in the scene are not yet understood despite the strong categorization of humans. Literature shows that most of the theories belonging to the capability of humans are either object-based or scene-based [5]. Object-centered or object-based scene categorization primarily recognizes a single prominent object significant to the scene category [6-7]. Scene-centered or scene-based categorization on the other hand focuses on global or overall scene properties (such as information about spatial placements exploiting

*Corresponding author.
*E-mail address: ptnagrale@gmail.com

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

object identities [8-9]). In reality, scene categorization is understood as either type or both and several hybrid approaches [10-12].

Outdoor scenes are responded to by selective neurons [13-14] of the human brain however, their sufficiency in categorization is ambiguous. Computationally useful, outdoor scenes depends on global properties that only account scenes with changing spatial layouts. Nevertheless, indoor scenes have identical spatial layouts and mostly differ in their class-defining objects. Scene-recognizing frameworks based on global properties within the image are not efficient since their performance drops significantly when they are subjected to indoor scenes. Secondly, when human subjects are stimulated using the global scene properties for a free response [15], they are unable to recognize and thus show poor performance. Estimates rated by humans based on spatial layouts are even poor performers [4]. Therefore, scene global properties independently are insufficient enough to describe the scenes effortlessly.

Human scene distinguishing capability is driven by object information and suggested as the most plausible candidate [6-7]. Humans are familiar with various kinds of objects in everyday life due to their frequent ceaseless interactions. They are capable of identifying the scene category merely by watching a single object. Thus, we can make assumptions based on such key objects when they are perceived at first glance. Different theories have supported that object details form the central part of scene categorization. However, in the absence of objects or the presence of semantically inconsistent objects, humans are worse at categorizing scenes [10,11,16,17]. Neural findings show that brain regions identifying the scenes are activated by a single specific object [18] and modulated by the object's properties [19]. Noticeable improvement in scene categorization performance can be achieved by employing object detectors in the frameworks [20,21]. Friedman [7] and Biederman [6] validated that certain objects are sufficient to diagnose the type of scenes and act as a tool to instantiate the reference, but they are not necessary.

The authors in [22] addressed inter-class similarity and variability in the case of indoor environments. They proposed a dual-stream model that can efficiently extract overall contextual details and local information for improved scene recognition. The former details capture high-quality information and correlations across the scenes, while fine-grained were part of the latter local features. A convolutional network was employed to uplift the local features. Their model effectively classified scenes that integrate similarly matching global contexts with distinct localized objects. They used the SRIN dataset and augmented the 1600x1200 resolution images to prevent overfitting to distinguish five indoor classes that included living rooms, bathrooms, bedrooms, dining rooms, and kitchens. They obtained 100% accuracy in the case of three classes, while 91.7% and 93.8% for the kitchens and living rooms classes, respectively.

Ningbo Guo et al. [23] worked on remote sensing images and classified seven categories of outdoor images. Their dual XE-Net model with multi-level features was designed to discriminate land use and land cover (LULC) images. The Xception and Efficient-V2 networks were employed to extract high-quality and low-level features, respectively using a transfer learning approach. The low, medium, and high-level features were fused sequentially and performed well on seven different datasets.

The indoor-outdoor scene images show significant variations concerning color, textures, spectrum information, objects, scales, etc. Due to complex spatial arrangements, extracting semantic features requires effective computer vision (CV) methods. They also exhibit low inter-class variance, which requires good calibration CV techniques. On the other hand, high intra-class variation requires CV approaches that can extract similar pattern features regardless of their variance. The presence of varying illuminations in the indoor-outdoor scenes necessitates robust feature-learning techniques to mitigate the effect. The objective of the proposed multi-feature framework using CNN is to extract local, global, and object-level features from the scene images to improve classification accuracy.

The article contributes to the following:

1. The proposed scene classification framework obtains quality features using object-level, image-level, and statistical handcrafted features. Objects in the scene being the crucial attributes, they are recognized using the YOLOv5 network and represented by fixed dimension quality features extracted using VGG19.
2. Diverse conventional features using various quality descriptors are added to enhance the scene representation quality which includes fine and local attributes from the scenes.
3. Color-based blind features using VGG19 are added to ensure chroma variations relative to foreground and objects.
4. Experimental investigation on scene generated dataset reaveled that a finely tuned LSTM network showed better performance on 15% test images when it was trained and validated using 75%:10% sample features.

Related Work

The study of several types of research in scene recognition prominently followed three sequential steps in their design. It includes feature learning from distinct scales and positions of the scene objects, obtaining feature descriptions through pooling, and training a classifier. Conventional strategies focus on low-level attributes which include color, edge, and texture patterns [22], and use handcrafted techniques such as SIFT [23], GIST [24], HOG [25], and SURF [26], and are enhanced by adding a Bag of words [27]. The middle phase used pooling techniques to aggregate the descriptors using Fisher vectors [28] and Locally Aggregated Descriptors [29]. Whereas the last stage of classification involved popular and widely used

Support vector machines, custom neural networks, and the K-nearest neighbor. The tedious job of extracting several handcrafted features and then eliminating the redundant or non-significant ones and selecting the optimum features is nowadays exploited by blind features. Deep learning [30] has proven to be a better solution in many computer vision applications. DL is capable of learning raw images and extracting the features to represent the image through high-level details.

As a part of a social cause, the work carried out by authors in [31] developed a framework that could help humans with visual impairment to navigate efficiently in their residence and live a normal life. They used the NYU dataset [32] to recognize the indoor scenes for the evaluation using a transfer learning approach through the DenseNet201 model, accompanied by a deep Liquid State Machine model for extracting features and classification, respectively. Subjects affected by visual deformities were able to use the system due to the performance improvement in recognition and understanding of indoor scenes through fuzzy color skating techniques, segmentation, and an adaptive World Cup optimization algorithm. They obtained a classification accuracy (20 classes) of 96% over some specific indoor scene images from the dataset. The work was limited to some specific indoor scenes only and thus lacked generalization ability.

Authors in [33] worked to enhance the capability of indoor mobile robots to classify indoor scenes, limited to rooms and corridors, to different rooms. Their model was based on a multi-scale CNN combined with LSTM networks and whale optimization. The ablation experiment was carried out on data collected through the 2D LiDAR. The first two combined networks were utilized for scene classification, and the optimization algorithm governed the fine-tuning of network parameters and performance improvement of their model. A classification accuracy of approximately 99% was obtained on the real data, while 94% on the publicly available FR079 dataset. The FR079 dataset was limited to 11 rooms and a corridor; therefore, the authors added 3-CNN layers to their model. Their work offered low storage and clean data, and suffered misclassification for similar objects, especially in the office environment.

Indoor scene recognition to assist MAV navigation was proposed in [34]. They compared the performance of two commonly used classifiers on three classes, including the corridor, room, and staircase. Handcrafted features, which included the GIST, enhanced GIST, and HODMG (Histogram of Directional Morphological Gradient) were extracted from the dataset images on one hand, and vanishing point detection based on Canny edges and lines on the other hand. They found that SVM was superior in recognizing the three entities with an accuracy of 99.33% over K-NN.

An object feature-based scene classification model using computer vision and natural language processing was introduced in [35]. The authors at the first stage detected the objects from the scene image and then extracted object features for classification. For this, their model was built using the YOLOv5 network to recognize objects in the scene and the TF-IDF approach for classification. They trained the YOLO network using Open Images V6 [36], which reasonably included almost covered indoor objects for 90 and 155 classes. They chose 8 indoor rooms as classes to classify the images from the Places365 dataset. The proposed scene recognition framework was simple and cost-effective, while it suffered from a lack of semantic relationships among the scene objects and an absence of room composition learning.

Experiments carried out on NYU V2 and SUN RGB-D dataset images obtained efficient and accurate results [37]. The authors presented a feature-based, lightweight model for indoor scene parsing. They used the MobileNetV2 network to extract features and ensured that it formed the backbone to uplift the depth information from the color images. Features were concatenated from different layers and performed feature-level adaptive fusion to classify at the pixel level. Many other state-of-the-art methods found in the literature have efficiently used large datasets about scene classification with Deep Learning due to its popularity and the momentum it has gained in several areas of computer vision. DenseNet [38], SqueezeNet [39], ResNet [40], VGG16 [41], GoogleNet [42], etc. are some of the commonly used networks that have been successfully incorporated for indoor and outdoor scene classification. Large datasets prominently include MIT [43], NYU [44], Places365 [45], Scene [46], SUN [47], etc. Large Nets and large datasets are now part of the research of interest due to the availability of powerful processors and improved deep-learning networks [48,49].

Two well-known datasets (SUN RGB-D & NYU Depth V2) for indoor scenes were part of the research carried out by Ricardo Pereira et al. [50]. They proposed a novel segmentation-based method to extract meaningful segmentation-based semantic features (SSFs). The 2D-spatial distribution obtained using the segmentation was encoded and bypassed the CNN-relied image-level features obtained from the color image. The Image-Level Object-based Feature aggregation Approach exploited the normal CNN outputs. They claimed that their model achieved the highest classification accuracy of 62.3% and 77.8% over the two datasets. However, they failed to encompass correlated features and hand over semantic details regarding within-scene objects with annotations that exist in the scene images.

Authors in [51] worked to mitigate the failure of ImageNet dataset-based deep-learned networks. They introduced a Self-Supervised deep-learning model and trained it from scratch using unlabelled scene images in the first pretext stage. The model learned labeled scene images during the later or downstream stage. Their EfficientNet-B3 CNN model, which incorporated the online and target networks, acted as the backbone to encode scene features using a cross-view contrastive learning approach. Input data was

augmented using geometrical transformation and subjected to the dual network, and the cross–view distance between both networks was optimized for minimum. During the first stage, they used low-resolution images (large batch size) while a smaller batch size (including all types of resolution images) was used at the later stage by eliminating the target network. They obtained about 87% and 88% accuracy over AID and NWPU-RESISC45 datasets, respectively, for 20 classes.

MATERIALS AND METHODS

In the proposed indoor-outdoor scene classification framework, features from detected objects, global attributes, and diverse conventional features are combined to improve detection accuracy. Objects in the scenes are detected using the YOLOv5m network. However, due to the objects' different sizes, the feature vectors' length was compromised. Extensive experiments over several images in the dataset showed that a dimension of 32x32 was able to accommodate the smallest detected object and retain the quality of the largest object in any image. Therefore, the size of the detected object by the YOLOV5m model was reduced to a dimension of 32x32. Blind features were extracted using a VGG19 pre-trained network after removing its top layer. All the objects detected by the YOLOv5m network from an image were provided as input to the VGG19 network for feature extraction. The number of objects recognized and marked by the bounding box depended on the scene input. E.g., the length of the feature vector that would result from a scene image consisting of 5 objects would be 5x512, and for 10 objects, 10x512. Therefore, we added the features corresponding to all objects to make a uniform output of 1x512 dimensions. Eventually, the resultant vector elements reflected a higher magnitude for more objects than a scene consisting of a lower number of objects. The value of elements is a function of the distinctiveness of the objects, whether they are related to indoor or outdoor scenes. The object-based feature extraction mechanism is depicted in Figure 1 by the first branch of the scene-recognizing framework.

Global features were used to ensure the contribution of regions not belonging to the objects in the scene. Such regions carry useful information regarding the category of the scene and they explicitly conveys the overall scene contents. Therefore, information relating to the color depth was extracted using another VGG19 network separately as shown in Figure 1. A total of 512 feature elements were used to represent the complete image. Global-level information from scenes is important since there is a possibility that objects present may induce ambiguity. That is, a bicycle or a tea table may be present in either indoor (living room) or outdoor scenes (home garden). The overall details (other living room details or trees, sky, etc.) contributing to the features would then be useful for the classifier to distinguish the classes correctly.

Traditional approaches as studied in the literature showed that even handcrafted features are effective and can be efficiently used for classifying scenes. Along with the object-level features and the global features, we extracted handcrafted features from grayscale input images. The original input images were reduced to 128x128 to reduce the computational complexity. This was done without losing significant details of the image under consideration. Several quality features were added to the object-based feature and scene-based feature. This was to compensate for the loss due to the dimension reduction of the YOLOV5m detected objects in the object-based feature extraction process. To uplift more fine details, 2310 elements using various descriptors were added to the feature set. Table 1 shown below lists various descriptors used by the scene-recognizing framework.

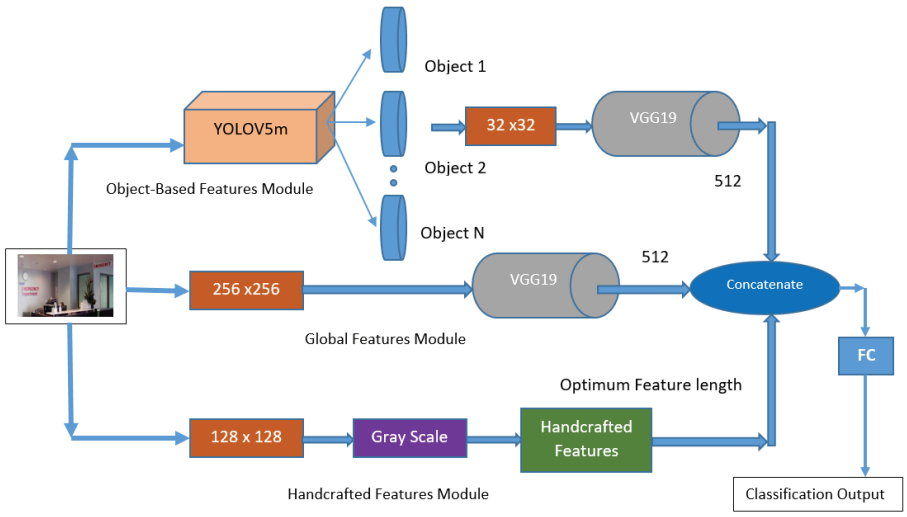

Figure 1. The framework of the proposed scene recognition system.

Table 1. Various descriptors used for Handcrafted features with respective dimension

Descriptors	Parameters	Dimension
Matched Filter (Edge-Based)	3x3 Kernel (Neighbourhood value = -1 and center pixel = 8)	256
Matched Filter (Sobel-Based)	Sigma = [0.5, 1, 1.5, 2] L = [9, 11, 13, 15] Orientation - 12	32
Wavelet	Magnitude and Energy of last two components [bior-3.1, 3.5, and 3.7, Daubechies and symlet 3, and haar]	24
GLCM	Parameters = 6	6
LBP	Radius = 3	256
LBP Fine	3x3 patch Averaging over 5x5 window	676
HOG	64x64 pixels per cell with 9 orientations	36
Original Image	Mean over 4x4 patch	1024
Total Feature Vector Length		**2310**

Features based on wavelets, matched filters, LBP, GLCM, and HOG form the feature set of handcrafted features.

Edge-based features are obtained using a 3x3 kernel whose center element value=8 while the neighbors have value=-1. A 2D filter using the kernel was used over the 128x128 dimension grayscale image. The coefficients of the filter were summed along both axis (axis=0 and axis=1) and concatenated to obtain a 128+128=256 element vector thus detailing the edge information from the input image. Another edge informative filter using the 'Sobel' operator was used with the value of sigma = [0.5, 1, 1.5, 2], length of the filter L = [9, 11, 13, 15] (length in Y-direction), and 12 orientations from [0 to 165 at an offset of 15) to obtain 32 elements in the features. Using two such edge operators, minimum loss due to edge miss was ensured.

The last two components of the first-order wavelet decomposition are used to extract energy and magnitude using three Bior mother wavelets (3.1, 3.5, and 3.7), one Daubechies (db3), one symlet (sym3), and Haar. The absolute and the square values of the vertical coefficients and the diagonal coefficients are summed up separately to measure the magnitude and energy of both components. These two measures provide information concentration in both directions. The effect due to the horizontal component was ignored since the measure was nearer to the vertical details. Using six mother wavelets with two measures resulted in 24 values contributing to the feature set. Image-level information, such as intensity disparities (contrast), overall energy, pixel consistencies (homogeneity), unlikeness, correlation, and uniformity (Angular second moment), is computed using the Gray Level Co-occurrence Matrix. These six image-level attributes were added to the feature set to boost the classifier's performance.

Textural aspects contained in the scene image were acquired using LBP. The textural information was obtained using a 3x3 kernel imposed over the image. The texture carrying values were added along rows (128) and columns (128) and concatenated to represent the textural aspect using 256 values. Another approach for uplifting more fine textures was also incorporated by representing the 3x3 window with a single clockwise pattern (read out pattern) using LBP. The texture pattern in 5x5 windows were then averaged to reduce the feature dimension and represented by a 26x26 array. The array was flattened into a 676 dimension vector to add local textural attributes.

A square window of 64 by 64 was used to extricate HOG features from the image. We used 9 orientation with Max-normalzed details to obtain 256 elements in [0, 1] range. To avoid loss of details from the original grayscale image, mean values corresponding to 4x4 distinct cells from the source image were also made part of the feature set. Thus the mean values added another 36 element to the feature representation set.

Therefore as Table 1 indicates, a feature vector of dimension 2310 using several handcrafted descriptor that was concatenated to other two features to form a vector of dimension 3334. Manual analysis showed that features from 1296 to 1311 in the vector showed no variations and were eliminated. The remaining feature vector with 3318 length was normalized using the Max-Normalization technique.

We attempted to train the network and found the problem of handling the zero entries in the feature set. The zero entries in the feature vector were filled by averaging the known entries in each column. Further, to reduce the burden of the classifier network, four consecutive features along each column (image) were averaged to lower the dimension of the representative vector. The final feature vector presented to the network governing a single real or a fake image was 830. The real and the fake samples were labeled with '0' and '1' for authentic and fake

Table 2. Parameters for the network

Parameter	Value
Activation function for the layers	'selu'
Samples used in a single batch	10
Maximum epoch	100
Learning Rate	0.01
Metric	'accuracy'

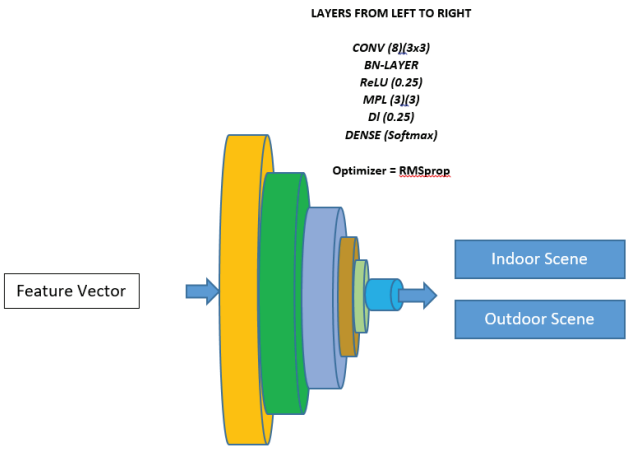

Figure 2. Fully connected (FC) network used for classification of ındoor/outdoor scenes.

images. Thus, an input of 830 elements was presented to a fully connected network shown in Figure 2. The network comprised of a convolutional layer (CONV) followed by a BatchNormalization layer (BN), a LeakyRelu layer (ReLU), a MaxPooling layer (MPL), and a Dropout layer (Dl). The final layer (Dense layer) used to classify two classes was used with the 'softmax' transfer function. The model was compiled using the 'RMSprop' optimizer with the 'SparseCategoricalCrossentropy' function. Table 2 lists the network information used for training.

Out of 5081 images, the first randomly selected 15% (763 combined from both classes) samples were used as test samples over which the performance was evaluated. Subsequently, 10% (432 samples from both classes) of random samples from the remaining were used for validation. The remaining samples (3886) after isolating test and validation were used for network training. The model was trained and tested 20 times by selecting random samples from the feature set for cross-validation. The following Table 3 shows training, testing, and validation samples.

Table 3. Randomized samples used for training, validation, and test

Sample Category	Value
Train	3886
Validation	432
Test	763

RESULTS AND DISCUSSION

The proposed indoor-outdoor scene classification framework consists of three feature-based sub-frameworks to extract image features and a fully connected network for classification. Each sub-frameworks uplift quality features concerning object-level, image-level, and fine plus coarse features to construct a robust representative vector of the input image. The final vector undergoes normalization and dimension reduction, which are presented to the properly tuned FC network. Experimental evaluation of the random isolated test samples shows that our framework achieved an average detection accuracy of 95.84% over the unbalanced dataset for 20-fold cross-validation. The dataset images are not pre-processed or subjected to augmentation.

We iterated the model for times. One of the training/validation and loss reponses with respect to epochs (100) are shown in Figures 3 and 4. We cross-validataed the samples by selelcting training (75%), validation (10%) and test (15%) sets randomly from the available feature set. The classifier used RMSprop optimizer with softmax transfer function to classify the scene categories. The best performance trained the network to reach 98.76% with a loss of 0.04%. The test samples were calssified with an accuracy of 95.84% even though the network was unable to score the 100% mark during training.

Figure 5 shows some examples of ambiguities for both classes. Both the classes have samples that cannot be

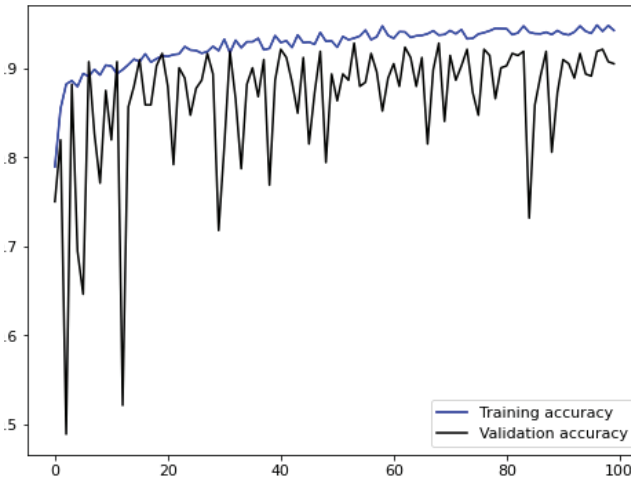

Figure 3. Plot showing the network performance as a function of epochs.

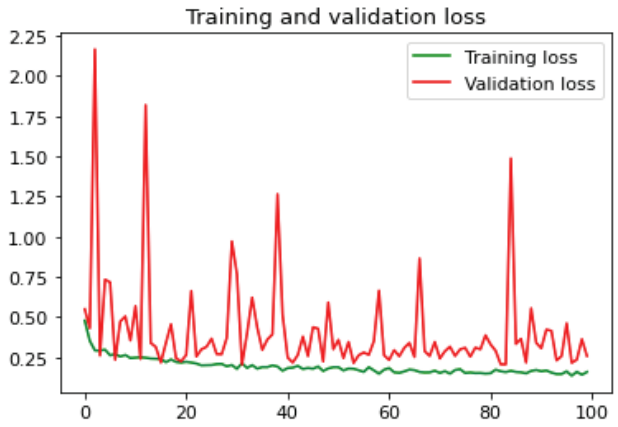

Figure 4. Plot showing network losses as a function of epochs.

	Indoor	outdoor	
Indoor	304	22	93.26
outdoor	7	425	98.42

Figure 6. Confusion Matrix for indoor/outdoor scene classification.

distinguished clearly even with naked eyes. The intra and inter-class similarity posses a vital challenge for a classifier. However, our proposed system worked better in distinguishing the representation accurately. Figure 6 represents the confusion matrix obtained over the dataset images. The indoor scenes were classified with lower accuracy (93.26%) than the outdoor scenes (98.42) due to object-only images as shown in Figure 5.

Several studies have employed different advanced learning models to carry out scene classification. Many such research works have used different dataset images for the classification task. Therefore, our work findings regarding performance metrics may not be fair and impartial. Still, the anaylsis shown in Table 4 offers an insight into various scene classification approaches with their performance while the same is graphically depicted in Figure 7.

The proposed framework outperforms other competing models except for the work introduced in [33] in which the author used the NYU depth dataset [34]. The self-generated dataset consists of diverse images from three datasets. The benchmark datasets used to generate our two-class scene dataset include CVPR09 datasets [45], Places365 [57], and UIUC Sports [58]. The motto behind the segregation of scene images from three distinct datasets was to intentionally induce complexity in recognizing the scenes and to test the validity of representative features. A few ambiguous samples were added to the classes as shown in Figure 7 to enhance the complexity further. Thus, the images from multiple datasets and ambiguous images creating a real-world problem impose difficulty in categorizing scenes on the framework. Since, diverse features were extracted using various descriptors, the CNN classifier was able to perform better in such a complex environment.

(a) Samples from indoor images

(b) Samples from outdoor images

Figure 5. Ambiguous samples from indoor (a) and outdoor (a) classes.

Table 4. Comparison with different baseline models

Reference	Year	Model	% Accuracy
[33]	2023	Deep Liquid State Machine	96%
[35]	2024	Customized Network and Whale Optimization	94.35%
[37]	2022	Object detection based TF-IDF Approach	83.63%
[52]	2015	Probabilistic Approach	45.71%
[53]	2019	Deep Learning	71.44%
[50]	2023	Deep Learning	75.8%
[54]	2021	Feature fusion + Graphical CNN	75%
[55]	2020	Deep Learning	95.6%
[56]	2019	Multi-modal attentive pooling network	67.7%
Proposed Approach	**2024**	**Diverse features + FC-CNN**	**95.84%**

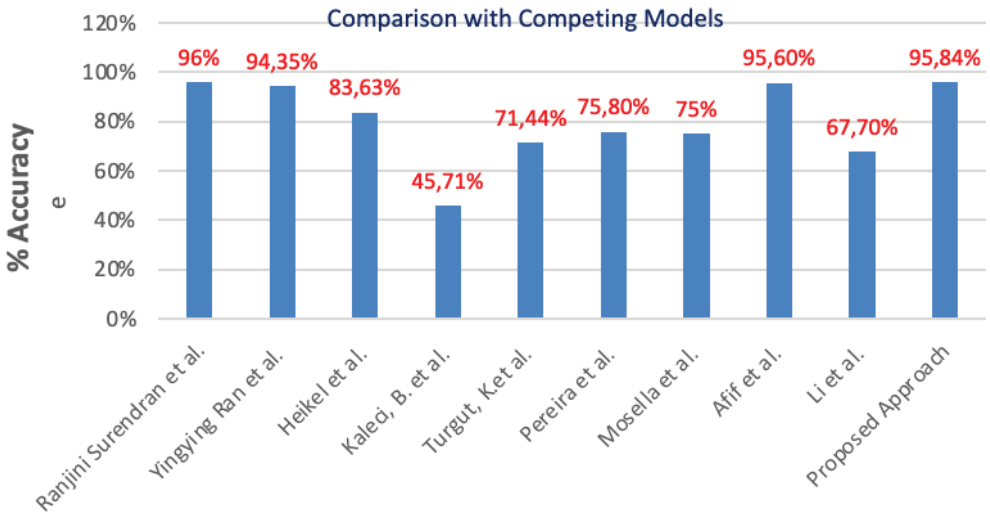

Figure 7. Comparison of proposed CNN-based framework on % accuracy.

CONCLUSION

The scene recognition framework introduced in this work uses content-based, scene-based, and a variety of handcrafted features to distinguish indoor/outdoor scenes from cross-domain datasets. The dataset used for the evaluation uses scene images from three distinct publicly available datasets to increase the complexity. Experimental analysis revealed that our scene recognition model obtained an average detection accuracy of 95.84% with 20-fold validation over 1000 epochs for each iteration. Thus, the proposed framework possesses a generalization ability over the unbalanced two-class samples. Other pre-trained models with a transfer learning approach can be tested over the features. Also, the fully connected CNN at the final stage can be replaced by a pre-trained network. The object-based features obtained for small and large objects were resized to 32x32 dimensions, which affects the quality of features, At the same time, the feature vector obtained for all the objects in the scene was summed up to obtain a single vector of dimension 512, which collectively represents all the objects in the image. Thus, larger objects eat up the contribution of smaller objects. Hence, a good approach is required that can add contributions made by smaller objects and larger objects by outputting a fixed dimension feature vector irrespective of the size of the objects and the number of objects in a given image. A single grayscale plane was considered for the handcrafted features. Using a suitable color space for the image would have improved the quality of the features. Even though the scene recognition framework introduced performed better in the complex scenario, the time required for training the network is somewhat large for 1000 epochs. The approach of sorting images into two classes was manual, however, the validity of the sorted images with respect to the classes has no ground truth. Lastly, all three-stage

feature extraction mechanisms can be put together, and a single network can be used to extract features and classify them as a future perspective. More images from different datasets can be added to the dataset or the existing dataset images can be augmented using suitable geometrical transformations. The custom CNN network can further be improved and tuned properly for better results.

ACKNOWLEDGEMENTS

This work was supported by the Research Department of the G.H. Raisoni, University, Amravati. The authors would like to thank for this support.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

[1] Oliva A. Gist of the scene. In: Itti L, Rees G, Tsotsos K, (editors). Neurobiology of Attention. New York: Elsevier Academic Press; 2005. p. 251–256. [CrossRef]

[2] Potter MC. Meaning in visual search. Science 1975;187:965–966. [CrossRef]

[3] Tversky B, Hemenway K. Categories of environmental scenes. Cogn Psychol 1983;15:121–149. [CrossRef]

[4] Anderson MD, Graf EW, Elder JH, Ehinger KA, Adams WJ. Category systems for real-world scenes. J Vis 2021;21:8. [CrossRef]

[5] Greene MR, Oliva A. Recognition of natural scenes from global properties: Seeing the forest without representing the trees. Cogn Psychol 2009;58:137–176. [CrossRef]

[6] Biederman I. On the semantics of a glance at a scene. In: Biederman I, editor. Percept Organ. Routledge; 1981. p. 213–253. [CrossRef]

[7] Friedman A. Framing pictures: The role of knowledge in automatized encoding and memory for gist. J Exp Psychol Gen 1979;108:316–355. [CrossRef]

[8] Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope. Int J Comput Vis 2001;42:145–175. [CrossRef]

[9] Oliva A, Torralba A. Building the gist of a scene: The role of global image features in recognition. Prog Brain Res 2006;155:23–36. [CrossRef]

[10] Furtak M, Mudrik L, Bola M. The forest, the trees, or both? Hierarchy and interactions between gist and object processing during perception of real-world scenes. Cognition 2022;221:104983. [CrossRef]

[11] Joubert OR, Rousselet GA, Fize D, Fabre-Thorpe M. Processing scene context: Fast categorization and object interference. Vision Res 2007;47:3286–3297. [CrossRef]

[12] Joubert OR, Fize D, Rousselet GA, Fabre-Thorpe M. Early interference of context congruence on object processing in rapid visual categorization of natural scenes. J Vis 2008;8:11. [CrossRef]

[13] Watson DM, Hartley T, Andrews TJ. Patterns of response to visual scenes are linked to the low-level properties of the image. Neuroimage 2014;99:402–410. [CrossRef]

[14] Watson DM, Hartley T, Andrews TJ. Patterns of response to scrambled scenes reveal the importance of visual properties in the organization of scene-selective cortex. Cortex 2017;92:162–174. [CrossRef]

[15] Wiesmann SL, Vo ML-H. What makes a scene? Fast scene categorization is a function of global scene information at different resolutions. J Exp Psychol Hum Percept Perform 2022;48:871–888. [CrossRef]

[16] Davenport JL, Potter MC. Scene consistency in object and background perception. Psychol Sci 2004;15:559–564. [CrossRef]

[17] Leroy A, Faure S, Spotorno S. Reciprocal semantic predictions drive categorization of scene contexts and objects even when they are separate. Sci Rep 2020;10:8447. [CrossRef]

[18] Henderson JM, Larson CL, Zhu DC. Full scenes produce more activation than close-up scenes and scene-diagnostic objects in parahippocampal and retrosplenial cortex: An fMRI study. Brain Cogn 2008;66:40–49. [CrossRef]

[19] Troiani V, Stigliani A, Smith ME, Epstein RA. Multiple object properties drive scene-selective regions. Cereb Cortex 2014;24:883–897. [CrossRef]

[20] Espinace P, Kollar T, Soto A, Roy N. Indoor scene recognition through object detection. In: Espinace P, editor. 2010 IEEE International Conference on Robotics and Automation. IEEE; 2010. p. 1406–1413. [CrossRef]

[21] Herranz L, Jiang S, Li X. Scene recognition with CNNs: objects, scales and dataset bias. 2016 IEEE Conf Comput Vis Pattern Recognit 2016;571–579. [CrossRef]
[22] Khan SD, Othman KM. Indoor scene classification through dual-stream deep learning: A framework for improved scene understanding in robotics. Computers 2024;13:121. [CrossRef]
[23] Gua N, Jiang M, Wang D, Zhou X, Song Z, Li Y, Gao L, Luo J. Scene classification for remote sensing image of land use and land cover using a dual-model architecture with multilevel feature fusion. Int J Digit Earth 2024;17:2353166. [CrossRef]
[24] Bosch A, Muñoz X, Martí R. Which is the best way to organize/classify images by content. Image Vis Comput 2007;25:778–791. [CrossRef]
[25] Brown M, Susstrun SK. Multi-spectral SIFT for scene category recognition. In: 2011 IEEE Conf Comput Vis Pattern Recognit; 2011 Jun 20–25; Colorado Springs, CO, USA. p. 177–184. [CrossRef]
[26] Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope. Int J Comput Vis 2001;42:145–175. [CrossRef]
[27] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proc IEEE Conf Comput Vis Pattern Recognit; 2005 Jun 20–26; San Diego, CA, USA. Vol. 1. p. 886–893. [CrossRef]
[28] Bay H, Ess A, Tuytelaars T, van Gool L. Speeded-up robust features (SURF). Comput Vis Image Underst 2008;110:346–359. [CrossRef]
[29] Yang J, Jiang YG, Hauptmann A, Ngo CW. Evaluating bag-of-visual-words representations in scene classification. In: Proc Int Workshop Multimedia Inf Retr; 2007 Sep 24–29; Augsburg, Germany. IEEE. p. 197–206. [CrossRef]
[30] Sanchez J, Perronnin F, Mensink T, Verbeek J. Image classification with the fisher vector: Theory and practice. Int J Comput Vis 2013;105:222–245. [CrossRef]
[31] Jegou H, Douze M, Schmid C, Pérez P. Aggregating local descriptors into a compact image representation. In: Proc IEEE Conf Comput Vis Pattern Recognit; 2010 Jun 13–18; San Francisco, CA, USA. p. 3304–3311. [CrossRef]
[32] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:7553. [CrossRef]
[33] Surendran R, Chihi I, Anitha J, Hemanth DJ. Indoor scene recognition: An attention-based approach using feature selection-based transfer learning and deep liquid state machine. Algorithms 2023;16:430. [CrossRef]
[34] Silberman N, Hoiem D, Kohli P, Fergus R. Indoor segmentation and support inference from RGB-D images. In: Proc 12th Eur Conf Comput Vis (ECCV); 2012 Oct 7–13; Florence, Italy. p. 746–760. [CrossRef]
[35] Ran Y, Xu X, Luo M, Yang J, Chen Z. Scene classification method based on multi-scale convolutional neural network with long short-term memory and whale optimization algorithm. Remote Sens 2024;16:174. [CrossRef]
[36] Anbarasu B, Anitha G. Vision-based position estimation and indoor scene recognition algorithm for quadrotor navigation. J Phys Conf Ser 2021;1969. [CrossRef]
[37] Heikel E, Espinosa-Leal L. Indoor scene recognition via object detection and TF-IDF. J Imaging 2022;8:209. [CrossRef]
[38] Kuznetsova A, Rom H, Alldrin N, Uijlings I, Pont-Tuset J, Kamali S, et al. The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale. Int J Comput Vis 2020;128:1956–1981. [CrossRef]
[39] Qian X, Lin X, Yu L, Zhou W. FASFLNet: Feature adaptive selection and fusion lightweight network for RGB-D indoor scene parsing. Opt Express 2023;31:27. [CrossRef]
[40] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proc IEEE Conf Comput Vis Pattern Recognit; 2017 Jul 21–26; Honolulu, HI, USA. p. 2261–2269. [CrossRef]
[41] Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. arXiv 2016; arXiv:1602.07360.
[42] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc IEEE Conf Comput Vis Pattern Recognit; 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–778. [CrossRef]
[43] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ICLR. arXiv 2014; arXiv:1409.1556.
[44] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proc IEEE Conf Comput Vis Pattern Recognit; 2015 Jun 7–12; Boston, MA, USA. p. 1–9. [CrossRef]
[45] Quattoni A, Torralba A. Recognizing indoor scenes. In: Proc IEEE Conf Comput Vis Pattern Recognit; 2009 Jun 20–25; Miami, FL, USA. p. 413–420. [CrossRef]
[46] Silberman N, Hoiem D, Kohli P, Fergus R. Indoor segmentation and support inference from RGB-D images. In: Proc 12th Eur Conf Comput Vis (ECCV); 2012 Oct 7–13; Florence, Italy. p. 746–760. [CrossRef]
[47] Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: A 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell 2017;40:1452–1464. [CrossRef]
[48] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit (CVPR); 2006 Jun 17–22; New York, NY, USA. Vol. 2. p. 2169–2178. [CrossRef]

[49] Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A. SUN database: Large-scale scene recognition from abbey to zoo. In: Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit (CVPR); 2010 Jun 13–18; San Francisco, CA, USA. p. 3485–3492. [CrossRef]

[50] Pareira R, Barros T, Garrote L, Lopes A, Nunes UJ. A deep learning-based global and segmentation-based semantic feature fusion approach for indoor scene classification. Pattern Recognit Lett 2024;179:24–30. [CrossRef]

[51] Alosaimi N, Alchicri H, Bazi Y, Ben Youssef B, Alajlan N. Self-supervised learning for remote sensing scene classification under the few shot scenario. Sci Rep 2023;13:433. [CrossRef]

[52] Kaleci B, Senler CM, Dutagaci H, Parlaktuna O. A probabilistic approach for semantic classification using laser range data in indoor environments. In: Proc Int Conf Adv Robot; 2015 Jul 27–31; Istanbul, Turkey. [CrossRef]

[53] Turgut K, Kaleci B. A deep learning architecture for place classification in an indoor environment via 2D laser data. In: Proc 3rd Int Symp Multidiscip Stud Innov Technol (ISMSIT); 2019 Oct 11–13; Ankara, Turkey. [CrossRef]

[54] Mosella-Montoro A, Ruiz-Hidalgo J. 2D–3D geometric fusion network using multi-neighborhood graph convolution for RGB-D indoor scene classification. Inf Fusion 2021;76:46–54. [CrossRef]

[55] Afif M, Ayachi R, Said Y, Atri M. Deep learning-based application for indoor scene recognition. Neural Process Lett 2020;51:2827–2837. [CrossRef]

[56] Li Y, Zhang Z, Cheng Y, Wang L, Tan T. MAPNet: Multi-modal attentive pooling network for RGB-D indoor scene classification. Pattern Recognit 2019;90:436–449. [CrossRef]

[57] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proc IEEE Conf Comput Vis Pattern Recognit (CVPR); 2015 Jun; Boston, MA, USA. p. 1–9. [CrossRef]

[58] [Dataset] UIUC Sports Event Dataset. Available at: https://www.kaggle.com/datasets/trolukovich/uiuc-sports-event-dataset Accessed Jul 23, 2025.