Journal of Information Systems Engineering and Management
2025, 10(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

An LSTM and Conventional, Global, and Object-Based Semantic Feature Fusion Framework for Indoor-Outdoor Scene Classification

Pandit T. Nagrale[1], Sarika Khandelwal[2]

[1]Ph.D. Scholar, Computer Science & Engineering
G H Raisoni University, Amravati, Maharashtra, India
ptnagrale@gmail.com
[2]Associate Professor, Computer Science & Engineering
G H Raisoni College of Engineering, Nagpur, Maharashtra, India
sarikakhandelwal@gmail.com

ARTICLE INFO

Received: 05 Oct 2024

Revised: 02 Dec 2024

Accepted: 16 Dec 2024

ABSTRACT

This article proposes a novel approach that uses diverse features from a scene image using a variety of local, global, and object-based descriptors. The regional and image-based features are obtained using various conventional statistical descriptors, whereas a deep network VGG19 is used to extract the scene's global features. Scene objects based on specific features are concatenated to regional and global features after they are segmented using another deep network, YOLOV5m. While conventional features represent intensity-based characteristics, global and object features carry color depth details of the scene image. A long-term short memory (LSTM) network with a fully connected (FC) dense layer network is trained over images from four benchmark cross-datasets. Experimental evaluation over 5081 indoor-outdoor scene images showed that the proposed scene classification approach obtained 96% accuracy in identifying both categories on 15% test sample images.

Keywords: Object-based descriptors, image-based features, statistical descriptors, deep network, LSTM, fully connected network.

Introduction

Scene classification is primarily analyzed based on the objects in the scene and the background [1]. The computer vision task is more complicated and challenging due to several commonness of the objects belonging to both indoor and outdoor categories [2]. For example, a bicycle may be parked outside in the parking area and may be kept in the living room of a house or hotel. Today, a perception module to categorize scenes is integrated into a robot [3, 4]. Several contents of a scene image can cover the aforementioned conditions that may belong to both categories [5]. For an efficient indoor-outdoor scene classification, intra-category differences or variations are crucial and need careful consideration. This is especially important when there are large changes in the same category and very few common patterns are available. The appearance similarities increase as a function of the number of scene classes. Therefore, the accurate scene discrimination task becomes more difficult due to the shared appearances across scene categories as a result of diluted inter-scene boundaries [6]. This issue leads to ambiguity in inter-class scenes.

Human responds to objects and their distribution across the scene images to discriminate scene categories [7]. Two major issues including the intra-class variations and inter-class ambiguity inspired researchers [5, 6, 8, and 9] to overcome them by exploiting object details and their correlations called semantic information across the scene. Recent work thus made use of object detection networks to locate objects in the scene using bounding boxes and identify their categories. Conversely, obtaining a well-defined object mask using semantic segmentation can be a better approach [6] that utilizes pixel-level object detection (MASK RCNN, YOLOV5, etc. detects 80 objects such as bicycles, cats, pens, knives, mobile, and chairs…). This approach results in more accurate object-based spatial distribution.

This paper introduces a novel approach, object segmentation-based semantic features, image-level global features, and eight different handcrafted or conventional features that are finally concatenated to form a meaningful feature representation of the scene. Objects in the scene image are segmented using a YOLV5m network and resized to 32x32 dimension to obtain 512 feature elements through a VGG19 network. This is because different objects in the scene are of different dimensions. For global or image-level features, another VGG19 network is used to obtain 512 features at the image level. Both the VGG19 networks are directly used to detect objects and extract image-level features trained using the ImageNet dataset. Thus, an image is represented by a wide variety of feature distributions forming a row vector of 3318 elements. Moreover, a custom network comprised of an LSTM layer followed by three dense layers is used to distinguish between the indoor and the outdoor scenes with properly tuned hyper-parameters.

The main contribution of the proposed scene classification framework can be summarized as:

- Object-level semantic details are extracted by segmenting the scene objects using the YOLOV5m network, further resized to a common dimension of 32x32, and represented with a 512-element feature vector using the VGG19 network.
- Image-level global features are obtained by subjecting the scene image to a VGG19 network. For handcrafted features, eight different descriptors operating on patch and image levels are incorporated to contribute.
- Experiments conducted on the concatenated features using a sequential LSTM network showed that the proposed scene classification framework showed remarkable classification accuracy on cross-dataset images.

Related Work

Researchers have been piqued by the overwhelming response of various machine and deep learning techniques in vast applications across a variety of domains. As a result, several researchers have shown interest and accepted the challenge in the area of scene classification problems [2]. An extensive review of past literature showed that the majority of methodologies adhere to a three-stage framework [2]: extraction of features, feature transformation and concatenation, and prediction. However, scenes are prone to ambiguity and variations, therefore networks-based blind features are not sufficient to represent scene details to a distinguishable level [13]. To overcome these issues, a new scene representing features is required, especially the semantic and multimodal features.

Since color depth features contribute little semantic information that includes the shape of the objects, color-based features using a multimodal approach have been suggested in [7, 9, and 14]. The work introduced in [7] a correlation approach for features that inculcate the knowledge of correlating features from each modality. An improvement technique was the focus of work in [9] where RGB-D semantic cues acquired from the region proposal technique were integrated. A high-level representation from multilevel features of the CNN-based RGB-D features was obtained using a randomized recursive network [14]. Thus the multi-modality approaches showed more discriminative characteristics than the single-modal features for the scene images. However, such approaches relied on the shapes of the scene objects while neglecting the type or categories and their respective distributions in the scene. The work proposed in [7, 9, 14] neglected the important aspect which possesses the ability to disentangle prediction.

Scenes were better represented by meaningful details by adding object-based features to blind features obtained using the CNN-based global features in [5, 8, 15-17]. The former CNN-based global features were exploited by object occurrence [18]. An improvement to the occurrence of the object, the work suggested in [5] computed the distance between the objects in the scene and used an inter-object separation relationship. Authors in [15] focused on object-to-object correlation for representation based on spatial layout. On the other hand, co-occurring objects and object occurrence relationship was exploited in [8] and an improved model based on two object-based scene representations was utilized. Moreover, despite using CNN-based blind features and object-based features and obtaining promising results, the latter features failed to provide significant information about the complete or overall scene such as floors, sky, walls, backgrounds, etc. In other words, these techniques failed to capture the pixel-level object's actual structures which can significantly contribute to disentangling the predictions.

Semantic Segmentation masks are commonly used in most scene classification techniques other than object detection for object-based features [6, 19, and 20]. Object representation using multi-object categorization was carried out in [20]. Objects in the scene were located using segmentation masks and their pixel representation was obtained using a histogram-like method. Work introduced in [6] extracted two sets of features using two separate branches of a network exploiting global and local features. The RGB image was used for global features, and the segmentation mask simultaneously worked out the local semantic features from the network. Adding object-based and scene-based features to conventional features, the overall image representation was enriched to improve the prediction accuracy in this article. Object-based features are acquired firstly by segmenting the objects by YOLOV5m network from the scene and then feature extraction is carried out from a predefined size of the object. Image-level information is added to the object-based set by acquiring the image features using the VGG19 network. Also, patch-based and global statistical features are added to enhance the image representation quality through conventional features obtained on the grayscale version of the scene image.

Material and Method

Figure 1 provides an overview of the proposed framework, consisting of three branches: the object-based semantic, the global, and the conventional branch. The output features from all three branches undergo a feature fusion mechanism and are finally fed to the predictor network.

The first two branches are operated on RGB images while the third branch computes the conventional features from the grayscale equivalent of the original scene image. The object-based features are extracted using the original dimension of the image (512 x 512), whereas global features are acquired from the color image with half the dimension (256 x 256). For eight conventional features, the original image is reduced to a dimension of 128 x 128 and then converted to grayscale.

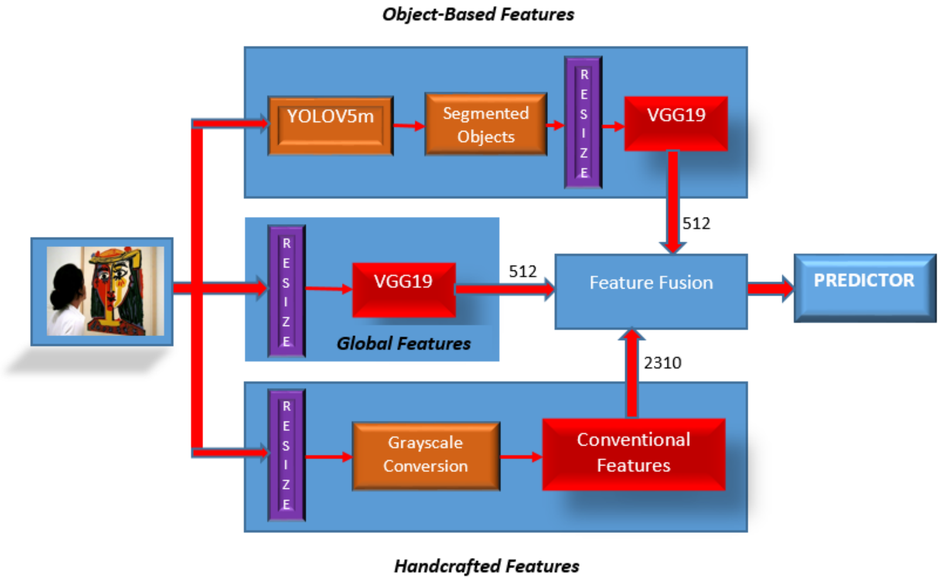


Figure 1 – The proposed scene classification framework

Object Segmentation and Semantic Features

Experiments conducted over several images from the generated datasets showed that the 32x32 dimension is a compromise between the largest and smallest object detected from the scene images. The possibility of losing significant information from resizing operations of the segmented objects using the YOLV5m network was considered while deciding the resize dimension. The resize dimension selected was a better compromise between all the segmented objects from various scene images. Unlike other approaches such as in [21], instead of using statistical descriptors (2D average position and standard deviation), blind features on the resized segmented objects were used to represent the semantic information. To extract the semantic information, a VGG19 network that was trained on the ImageNet dataset was used, and features from the last fully connected layer were added to the fusion module. Each segmented object was represented by a 512-element vector from the fully connected layer of the VGG19 network. Figure 2 below shows the VGG19 network, which is modified and utilized to extract blind features from YOLOV5m segmented objects.

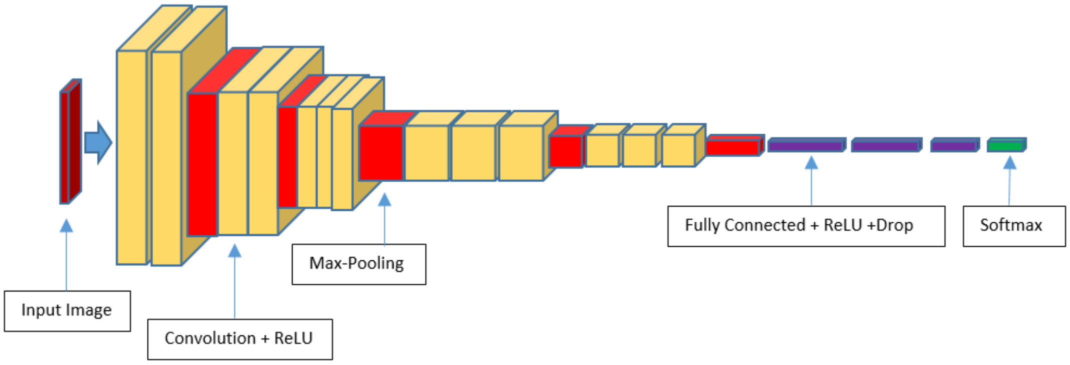


Figure 2 – Basic structure of VGG19 Network

Global Features

The original image of 512x512x3 dimension was resized to 256x256x3 to extract the global aspects of the scene image such as objects in the scene, their spatial correlations, structures, and characteristics. Global features play a significant role in the actual scene representation. Therefore, RGB image-based global features are extracted through the VGG19, a deep-learned network that acts as the backbone of the fully connected layer. The output of the fully connected layer is reshaped into 8 x 8 x 512 and the sum along both axes is considered to output a 1x512 feature vector for the input scene image.

Grayscale image-based Conventional features

Object-based features no matter are the best form of features to represent the semantic details of a scene image. However, YOLOV5m is limited to detecting a few certain objects from real-world scenarios. The ability of the YOLOV5m network can be improved through transfer learning but it requires being trained on numerous real-world existing objects which is complicated.

To compensate, eight different types of diverse patch (fine) and coarse-level features are added to the object-based and global features. The importance of conventional features in classification problems has already been proven in many applications. The RGB image was converted to grayscale and resized to 128 x 128 and various descriptors were applied to obtain a 2310-element feature vector for representing the image in different aspects. Edge information was contributed using two edge-based filters: the Sobel filter and a matched filter. The matched filter uses the following kernel for extracting the edge details from the image which is depicted below in Figure 3. A total of 256 elements using the matched filter and 32 elements using the Sobel operator were extracted from the scene image.

-1	-1	-1
-1	8	-1
-1	-1	-1

Figure 3 – Kernel used for the Matched filter-based descriptor

Wavelet-based features are extracted using six different mother wavelets. The representation includes the energy and the magnitude of the high-low and the high-high component of first-order wavelet coefficients obtained using 'bior3.1', 'bior3.5', 'bior3.7', 'db3', 'sym3', and 'haar' wavelets. The following expressions (1) and (2) were used to compute the magnitude and the energy of the high-low and the high-high components.

$$\text{Magnitude, M} = \frac{1}{mxn} \sum_{x=1}^{m} \sum_{y=1}^{n} \left| C_{x,y} \right| \qquad (1)$$

$$\text{Energy, E} = \frac{1}{mxn} \sum_{x=1}^{m} \sum_{y=1}^{n} C_{x,y}^{2} \qquad (2)$$

Where ‘m’ and ‘n’ are the height and the width of the wavelet component. ‘$C_{x,y}$’ is the first-order wavelet coefficient at (x, y) spatial location. For six different wavelets, two magnitudes and two energy features resulting in 24 elements were extracted from the grayscale image.

Texture information has always been crucial in representing the global information of any image. Fine details from the grayscale image were extracted using an LBP descriptor using a radius of 3. The texture pattern resulting from the LBP descriptor was summed along rows and columns and then concatenated to form a 256-element feature vector. More fine details were added using a 3x3 patch-based LBP image. A textured image was obtained after thresholding the 3x3 neighborhood using the center pixel and by following a single read-out pattern [out of 55 significant patterns]. The textural image was reduced in dimension by averaging values over a 5x5 window to result in a 26x26 array. The 2D values were flattened to generate a 676 dimension vector thus adding the textural details. The following Figure 4 shows how the LBP features were used to add texture information.

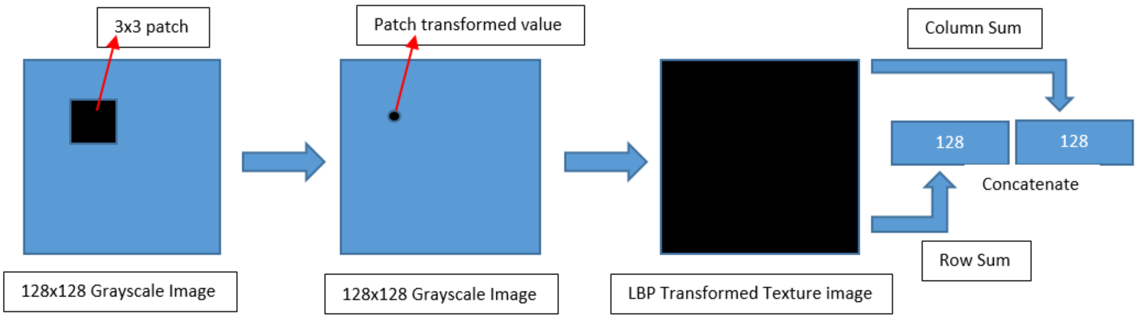


Figure 4 – LBP-based feature extraction. 3x3 patch to compute the LBP equivalent for the center pixel. The value computed for the 3x3 mask. Equivalent texture image. Feature vector considered.

A similar approach was used in the second case, except the LBP values over a 5x5 window were averaged from the texture.

HOG features using 64x64 pixels per cell with 9 orientations were obtained and normalized using the max-normalization technique. Further, they were transformed in the range [0 255] to contribute a 256 dimension vector to the feature set. Lastly, the grayscale image is represented in its original form by averaging values in 4x4 non-overlapping blocks. A feature vector of 6 x 6 = 36 dimensions is thus added to the set of handcrafted features.

To ensure the contribution of the pixel values of the scene image in their original form, mean values corresponding to 4x4 non-overlapping windows were computed and added to the feature set. Thus, a feature vector of 1024 values is used to directly represent the scene image. Lastly, the overall scene image contrast, energy, homogeneity, correlation, ASM, and dissimilarity were measured using the Gray Level Co-occurrence Matrix. Global-level indicators concerning 6 parameters were used to enhance the performance of the classifier.

The following Figure 5 shows the distribution of coarse and fine features along their dimensions.

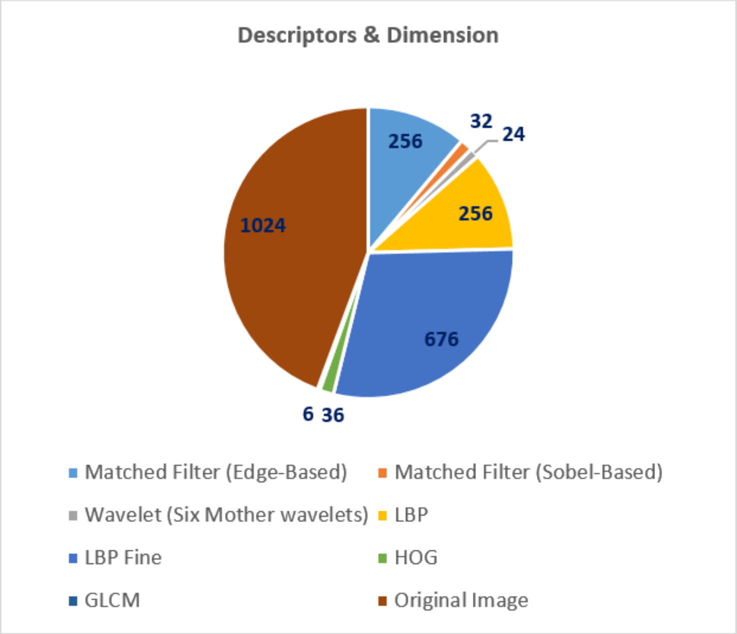


Figure 5 – Descriptors and their feature lengths

Finally, the dimensions of all the features extracted using the object information, scene details, and statistical descriptors are shown the Table 1 below.

Table 1 – Various features, descriptions, and their dimension

Feature	Description	Dimension
Object-Based	512x812-YOLOV5m-32x32-VGG19	1x512
Scene-Based	512x512-256x256-VGG19	1x512
Conventional	512x512-128x128-Grayscale-Descriptors (Edge descriptors, wavelets, LBP, HOG, image-level, and GLCM)	256+32+24+256+676+36+1024+6= 1x2310
	Total Feature Vector	**3334**

Experiments

The proposed approach was evaluated over a self-generated dataset. The dataset used in this work includes indoor and outdoor images from multiple datasets. We have used images from Places365 [22], UIUC Sports [23], and indoor CVPR09 datasets [24]. This was done to increase the classification complexity and test the robustness of the diverse features obtained through the 3-parallel feature extraction stages with our scene recognition framework. On the other hand, a few images in the generated dataset were ambiguous regarding the class but were considered in either of the classes. Some of the ambiguous images from both indoor and outdoor classes are shown in Figures 6 and 7 respectively. A total of 5081 images were collected from three different datasets [22-24] and partitioned manually into two classes. Table 2 shows the distribution of indoor and outdoor concerning their samples.

Table 2 – Number of samples in two categories: Indoor and Outdoor

Category	Number of samples
Indoor	2240
Outdoor	2841
Total	5081

Another complexity introduced while generating the dataset was the uneven number of sample images in the categories. The two classes clearly show that the dataset constructed is unbalanced. To ensure that a scene image is accurately placed in respective categories, the samples from both categories were scanned by three different subjects. There were no issues regarding the samples acquired from the CVPR09 dataset since they belonged to the indoor class only. It is difficult to classify the scenes correctly. Also, inter-class and intra-class similarities pose a great challenge for a classifier to discriminate between both categories. However, the proposed computer vision method can achieve better performance in such scenarios.

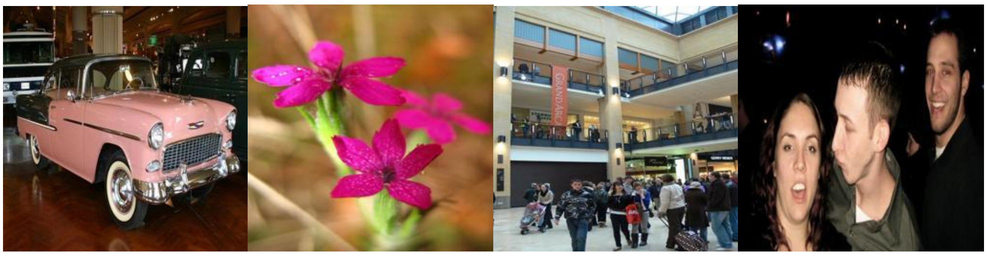


Figure 6 – Some ambiguous images from indoor class. A car is parked inside. Flowers surrounding unknown. People are seen in the surrounding environment. The family's surroundings are unknown.

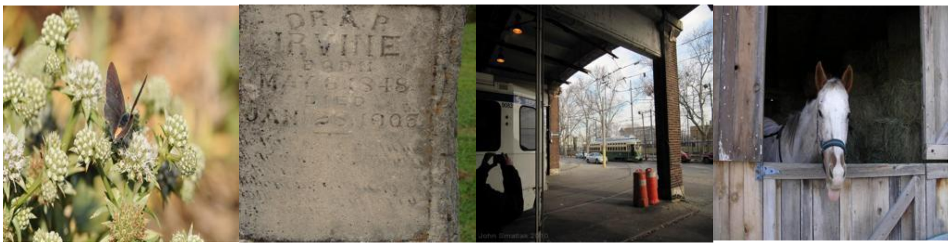


Figure 7 – Some ambiguous images from outdoor class. Background blurred for flowers. Epitaph with missing environment. Mixed surrounding environment. Horse partly inside stable.

Implementation Details

The multi-modal features extraction framework & LSTM network was coded in Python 3.9 on an Anaconda-based Spyder environment. The system properties were: 11th Gen Intel(R) Core(TM) i5-11500 @ 2.70GHz, 16.0 GB, 512 GB SSD, Windows 11 operating system. The main focus of the work was not only to test the robustness of the features but also to assess the generalization capabilities over an unbalanced dataset. We trained and tested the network several times for cross-validation to test the network with different sets of training, validation, and test samples. The sequential LSTM-based fully connected network employed was trained for 30 epochs with 75% training and 10% validation samples. The rest of the 20% samples from the dataset were used to test the performance of the network. The custom sequential LSTM-Fully connected network is shown in Figure 8.

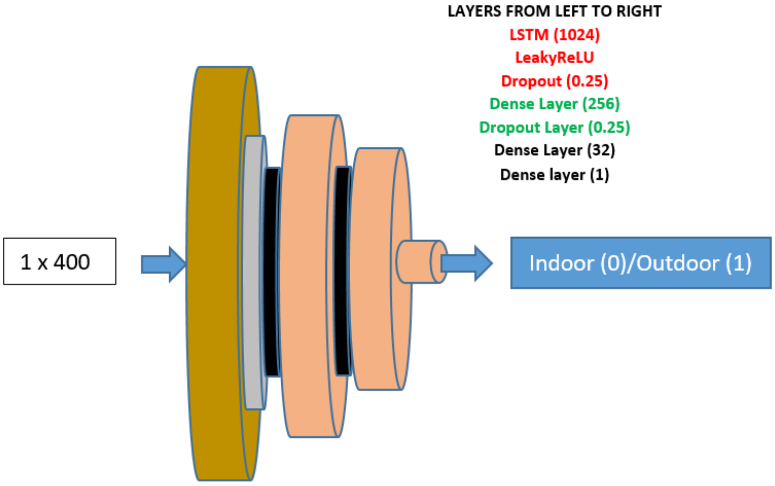


Figure 8 - The custom sequential LSTM-Fully Connected Network

Manual inspection showed that out of 3334 elements in the feature vector obtained from a single scene image, it was found that the features from 1296 to 1311 included significant zero values and therefore neglected. However, the remaining 3318 elements were subjected to principal component analysis (PCA) for dimension reduction. Conducting several experiments on the available 5081 samples, and choosing random samples for training, testing, and validation, it was found that 400 components of PCA were minimum and sufficient for better performance. The batch size was set to 120 and the network was trained using 30 epochs.

Figures 9 and 10 show training and validation accuracies and loss as a function of epochs for the best performance out of 50 iterations. We have used the 'RMSprop()' optimizer, with 'accuracy' as the metrics, and 'sigmoid' as the activation in the last dense layer.

The network reached was trained to 100%. The maximum accuracy obtained on the test set (from 30 iterations) was found to be 97.04%. The network succeeded in learning at its best (100%), and the performance in classifying the unknown samples was better (97.04%) as compared to other competing scene prediction techniques.

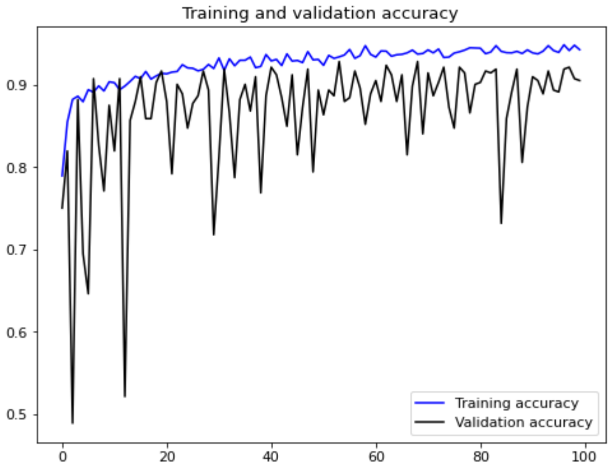


Figure 9 – Training and validation accuracies as a function of epochs.

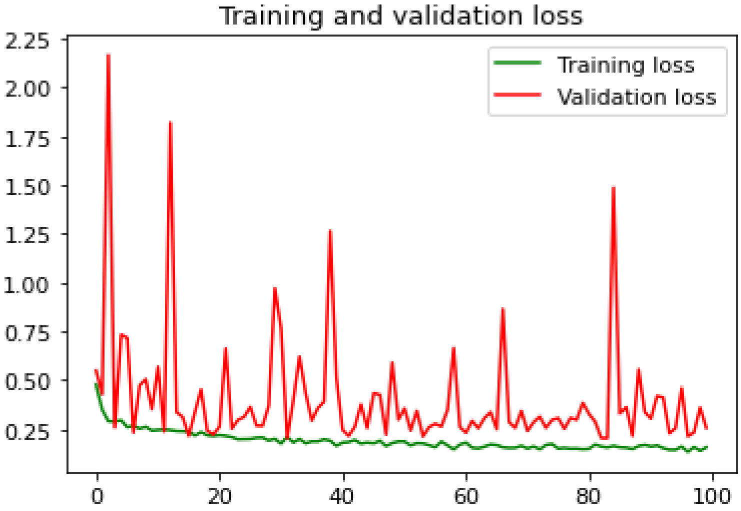
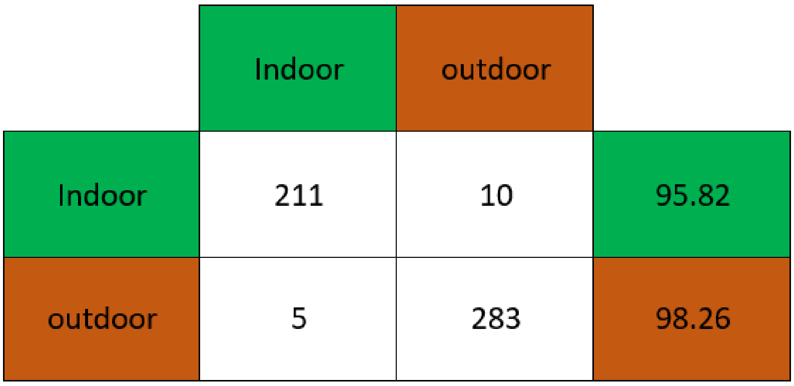


Figure 10 - Training and validation losses as a function of epochs.

The 20% random samples to test selected from indoor/outdoor categories for an iteration about the best performance were 221/288 respectively from 509 total samples. The confusion matrix shown in Figure 10 shows that 211 samples from the indoor category were classified accurately and 10 were placed in the wrong group. On the other hand, out of 288 outdoor samples, 283 samples were placed in the outdoor class and 5 were misclassified.



	Indoor	outdoor	
Indoor	211	10	95.82
outdoor	5	283	98.26

Figure 10 – Confusion Matrix for the test samples

Comparison

We selected nine other competing techniques to compare the performance of the proposed scene classification framework. Most of the techniques found in the literature dealt with indoor classification problems. Also, the datasets used in their work differ, and therefore our findings relating to the classification accuracy may not be truly equitable. However, the comparison shown in Table 3 provides insight regarding the performance of various scene classification methods in ascending order concerning the publication. Even though, our self-generated dataset includes cross-dataset images from three different benchmark datasets and includes ambiguity in samples belonging to both categories, the classification accuracy with the proposed scene classification framework is higher than other methods.

Table 3 - Comparison of existing approaches with the proposed scene classification framework.

Author (Reference)	Year	Methodology Used	Accuracy
Kaleci, B. et al. [25]	2015	K-Means	45.71%
Li et al. [9]	2019	Network + Attentive pooling	67.7%
Turgut, K.et al. [26]	2019	MLP	71.44%
Afif et al. [27]	2020	CNN + scaling	95.6%
Mosella et al. [28]	2021	Feature fusion + Graphical CNN	75%
Heikel et al. [29]	2022	YOLO + TF+IDF	83.63%
Pereira et al. [21]	2023	Segmentation + Multiple Networks + Fused features	75.8%
Ranjini Surendran et al. [30]	2023	Segmentation + DenseNet201 + World Cup organization + LSM Classifier	96%
Yingying Ran et al. [31]	2024	Multi-scale CNN + LSTM + Whale optimization	94.35%
Proposed Approach	**2024**	**Multi-scale features + (LSTM + Fully connected Network)**	**97.04%**

Conclusion

The multi-scale scene feature-plus LSTM network–based recognition framework introduced in this article is evaluated on self-generated cross-dataset samples to discriminate indoor and outdoor scene images. The multiscale features include object-based blind features, scene-based global features, and eight distinct conventional features. Object-oriented features are employed to capture the semantic information from the scene, whereas scene-based global features add spatial information relating to the contents of the scene. On the other hand, the handcrafted features are incorporated to ensure fine and coarse details are added. Three patch level and three image level descriptors are used to capture the local details to enhance the feature set and assist the predictor for better performance. Experiments conducted on the self-generated dataset images showed that the best results obtained on 20% of test samples randomly chosen were classified with 97.04% accuracy which is found to be superior to other state-of-the-art work.

The framework exploits the unbalanced nature of the dataset and inter-class ambiguity and performs well with small epochs and batch size. The feature vector was reduced in dimension using the PCA and 400 components were used to represent the complete feature vector. The work can be extended by adding more images to the datasets. The YOLOV5m network is limited to identifying objects. More objects can be identified through transfer learning using a large scene dataset. Lastly, the LSTM network can be fine-tuned to improve classification accuracy.

References

[1] Z. Yi, T. Chang, S. Li, R. Liu, J. Zhang, A. Hao, Scene-aware deep networks for semantic segmentation of images, IEEE Access 7 (2019) 69184–69193.

[2] L. Xie, F. Lee, L. Liu, K. Kotani, Q. Chen, Scene recognition: A comprehensive survey, Pattern Recognit. 102 (2020) 107205.

[3] Y. Zhang, H. Chen, K. Yang, J. Zhang, R. Stiefelhagen, Perception framework through real-time semantic segmentation and scene recognition on a wearable system for the visually impaired, in: IEEE International Conference on Real-Time Computing and Robotics, RCAR, 2021.

[4] R. Pereira, A. Cruz, L. Garrote, G. Pires, A. Lopes, U.J. Nunes, Dynamic environment-based visual user interface system for intuitive navigation target selection for brain-actuated wheelchairs, in: IEEE International Conference on Robot and Human Interactive Communication, RO-MAN, 2022.

[5] R. Pereira, L. Garrote, T. Barros, A. Lopes, U.J. Nunes, A Deep Learningbased Indoor Scene Classification Approach Enhanced with Inter-Object Distance Semantic Features, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2021.

[6] A. López-Cifuentes, M. Escudero-Viñolo, J. Bescós, A. García-Martín, Semanticaware scene recognition, Pattern Recognit. 102 (2020).

[7] Y. Li, J. Zhang, Y. Cheng, K. Huang, T. Tan, DF2Net: Discriminative feature learning and fusion network for RGB-D indoor scene classification, AAAI Conf. Artif. Intell. (2018).

[8] X. Song, S. Jiang, B. Wang, C. Chen, G. Chen, Image representations with spatial object-to-object relations for RGB-D scene recognition, IEEE Trans. Image Process. 29 (2020) 525–537.

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

[10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: European Conference on Computer Vision, ECCV, 2018.

[11] S. Song, S.P. Lichtenberg, J. Xiao, SUN RGB-D: A RGB-D scene understanding benchmark suite, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015.

[12] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from RGB-D Images, in: European Conference on Computer Vision, ECCV, 2012.

[13] Z. Xiong, Y. Yuan, Q. Wang, RGB-D scene recognition via spatial-related multi-modal feature learning, IEEE Access 7 (2019).

[14] A. Caglayan, N. Imamoglu, A.B. Can, R. Nakamura, When CNNs meet random RNNs: Towards multi-level analysis for RGB-D object and scene recognition, Comput. Vis. Image Underst. 217 (2022).

[15] X. Song, C. Chen, S. Jiang, RGB-D scene recognition with object-to-object relation, in: ACM International Conference on Multimedia, 2017.

[16] X. Cheng, J. Lu, J. Feng, B. Yuan, J. Zhou, Scene recognition with objectness, Pattern Recognit. 74 (2018) 474–487.

[17] L. Zhou, J. Cen, X. Wang, Z. Sun, T.L. Lam, Y. Xu, BORM: Bayesian Object Relation Model for Indoor Scene Recognition, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2021.

[18] R. Pereira, N. Gonçalves, L. Garrote, T. Barros, A. Lopes, U.J. Nunes, Deep-learning based global and semantic feature fusion for indoor scene classification, in: IEEE International Conference on Autonomous Robot Systems and Competitions, ICARSC, 2020.

[19] C. Herranz-Perdiguero, C. Redondo-Cabrera, R.J. López-Sastre, In pixels we trust: From pixel labeling to object localization and scene categorization, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2018.

[20] A. Ahmed, A. Jalal, K. Kim, A novel statistical method for scene classification based on multi-object categorization and logistic regression, Sensors 20 (14) (2020).

[21] Ricardo Pereira, Tiago Barros, Luis Garrote, Ana Lopes, and Urbano J. Nunes, “A ddep learning-based global and segmentation-based semantic feature fusion approach for indoor scene classification,” Pattern Recognition Letters, 179, 2024, pp. 24-30.

[22] C. Szegedy et al. (2015) Going deeper with convolutions. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 1–9. doi: 10.1109/CVPR.2015.7298594

[23] https://www.kaggle.com/datasets/trolukovich/uiuc-sports-event-dataset.

[24] Quattoni, A.; Torralba, A. Recognizing indoor scenes. In Proceedings of the IEEE Conference on Computer and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 413–420.

[25] Kaleci, B.; Senler, C.M.; Duta˘gacı, H.; Parlaktuna, O. A probabilistic approach for semantic classification using laser range data in indoor environments. In Proceedings of the 2015 International Conference on Advanced Robotics, Istanbul, Turkey, 27–31 July 2015.

[26] Turgut, K.; Kaleci, B. A Deep Learning Architecture for Place Classification in an Indoor Environment via 2D Laser Data. In Proceedings of the 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 11–13 October 2019.

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

[28] 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.

[29] Heikel, E.; Espinosa-Leal, L. Indoor Scene Recognition via Object Detection and TF-IDF. J. Imaging 2022, 8, 209.

[30] Surendran, R.; Chihi, I.; Anitha, J.; Hemanth, D.J. Indoor Scene Recognition: An Attention-Based Approach Using Feature Selection-Based Transfer Learning and Deep Liquid State Machine. Algorithms 2023, 16, 430.

[31] 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.