IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 39, No. 4, August, 2022, pp. 1255-1265
Journal homepage: http://iieta.org/journals/ts

Reliable Scene Recognition Approach for Mobile Robots with Limited Resources Based on Deep Learning and Neuro-Fuzzy Inference

Aditya Singh[1,2], Padmakar Pandey[1*], Domenec Puig[2], Gora Chand Nandi[1], Mohammed Abdel-Nasser[3]

[1] Center of Intelligent Robotics, IIIT Allahabad, Prayagraj 211015, India
[2] Departament d'Enginyeria Informatica i Matemátiques, Universitat Rovira i Virgili, Tarragona 43007, Spain
[3] Department of Electrical Engineering, Aswan University, Aswan 81542, Egypt

Corresponding Author Email: rs175@iiita.ac.in

https://doi.org/10.18280/ts.390418

Received: 1 May 2022
Accepted: 26 July 2022

Keywords:
indoor scene recognition, deep learning, CNNs, neuro-fuzzy, transfer learning

ABSTRACT

Indoor scene recognition is complex due to the commonality shared between different spaces. Still, when it comes to robotics applications, the uncertainty increases due to illumination change, motion blur, interruption due to external light sources, and cluttered environments. Most existing fusion approaches do not consider the uncertainty, and others have a high computational cost that may not suit robots with limited resources. To mitigate these issues, this paper proposes a reliable indoor scene recognition approach for mobile robots with limited resources based on robust deep convolutional neural networks (CNNs) feature extractors and neuro-fuzzy inference to consider the uncertainty of the data. All CNN feature extractors are pre-trained on the Imagenet dataset and used in the manner of transfer learning. The performance of our fusion method has been assessed on a customized MIT-67 dataset and for real-time processing on a Locobot robot. We also compare the proposed method with two standard fusion methods---Early Feature Fusion (EFF) and Weighted Average Late Fusion (WALF). The experimental results demonstrate that our method achieves competitive results with a precision of 94%, and it performs well on the Locobot robot with a speed of 3.1 frames per second.

1. INTRODUCTION

Household robots are used for semantic-rich applications in a user-centric environment. Robots have been utilized in various applications, ranging from room cleaning [1] to elderly care [2, 3]. For such applications [4], scene recognition is a prime requirement for robot navigation and localization tasks. Scene recognition is important for providing high-level semantic information about a scene: It provides information about the robot's current locality and improves the quality of Human-Robot interaction [5, 6]. Semantic cues such as the presence of a specific object class, the layout of the contour, and the topological connection of spaces can all be used to classify a space. However, these systems have a restricted precision range. There is a huge requirement for correct recognition between similar traits [5] of two different spaces. According to Yan et al. [7], understanding spatial relations of objects is critical in many robotic applications such as grasping, manipulation, and obstacle avoidance. On the other hand, Humans can simply reason about an object's spatial relations from a glimpse of a scene based on prior knowledge of spatial constraints.

Several indoor scene recognition techniques including robots have been presented in recent years [5, 7-14]. RGB-D data is utilized to detect object orientation in many existing recognition methods, and it is incorporated with the definition of objects in the scene [15, 16]. The use of depth data always creates a dependence on a new hardware and depth sensor (infrared) suffers with distorted measurement due to rapid illuminance change and motion blur [17, 18]. Deep convolutional neural networks (CNNs) have been used for scene classification, achieving accuracy rates of 80.38 and 68.24 for 3 and 15 classes, respectively [19, 20]. The study [21] found that CNNs are biased toward identifying textures rather than shapes and that they perform badly when faced with a variety of visual distortions. As shown in Figure 1, indoor scene recognition involves several visual distortions like illumination changes [22], cluttered scenes, and similar traits [23] that impose a sort of uncertainty when classifying indoor scenes by a CNN model. Hence, the biased nature of CNNs and such sources of uncertainty significantly limit the accuracy of the existing indoor scene recognition method.

Noticeable efforts have been made in the literature to enhance the performance of indoor scene recognition methods. For instance, Glavan and Talavera [5] have used multi-modal learning on video data gathered from social media and employed different fusion strategies to aggregate models trained on multi-modal data. However, these methods obtained a limited accuracy of 70%. A multi model-fusion approach based on adaptive discriminative metric learning has been presented in the study of Wang et al. [24] for indoor scene recognition. This fusion approach achieved an accuracy of 88.43% on MIT-67 dataset. The studies [25-27] have applied different fusion approaches to deep CNNs features to recognize human activity in indoor scenes and to find semantic matching between two scenes. These fusion approaches adopt a weighted addition approach for fusing the features at different stages (i.e., joint fusion or late fusion). Although the weighted fusion approaches increase the probability score for detected classes, they fail to penalize the non-prior predictions,

which can give rise to uncertainty in predictions.

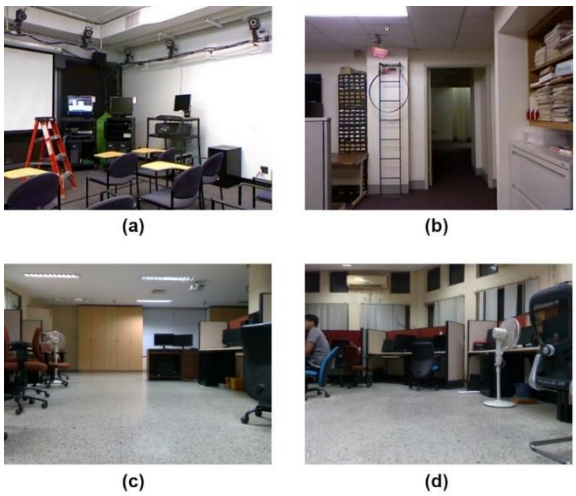


Figure 1. Visual distortions in indoor scene recognition framework. (a) and (b) are taken from an indoor scene recognition dataset. (c) and (d) are captured by a Locobot robot in a laboratory at IIIT-Allahabad

Indeed, most existing fusion approaches do not consider the uncertainty due to illumination changes, cluttered scenes, and similar traits. Besides, some models have a high computational cost that may not suit robots with limited resources. The main issue in most multi-model fusion approaches is the size of the network. The number of parameters will increase multi-fold by fusing multiple heavy networks, which creates a lag in applying on robot hardware due to the high computational cost.

To address the issues discussed previously, in this paper we propose a reliable indoor scene recognition approach for mobile robots with limited resources based on multiple lightweight deep CNN feature extractors and neuro-fuzzy inference. Figure 2 presents a schematic diagram of the proposed indoor scene recognition approach. The proposed approach is implemented and tested on Locobot, a limited resource robot, to recognize indoor scenes in IIIT-Allahabad. Locobot robot has an Intel NUC computer with 8 GB RAM capacity. In this approach, we employ transfer learning and different lightweight deep CNN feature extractors such as VGG-16 [28], ResNet50 [29], and EfficientNet-B3 [30] to extract robust features from the images of the indoor scenes. *These representative ConvNets focus on different aspects of accuracy, efficiency and scalability.* Transfer learning allows us to employ deep CNNs trained on a large dataset (e.g., ImageNet [31]) to extract robust and descriptive features from the indoor scene recognition dataset. Each deep CNN feature extractor helps extract robust and powerful features against one or multiple kinds of visual distortions in indoor scene recognition images.

In this paper, an efficient fusion mechanism is proposed to aggregate the individual feature extractors to consider the uncertainty of the data mentioned above. In particular, the predictions of the individual deep feature extractors are aggregated using the neuro-fuzzy inference technique and compared with weighted average fusion technique to obtain reliable scene recognition results. The proposed multi-model fusion approach provided an average indoor scene recognition rate of 3.1 frames per second, which is sufficient for the application of contour or space localization [32].

This study makes the following key contributions:

- Proposing a reliable indoor scene recognition approach for mobile robots with limited resources based on robust deep CNN feature extractors and neuro-fuzzy inference to consider the uncertainty of the data;
- Investigating the performance of various fusion mechanisms for indoor scene recognition approach;
- Providing comparisons with state-of-the-art methods and achieving a superior indoor scene recognition precision of 94%.
- Achieving an average indoor scene recognition rate of 3.1 frames per second in a limited resource robot---Locobot robot.

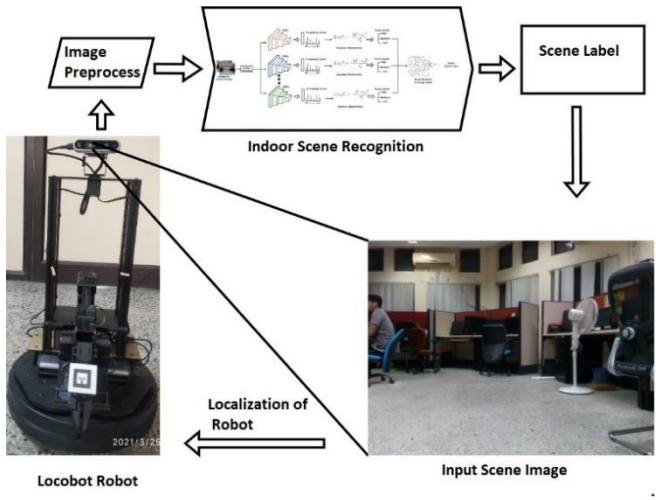


Figure 2. The schematic diagram for indoor scene recognition using a Locobot robot. Images were captured in a laboratory at IIIT-Allahabad

The remainder of this research is divided into four sections. The state-of-the-art approaches are discussed in Section 2. The proposed indoor scene recognition approach and fusion mechanisms are explained in Section 3. The experimental results are presented and discussed in Section 4. The study's findings and future work are presented in Section 5.

2. RELATED WORK

Several conventional image processing and machine learning techniques were proposed in the literature to handle scene recognition or classification problems. With the evidence of deep learning, using convolution networks is well seen in practice. This section presents and discusses several deep learning-based indoor scene classification techniques. These techniques include transfer learning methods, complex statistics-based classification models, and multi-model fusion methods. A detailed comparison of existing indoor scene classification techniques was presented in the study of Afif et al. [33], where the MIT-67 dataset was used to assess the performance of the compared models. Table 1 presents a summary of different techniques used for scene classification.

The authors of Hanni et al. [34] claimed that the indoor scene recognition [35] could not achieve high accuracy by using traditional neural networks, and therefore they proposed to use deep transfer learning techniques to enhance the accuracy. They proposed to use a deep learning-based model that consists of 3 inception layers, a mix of 1×1, 3×3, and 5×5 convolution layers, 3 max-pooling layers, and 10 mixed layers,

i.e., a model that fabricates a different neural network for each type of data provided to the neural network. The major operation of the input layer of this model is to resize the image to the optimum size for the pre-trained model. The mixed layer activity includes the extraction of various features for each type of image provided, and the max-pooling layer is used to realize the complex division of pixels into various segments. Following this, a pooling layer is used to retain the maximum pixel value. All the pixels in the input image were converted to small patches to obtain the maximum possible value of the pixel. A pre-trained Inceptionv3 [36] was used for classification.

To evaluate this method, an RGB Depth (RGB-D) images dataset had 3 indoor scene classes (Bedroom, Kitchen, Study), where each class had 150-200 images. The second dataset used was NYUv2 [37], where 6 indoor scene classes (Bedroom, Bathroom, Furniture Store, Cafe, Living Room, Kitchen) were considered. It is important to note that the model was trained and tested for each dataset individually, resulting in limited accuracy of 86%.

Besides, two problems, namely dataset bias in CNN and performing a combination of both scene and objects, were addressed by Herranz et al. [38]. Previously, the authors worked on a single Hybrid-CNN, which was trained on places and object classes of ImageNet simultaneously. However, it was observed that there was an induction of dataset bias when only one (Hybrid-CNN) model was used. Then, they proposed scale-specific CNN, which resulted in higher accuracy. The authors tested their model on three different datasets-15 scenes [39], MIT67 [40], and SUN397 [41]. Of these datasets, the SUN397 dataset was used as the target dataset for evaluating the dataset bias and achieved an accuracy of 85.81%. The study of Hanni et al. [42] focused on the scale ranges of the objects found in scenes and tuning CNNs to reduce indoor scene dataset bias.

Various methods are discussed for the application of scene recognition methods for the localization of robots [43]. It discusses a real-time scene recognition scheme to use objects segmented in it as the natural landmarks and explores the suitability of configured representation for automatic scene recognition in robot navigation. It discusses the uncertainty issue associated with the problem of scene recognition problem. Wang et al. [43] obtained a limited indoor scene classification accuracy of 90.1% on 8 indoor classes.

In an attempt to improve the indoor classification accuracy, various multi-model approaches were proposed [5, 24, 44], in which the input indoor scene image was fed into multiple CNN models, and a fusion technique (e.g., early or late fusion) was employed to produce the final prediction (i.e., indoor scene class). Miao et al. [45] proposed employing object information from the scene to enhance the prediction of the CNN model. In particular, they proposed an object-to-scene (OTS) module that extracts object features based on a segmentation network and an object feature aggregation module (OFAM). Afterward, the object relations are calculated, and the scene representation is constructed based on the proposed object attention module (OAM) and global relation aggregation module (GRAM). The study demonstrated that OTS successfully extracts object features and learns object relations from the segmentation network. They have tested it on 5 indoor classes (Bedroom, Corridor, Kitchen, Living Room, and Bathroom) of ICR5-23 dataset and claim an accuracy of 91.098%. The *main disadvantage* of this method is that it is limited to static images, and they give a limited performance on a real robot.

Three main issues [46] were discussed for scene recognition from a robot camera: 1) the indoor places include typical home objects, 2) a sequence of images instead of an isolated image is provided because the images are captured successively by a cleaning robot, and 3) the camera of the cleaning robot has a different view compared with those of cameras typically used by human beings. It points out an uncertain situation while applying CNN models to a real robot. Also, these models are heavy to be executed on limited-resource mobile robots.

Neuro-fuzzy [35] is proposed as a potential solution in case of uncertainty. It adjusts the relevance of predictions from different sources using fuzzy rules. It converts the inputs (predictions from different sources) into fuzzy values and uses a neural network for rule training.

In this paper, we propose an efficient fusion mechanism to aggregate the individual CNN feature extractors to consider the uncertainty of the data mentioned above. Specifically, we propose a neuro fuzzy-based fusion mechanism to fuzzify the predictions of robust CNN-based indoor scene classifiers. Besides, we compare the proposed fusion mechanism with early and late fusion methods.

Table 1. A summary of different scene classification approaches

Work	Dataset	Number of classes	Methods	Accuracy (%)
[47]	[39, 40]	4, 3	EfficientNet-B3	95, 97
[48]	[39]	15	CNN + SVM	86
[49]	NA	4	Inception v3	73.3
[50]	[40, 41]	10, 15	VGG-19 hybrid + VGG-11 P205+G P205	85.97, 70.69
[51]	[41]	19	RGB-D CNN	52.4
[52]	[37, 41]	10	Key local feature	55.9, 67.8
[38]	Custom Data	3	Multiple classifiers	80.38
[19]	[39]	15	Places- CNN features	68.24
[5]	InstaIndoor	18	Multi-model fusion	70
[42]	[40]	10	ADM learning	88.43

3. PROPOSED METHOD

In this section, we first define the problem of uncertainty in indoor scene classification. Then, we present the proposed neuro-fuzzy based fusion technique in detail, followed by a description of the employed CNN feature extractors. Ultimately, we explain the implementation of the proposed methods on a robot.

3.1 Problem definition

In any CNN-based classification approach, a RGB image is classified corresponding to a label using the predicted probability $Z_i = [z_1, z_2, \dots z_n]$ by a CNN model CNN_i, where 'n' is the number of class labels. Zi should fulfill the following condition:

$$\sum_{i=1}^{n} Z_i = 1 \quad (1)$$

Usually, a softmax function (P) is used to derive the final probability of class label using the prediction vector Z_i. The standard softmax is defined as follows:

$$P(Z)_i = \frac{\epsilon^{z_i}}{\sum_{j=1}^{n} \epsilon^{z_j}} \quad (2)$$

where, $'n'$ is the number of classes in multi-class classification.

In case of fusion of $'m'$ CNN models, weighted average is commonly used to aggregate the predicted probabilities $(P_{i_1}, P_{i_2}, .. P_{i_m})$ into a consensus prediction. However, the weighted average does not acknowledge the disparity in prediction by all models. The variation in probabilities creates uncertainty for normal fusion mechanisms.

To efficiently fuse the probabilities of the CNN models, we propose a neuro-fuzzy-based approach to aggregate them. The prediction probabilities of each model are first changed to the respective membership function (fuzzification) and then passed with a rule base to decide the final indoor scene class label. Notably, the fuzzy weights and rule base are trained in a supervised manner.

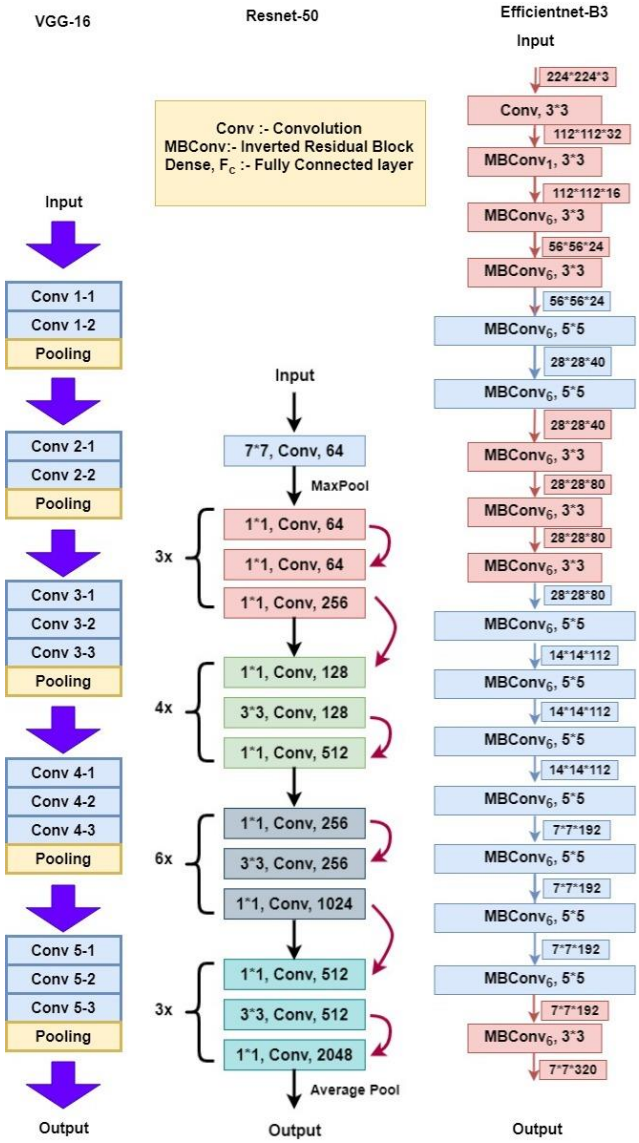


Figure 3. Individual CNN feature extractors. VGG-16, Resnet-50, Efficientnet- B3 have variety in mechanism

3.2 Individual deep CNNs feature extractors

To extract indoor scene relevant features from input images, we employ three pre-trained deep convolution feature extractors, namely VGG-16 [28], ResNet50 [29], and EfficientNet-B3 [30]. All pre-trained models were trained on ImageNet dataset [53]. It should be noted that the use of different robust CNN feature extractors can generate multiple feature representations, which can reflect different "views" of the data, meaning that such representations make a complete characterization of the input scene image. In Figure 3, individual CNN extractors are shown. Below, we present a brief description for each CNN feature extractor.

- VGG-16 is a deep CNN architecture that uses 13 convolutional layers and 3 fully connected layers. It uses the same size kernel of 3 × 3 to the whole network to extract features at a large scale. It uses a pooling of 2 × 2 with a stride of 2 for the entire network. The pre-trained VGG-16 model can be found at https://github.com/fchollet/deep-learning-models/blob/master/vgg16.py.
- ResNet50 uses skip connections as a unique concept to preserve the features extracted from previous network layers. It uses a 5 stage convolution process for feature extraction. The pre-trained ResNet50 model can be found at https://github.com/keras-team/keras-applications/blob/master/keras_applications/resnet50.py.
- EfficientNet-B3 is CNN architecture and scaling technique that uniformly scales all depth, width, and resolution dimensions utilizing a compound coefficient. It employs a stage-wise parallel network with varying sizes of convolution kernels, which makes it more efficient in extracting relevant features. The pre-trained EfficientNet-B3 model can be found at https://github.com/qubvel/efficientnet.

3.3 Proposed indoor scene recognition using neuro-fuzzy-based fusion

Figure 4 shows the proposed indoor scene recognition based on a neuro fuzzy network. The input indoor scene image is fed into a set of deep feature extractors $(CNN1, CNN2, \ldots, CNNn)$ to classify it as one of the n indoor scene classes. Given CNN_i, the probabilities of each indoor scene class $Pi = [p1, p2, \ldots, pn]$ are computed. The indoor scene probability vectors generated by all CNNs are inputted to the proposed neuro-fuzzy fusion mechanism to fuse them and predict the final indoor scene class label.

The fusion of output prediction vectors of different CNNs is a complex task due to each model's range and scale value variation. Each prediction vector P_i is computed after making a specific feature extraction mechanism by a CNN_i, which varies from one CNN to another due to the variation in the number of layers and way of processing. Most existing early and late fusion methods assume that the predictions from all CNN models have the same in nature or are identical, which is not practically correct. Considering this fact, we can conclude that handling uncertainty requires a more robust learning system that adapts to a complex environment and the fuzzy inference system, which disposes of a fuzzy inference system.

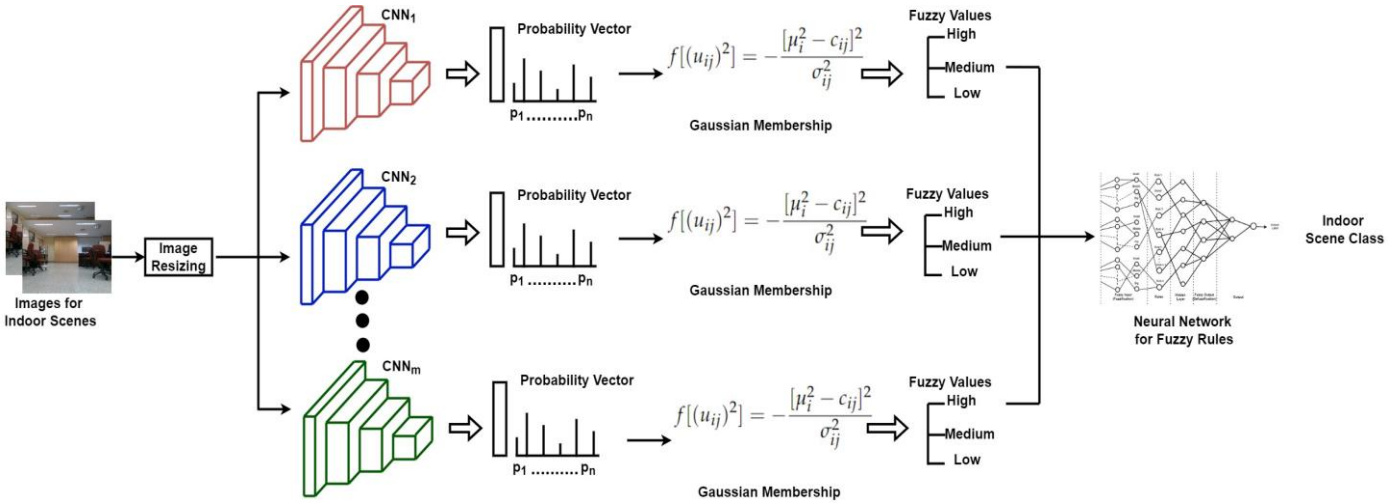


Figure 4. Schematic diagram of the proposed neuro-fuzzy based fusion mechanism

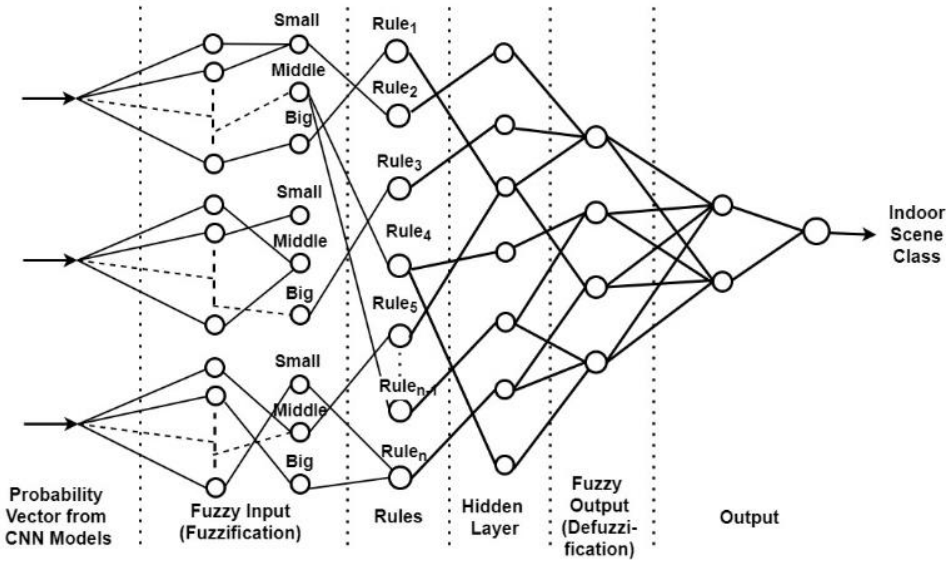


Figure 5. The architecture of the proposed neuro-fuzzy based fusion model

Here, we propose using a neuro-fuzzy system to find the best weights to fuse the predictions of different robust CNNs. As shown in Figure 5, the inputs to the neuro-fuzzy system are the predictions of the m CNN models (i.e., P_1, P_2,...., P_m). The architecture of the proposed neuro-fuzzy model contains six layers, including the input layer, fuzzification, rules, hidden layer, fuzzy output and output. The output from CNN models is analyzed through a gaussian distribution and converted to respective fuzzy values using the Gaussian membership function. The reason for selecting gaussian distribution is because of its smooth nature and concise notation. It can be defined as follows:

$$f\left[\left(u_{ij}\right)^2\right] = -\frac{[\mu_i^2 - c_{ij}]^2}{\sigma_{ij}^2} \tag{3}$$

where, c_{ij} and σ_{ij} are center and width of gaussian membership function. The error computation at output layer takes place using cross entropy loss and backpropagates the dewhere, c_{ij} and σ_{ij} are center and width of gaussian membership function. The error computation at output layer takes place using cross entropy loss and backpropagates the derivative δ_t,

$$\delta_t = -\frac{\delta D}{\delta a} \tag{4}$$

where, D is the difference (loss) between target and predicted values, and a is the weights of the last layer of neuro fuzzy network. The loss (D) is calculated as follows:

$$D = -\sum_{i=1}^{N} y_i \dot{log} \hat{y} \tag{5}$$

In the training phase of the fusion model, a hybrid optimizer, grid partition algorithm [54] with Adaptive Neuro-Fuzzy Inference System (ANFIS), are used. The hybrid optimizer uses both the steepest descent algorithm and the least-squares algorithm to fit the data faster than the traditional back propagation optimizer, which uses a least-square algorithm. The number of epochs used for training the fusion model is 150 (the model reaches an error range of 0.0005 before 100 epochs). Figure 6 presents the setting for training the fusion model. As shown, the number of inputs is 30 (10-elements of class probability vectors produced by each of the 3 CNN models) and one output label. The number of input membership functions is 5 for each input.

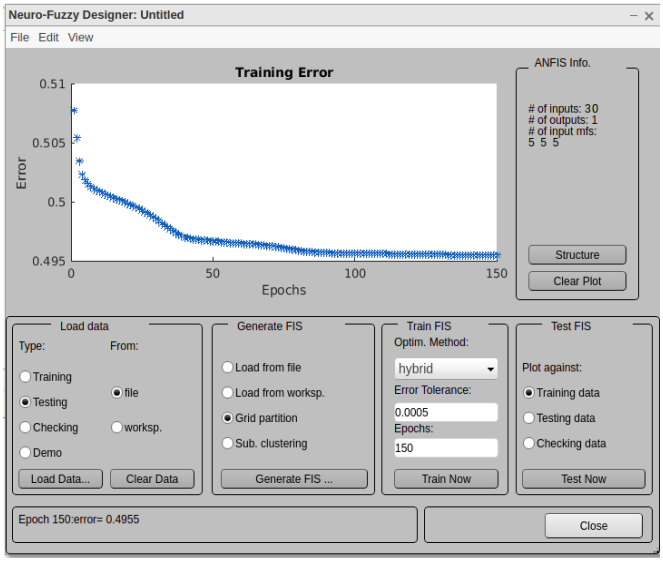


Figure 6. The setting for training the fusion model

3.4 Implementation of the proposed method

Algorithm 1 presents the steps to implement the proposed indoor scene classification method. Given that the number of individual feature extractors is 3. The input scene image I is fed into the three CNN models CNN_1, CNN_2, and CNN_3 to produce three class probability vectors P_1, P_2, and P_3. Then the neuro-fuzzy fusion-based model is trained.

In the neuro-fuzzy model, the class probability vectors are fuzzified using a Gaussian membership function as in Equation 3. The corresponding fuzzy values are then mapped against the final output using fuzzy rules, which are trained using a neural network. The output of the neural network is defuzzified to a crisp value as final output (in our case, scene label).

3.5 Application on robot

The validity of the proposed method is tested on a Locobot robot. The robot is configured with an Intel NUC - NUC8i3BEH computer [55]. This computer works with a ROS operating system and continuously integrates the information from different sensors. Locobot robot has a RAM capacity of 8 GB, SSD capacity of 250 GB and is equipped with an 8th Generation Intel Core i3 8109U processor. Locobot robot has an Intel Realsense [56] camera for capturing the indoor environment images. We have discussed the specification of the robot environment in Table 2. The Robot Operating System is used as a middleware for running various sensors and processes.

Table 2. Specification of robot hardware and software

Parameter	Value
RAM	8 GB
Operating System	Ubuntu 18.04
Middleware	ROS Kinetic
Processor	Intel Core 248 i3 8109U
Process used simultaneously	Odometer, ORB-SLAM

The proposed model is built on the Tensorflow framework using Keras API. The model is trained on an external machine (GPU) and then the weights configuration is saved for the final model. A protobuf file is used for the deployment of the model on the robot computer. A docker configuration is created for the required configuration to run the model on any other machine. The docker's compatibility in the ROS environment is a considerable issue while deployment.

4. EXPERIMENTS AND RESULTS

Here, we introduce the customized dataset we prepared for our application and evaluation metrics. Then, we analyze the performance of the individual CNN models with the indoor scene recognition, the performance of the proposed neuro-fuzzy fusion method and compare it with two well-known fusion techniques.

Algorithm 1 Algorithm for Proposed Late Fusion with Neuro Fuzzy Method

Input: Preprocessed image I of size 224 × 224 × 3
Output: Indoor scene class
Step:1 Load pretrained $CNN_1, CNN_2, \ldots, CNN_m$
Step:2 Compute the probability vectors for I: $P_1, P_2, \ldots, P_m$.

$$P_1 \leftarrow CNN_1(I)$$
$$P_2 \leftarrow CNN_2(I)$$
$$P_m \leftarrow CNN_m(I)$$

Step:3 Fuzzify P_1, P_2, P_3. *using Gaussian membership function (Eq. (3))*
Step:4 Train a Neural network for fuzzy rules for classification
Step:5 Defuzzify inference output to a crisp value
Step:6 Calculate the loss (Eq. (5)) and backpropagate it

4.1 Customized indoor scene classification dataset

It should be noted that the indoor classes that LocoRobot will recognize in our environment are limited. To train the proposed indoor scene recognition model, we used the MIT-67 dataset [22] to create a customized dataset. The original MIT-67 database contains 67 indoor categories and 15,620 images of varying sizes. We selected 10 classes of 67 indoor categories: Kitchen, Living Room, Bedroom, Airport Inside, Casino, Warehouse, Bakery, Book Store, Toy Store, and Bathroom. The selection of classes is based on the commonness of the scenes and more probability to appear in the daily life of our environment. Our customized dataset has 5,059 images representing the ten classes. We split the dataset as follows: 80% for training and 20% for testing.

4.2 Evaluation metrics

The performance of the proposed indoor scene recognition method is assessed in terms of Precision, Recall, and F1-score, which are formulated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$F1_{score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

where, TP, FP, TN and FN stand for True Positive, False Positive, True Negative, and False Negative, respectively.

4.3 Performance analysis

First, we assess the performance of the individual CNN models for indoor scene classification. As shown in Table 3, the three CNN models obtain more than 80% precision. The EfficientNet-B3 model achieves the best recall and F1-score values (87%). EfficientNet-B3 competes in terms of performance thanks to the model scaling mechanism that prevents unwanted information loss, which helps in improving accuracy. Figure 7 presents the confusion matrix of VGG-16, ResNet-50 and EfficientNet- B3 to visualize their class-wise performance. As one can see, in the case of class 3, VGG-16 and ResNet-50 wrongly classified 23 samples as class 7; in turn, EfficientNet-B3 produced a lower misclassification rate, where it wrongly classified 22 samples as class 7. Besides, in the case of class 7, VGG-16 mis-classified 15 samples as class 3, ResNet-50 misclassified 17 samples as class 3, and EfficientNet-B3 misclassified 12 samples as class 3. The main confusion is in class 3 'Airport Inside' and class 7 'Book Store' because their images have the same kind of furnished setup and illumination level. It usually confuses the CNN network, and it behaves uncertainly.

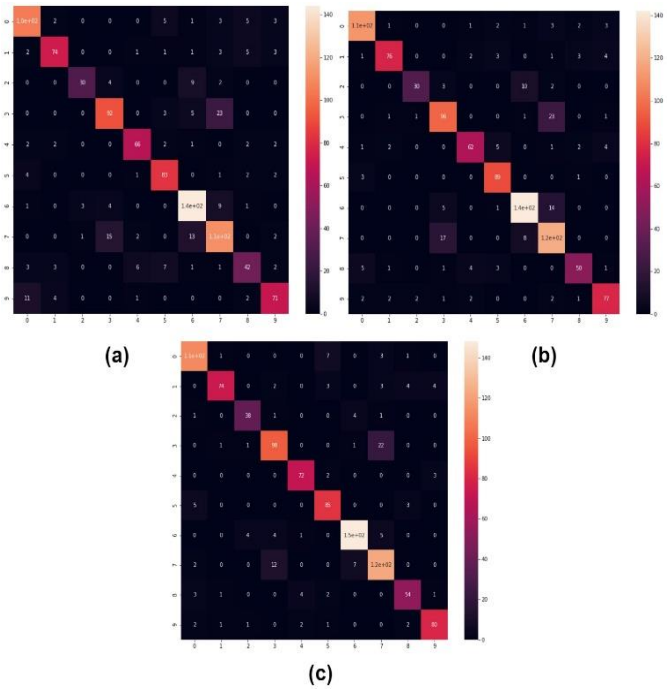


Figure 7. Plots of the confusion matrix of the individual CNN models. (a) VGG-16, (b) ResNet-50 and (c) EfficientNet-B3

Table 3. Performance analysis of individual deep CNN for indoor scene recognition

Model	Precision	Recall	F1-Score
VGG-16	81	79	80
ResNet-50	85	83	84
EfficientNet-B3	88	87	87

Table 4 presents the classification results of the proposed neuro-fuzzy fusion method for indoor scene classification. It achieved a precision of 94%, a recall of 94%, and an F1-score of 93%. The proposed neuro-fuzzy fusion method significantly enhances the results of the individual CNN models tabulated in Table 3. As one can see, our method increments the precision, recall and F1-score rates by more than 5%.

Table 4. Classification results for fusion models. Early fusion mechanism, late fusion with weighted average and neuro fuzzy (proposed model)

Model	Precision	Recall	F1-Score
Early Fusion Method	90	89	90
Weighted Average	92	91	91
Neuro Fuzzy	94	93	93

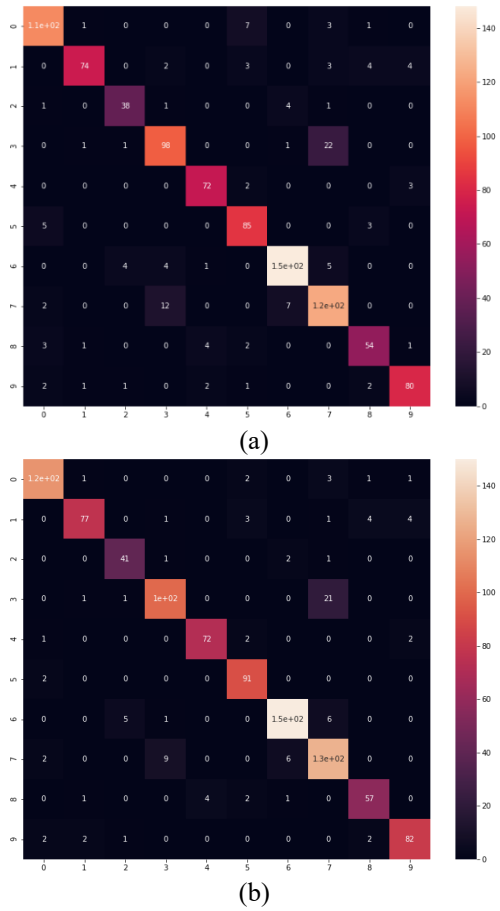
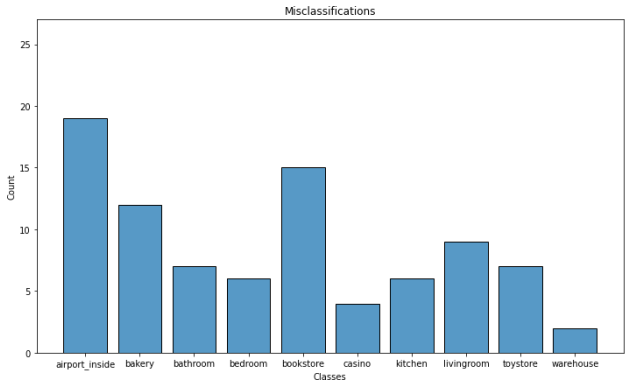


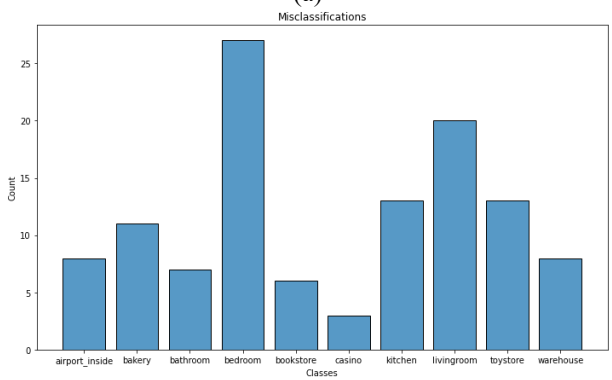
Figure 8. Plots of the confusion matrix of (a) the proposed neuro-fuzzy fusion, and (b) EFF method

Furthermore, we compare the proposed neuro-fuzzy fusion method against two different fusion algorithms, namely early feature fusion (EFF) and weighted average-based late fusion (WALF). In EFF, the output feature spaces from VGG-16, ResNet-50 and EfficientNet-B3 are resized and reduced to identical size by using principal component Analysis (PCA). PCA also ensures the selection of most relevant features. The final feature vector size of F_1, F_2, and F_3 from all three models is 256 × 1. Feature vectors (F_1, F_2, and F_3) are then fused into a final feature vector F using vector addition (F = F1 $\oplus$ F2 $\oplus$ F3). Finally, the fused feature vector (256 × 1) is used for indoor scene classification. In our experiments, a fully connected network is used for indoor scene classification. A softmax function is used to convert final vector output to a probability function. In the case of WALF, VGG-16, ResNet-50 and EfficientNet-B3 are used for predicting class probability vector (P_1, P_2, P_3), where the length of each vector equals the number of classes (10 × 1). The weighted-average technique is used to fuse the three class probability vectors into

one probability vector as follows $W = a_1\dot{P}_1 + a_2\dot{P}_2 + a_3\dot{P}_3$, where $a_1, a_2, and\ a_3$ are the weights of WALF with the constraint of $a_1 + a_2 + a_3 = 1$ It should be noted that $a_1, a_2, and\ a_3$ were empirically set to 0.27, 0.25, and 0.48 for VGG-16, Resnet-50 and EfficientNet-B3, respectively. As presented in Table 4, the proposed neuro-fuzzy fusion method achieves a precision score 4% higher than EFF and 2% higher than WALF. EFF increments the precision score of the individual feature extractors with 2%. The WALF obtains classification results better than the early fusion method with increments of 2% on all evaluation metrics. EFF applies feature fusion after the convolution blocks and feature reduction; in turn, WALF employs the whole CNN network and fuses the probability vectors of all networks, which has undoubtedly increased the precision. Figure 8 shows the confusion matrix of the proposed neuro-fuzzy fusion method and EFF to visualize their class-wise performance. We can see in Figure 8(a), in the case of class 3, our method wrongly classified 17 samples as class 3, while EFF mis-classified 18 samples (Figure 8(b)). The misclassification of class 6 to class 2 is also decreased. Figure 9 shows the distribution of misclassification of all indoor scene classes with the proposed method and EFF. As shown in Figure 9 (b), the highest EFF class misclassification probability happens with the 'bedroom' class. The neuro-fuzzy fusion method could reduce such misclassification bias toward one class. The proposed fusion method employs neuro-fuzzy to fuse the indoor scene probability vectors, and to handle uncertainty of data that originates due to illumination changes, cluttered scenes, and similar traits. As we can see in Table 4 and Figures 9(a) and 9(b), the proposed method outperforms the other fusion methods, which proves its application in handling data uncertainty.



(a)



(b)

Figure 9. The distribution of mis-classification of all indoor scene classes. (a) the proposed neuro-fuzzy fusion, and (b) EFF method

4.4 Performance on Locobot robot

Here, we present the performance of the proposed neuro-fuzzy fusion method on the Locobot robot in IIIT-Allahabad. The Robot works in a ROS environment and also uses multi-sensor information. Figure 10 shows a Live feed indoor scene classification result from the Locobot robot. Table 5 shows the detection speed of the proposed method on Locobot robot. As demonstrated above, the precision of the proposed method is higher than 94%. As one can see, the proposed method achieved 3.1 FPS (frame per second), which is sufficient for our indoor scene classification application and comparable to the ones of EFF and WALF.

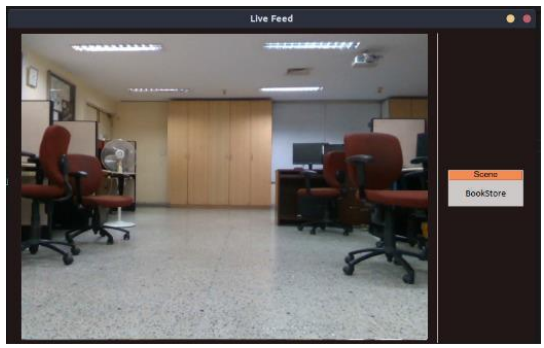


Figure 10. Live feed indoor scene classification results from Locobot robot based on the proposed neuro-fuzzy fusion method

Table 5. Performance on Locobot robot

Method	FPS
EFF	3.4
WALF	3.2
Proposed	3.1

4.5 Discussion

It is worth noting that there are alternative scene recognition approaches using robots, for instance, the cloud-based approach, where the deep learning model is hosted in a cloud server, and the robot sends the indoor scene image to it to analyse it. Then, the indoor scene class label is sent to the robot. Such approaches generally face three significant issues: peed, connectivity, and isolation. The speed is a considerable concern in robotics because it requires real-time processing with zero latency. Still, cloud processing sometimes assigns a queue to a process real-time data association gets affected.

Secondly, the connectivity issue happens due to a sudden loss of connection. In the case of real-time processing, it involves a lot, and it makes the robot an idle senseless machine. The connectivity issue is being solved in various ways, either by using a good 5G connection or Wi-Fi but in a limited scope.

The point of isolation comes when one does not want to connect the robot with the internet due to security and personal reasons. This issue matters in the case of many household robots and creates hurdles for cloud usage in robotics in an ethical manner.

5. CONCLUSION

This paper has proposed an efficient indoor scene recognition method based on deep CNNs and a neuro-fuzzy

fusion technique for a mobile robot with limited resources so-called LocoRobot. The proposed method has been compared with early feature fusion (EFF) and weighted average late fusion (WALF) methods to demonstrate its efficacy. The proposed method, EFF, and WALF have been tested on a Locobot robot in IIIT Allahabad. The proposed method has outperformed EFF and WALF with a precision of 94% with a speed of 3.1 FPS. It could reduce such misclassification bias toward particular classes, proving its application in handling data uncertainty from different sources like visual distortions in indoor images due to motion-related fluctuations, illumination changes, cluttered scenes, and similar traits. Such efficient indoor scene recognition effectively helps mobile robots with limited resources like LocoRobot to localize themselves in an unknown environment. In future work, we will optimize the individual CNN models and the proposed fusion method to further enhance the FPS rate with different mobile robots with limited resources. We will validate our method on more specific datasets for different use cases.

REFERENCES

[1] Suzuki, T., Sakata, Y., Yamaguchi, A. (2021). Toilet cleaning service by mobile robot equipped with RGB-D camera and single arm. In 2021 IEEE/SICE International Symposium on System Integration (SII), pp. 710-711. https://doi.org/10.1109/IEEECONF49454.2021.9382714

[2] Bardaro, G., Antonini, A., Motta, E. (2022). Robots for elderly care in the home: A landscape analysis and co-design toolkit. International Journal of Social Robotics, 14(3): 657-681. https://doi.org/10.1007/s12369-021-00816-3

[3] Kyrarini, M., Lygerakis, F., Rajavenkatanarayanan, A., Sevastopoulos, C., Nambiappan, H.R., Chaitanya, K.K., Makedon, F. (2021). A survey of robots in healthcare. Technologies, 9(1): 8. https://doi.org/10.3390/technologies9010008

[4] Singh, A., Pandey, P., Nandi, G.C. (2021). Influence of human mindset and societal structure in the spread of technology for Service Robots. In 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), pp. 1-6. https://doi.org/10.1109/UPCON52273.2021.9667652

[5] Glavan, A., Talavera, E. (2022). InstaIndoor and multi-modal deep learning for indoor scene recognition. Neural Computing and Applications, 34(9): 6861-6877. https://doi.org/10.1007/s00521-021-06781-2

[6] Labinghisa, B., Lee, D.M. (2021). A deep learning based scene recognition algorithm for indoor localization. In 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 167-170. https://doi.org/10.1109/ICAIIC51459.2021.9415278

[7] Yan, F., Wang, D., He, H. (2020). Robotic understanding of spatial relationships using neural-logic learning. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 8358-8365. https://doi.org/10.1109/IROS45743.2020.9340917

[8] Shah, A.A., Rana, K. (2019). A review on computer vision-scene classification techniques. In 2019 Third International Conference on Inventive Systems and Control (ICISC), pp. 558-566. https://doi.org/10.1109/ICISC44355.2019.9036472

[9] Liu, S., Tian, G. (2019). An indoor scene classification method for service robot Based on CNN feature. Journal of Robotics, 2019: Article ID 8591035. https://doi.org/10.1155/2019/8591035

[10] Li, Y., Zhang, J., Cheng, Y., Huang, K., Tan, T. (2018). Df^2 net: Discriminative feature learning and fusion network for RGB-d indoor scene classification. In Proceedings of the AAAI Conference on Artificial Intelligence, 32(1): https://doi.org/10.1609/aaai.v32i1.12292

[11] Pereira, R., Gonçalves, N., Garrote, L., Barros, T., Lopes, A., Nunes, U.J. (2020). Deep-learning based global and semantic feature fusion for indoor scene classification. In 2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), pp. 67-73. https://doi.org/10.1109/ICARSC49921.2020.9096068

[12] Sorkhi, A.G., Hassanpour, H., Fateh, M. (2020). A comprehensive system for image scene classification. Multimedia Tools and Applications, 79(25): 18033-18058. https://doi.org/10.1007/s11042-019-08264-y

[13] Aziz, S., Awais, M., Akram, T., Khan, U., Alhussein, M., Aurangzeb, K. (2019). Automatic scene recognition through acoustic classification for behavioral robotics. Electronics, 8(5): 483. https://doi.org/10.3390/electronics8050483

[14] Cebollada, S., Payá, L., Flores, M., Peidró, A., Reinoso, O. (2021). A state-of-the-art review on mobile robotics tasks using artificial intelligence and visual data. Expert Systems with Applications, 167: 114195. https://doi.org/10.1016/j.eswa.2020.114195

[15] Ismail, A.S., Seifelnasr, M.M., Guo, H. (2018). Understanding indoor scene: Spatial layout estimation, scene classification, and object detection. In Proceedings of the 3rd International Conference on Multimedia Systems and Signal Processing, pp. 64-70. https://doi.org/10.1145/3220162.3220182

[16] Song, X., Jiang, S., Herranz, L., Chen, C. (2018). Learning effective RGB-D representations for scene recognition. IEEE Transactions on Image Processing, 28(2): 980-993. https://doi.org/10.1109/TIP.2018.2872629

[17] Satish, P., Srikantaswamy, M., Ramaswamy, N.K. (2020). A comprehensive review of blind deconvolution techniques for image deblurring. Traitement du Signal, 37(3): 527-539. https://doi.org/10.18280/ts.370321

[18] Kubota, Y., Hayakawa, T., Ke, Y., Moko, Y., Ishikawa, M. (2020). High-speed motion blur compensation system in infrared region using galvanometer mirror and thermography camera. In Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems, 11379: 154-160. https://doi.org/10.1117/12.2558450

[19] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A. (2014). Learning deep features for scene recognition using places database. Advances in neural information processing systems, 27.

[20] Hernandez, A.C., Gomez, C., Derner, E., Barber, R. (2019). Indoor scene recognition based on weighted voting schemes. In 2019 European Conference on Mobile Robots (ECMR), pp. 1-6. https://doi.org/10.1109/ECMR.2019.8870931

[21] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W. (2018). ImageNet-trained

CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231. https://doi.org/10.48550/arXiv.1811.12231

[22] Mao, Y., Li, Y. (2022). PNN for indoor and outdoor scene recognition. In 2022 14th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), pp. 392-396. https://doi.org/10.1109/ICMTMA54903.2022.00082

[23] Khan, S.H., Hayat, M., Bennamoun, M., Togneri, R., Sohel, F.A. (2016). A discriminative representation of convolutional features for indoor scene recognition. IEEE Transactions on Image Processing, 25(7): 3372-3383. https://doi.org/10.1109/TIP.2016.2567076

[24] Wang, C., Peng, G., De Baets, B. (2020). Deep feature fusion through adaptive discriminative metric learning for scene recognition. Information Fusion, 63: 1-12. https://doi.org/10.1016/j.inffus.2020.05.005

[25] Zhang, P., Huang, X., Wang, Y., Jiang, C., He, S., Wang, H. (2021). Semantic similarity computing model based on multi model fine-grained nonlinear fusion. IEEE Access, 9: 8433-8443. https://doi.org/10.1109/ACCESS.2021.3049378

[26] Bi, Z., Huang, W. (2021). Human action identification by a quality-guided fusion of multi-model feature. Future Generation Computer Systems, 116: 13-21. https://doi.org/10.1016/j.future.2020.10.011

[27] Verma, K.K., Singh, B.M. (2021). Deep multi-model fusion for human activity recognition using evolutionary algorithms. International Journal of Interactive Multimedia and Artificial Intelligence, 7(2). https://doi.org/10.9781/ijimai.2021.08.008

[28] Tammina, S. (2019). Transfer learning using vgg-16 with deep convolutional neural network for classifying images. International Journal of Scientific and Research Publications (IJSRP), 9(10): 143-150. http://dx.doi.org/10.29322/IJSRP.9.10.2019.p9420

[29] Wen, L., Li, X., Gao, L. (2020). A transfer convolutional neural network for fault diagnosis based on ResNet-50. Neural Computing and Applications, 32(10): 6111-6124. https://doi.org/10.1007/s00521-019-04097-w

[30] Tan, M., Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning, pp. 6105-6114.

[31] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248-255. https://doi.org/10.1109/CVPR.2009.5206848

[32] Yu, H., Liang, L., Shi, P., Jiang, Q. (2021). A Direct approach of path planning using environmental contours. Journal of Intelligent & Robotic Systems, 101(1): 1-14. https://doi.org/10.1007/s10846-020-01271-4

[33] Afif, M., Ayachi, R., Said, Y., Atri, M. (2020). Deep learning based application for indoor scene recognition. Neural Processing Letters, 51(3): 2827-2837. https://doi.org/10.1007/s11063-020-10231-w

[34] Hanni, A., Chickerur, S., Bidari, I. (2017). Deep learning framework for scene based indoor location recognition. In 2017 International Conference on Technological Advancements in Power and Energy (TAP Energy), pp. 1-8. https://doi.org/10.1109/TAPENERGY.2017.8397254

[35] Katasev, A.S. (2019). Neuro-fuzzy model of fuzzy rules formation for objects state evaluation in conditions of uncertainty. Computer Research and Modeling, 11(3): 477-492. https://doi.org/10.20537/2076-7633-2019-11-3-477-492

[36] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-First AAAI Conference on Artificial Intelligence.

[37] Jiao, J., Wei, Y., Jie, Z., Shi, H., Lau, R.W., Huang, T.S. (2019). Geometry-aware distillation for indoor semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2869-2878.

[38] Herranz, L., Jiang, S., Li, X. (2016). Scene recognition with CNNs: Objects, scales and dataset bias. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 571-579.

[39] Sadeghi, F., Tappen, M.F. (2012). Latent pyramidal regions for recognizing scenes. In European Conference on Computer Vision, pp. 228-241. https://doi.org/10.1007/978-3-642-33715-4_17

[40] Afif, M., Ayachi, R., Said, Y., Atri, M. (2020). Deep learning based application for indoor scene recognition. Neural Processing Letters, 51(3): 2827-2837. https://doi.org/10.1007/s11063-020-10231-w

[41] Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3485-3492. https://doi.org/10.1109/CVPR.2010.5539970

[42] Hanni, A., Chickerur, S., Bidari, I. (2017). Deep learning framework for scene based indoor location recognition. In 2017 International Conference on Technological Advancements in Power and Energy (TAP Energy), pp. 1-8. https://doi.org/10.1109/TAPENERGY.2017.8397254

[43] Wang, X., Wang, X., Wilkes, D.M. (2020). An Automatic scene recognition using TD-learning for mobile robot localization in an outdoor environment. In Machine Learning-Based Natural Scene Recognition for Mobile Robot Localization in an Unknown Environment, 293-310. https://doi.org/10.1007/978-981-13-9217-7_15

[44] Akilan, T., Wu, Q.J., Yang, Y., Safaei, A. (2017). Fusion of transfer learning features and its application in image classification. In 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 1-5. https://doi.org/10.1109/CCECE.2017.7946733

[45] Miao, B., Zhou, L., Mian, A.S., Lam, T. L., Xu, Y. (2021). Object-to-scene: Learning to transfer object knowledge to indoor scene recognition. In 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2069-2075. https://doi.org/10.1109/IROS51168.2021.9636700

[46] Choe, S., Seong, H., Kim, E. (2021). Indoor place category recognition for a cleaning robot by fusing a probabilistic approach and deep learning. IEEE Transactions on Cybernetics, 52(8): 7265-7276. https://doi.org/10.1109/TCYB.2021.3052499

[47] Afif, M., Ayachi, R., Said, Y., Atri, M. (2020). Deep learning based application for indoor scene recognition. Neural Processing Letters, 51(3): 2827-2837. https://doi.org/10.1007/s11063-020-10231-w

[48] Yuan, Y., Mou, L., Lu, X. (2015). Scene recognition by manifold regularized deep learning architecture. IEEE Transactions on Neural Networks and Learning Systems, 26(10): 2222-2233. https://doi.org/10.1109/TNNLS.2014.2359471

[49] Guo, W., Wu, R., Chen, Y., Zhu, X. (2018). Deep learning scene recognition method based on localization enhancement. Sensors, 18(10): 3376. https://doi.org/10.3390/s18103376

[50] Xie, G.S., Zhang, X.Y., Yan, S., Liu, C.L. (2015). Hybrid CNN and dictionary-based models for scene recognition and domain adaptation. IEEE Transactions on Circuits and Systems for Video Technology, 27(6): 1263-1274. https://doi.org/10.1109/TCSVT.2015.2511543

[51] Song, X., Herranz, L., Jiang, S. (2017). Depth CNNs for RGB-D scene recognition: Learning from scratch better than transferring from RGB-CNNs. In Thirty-First AAAI Conference on Artificial Intelligence.

[52] Xiong, Z., Yuan, Y., Wang, Q. (2019). RGB-D scene recognition via spatial-related multi-modal feature learning. IEEE Access, 7: 106739-106747. https://doi.org/10.1109/ACCESS.2019.2932080

[53] Keras Applications. https://keras. io/api/applications/, accessed on 28 March 2022.

[54] Karaboga, D., Kaya, E. (2019). Adaptive network based fuzzy inference system (ANFIS) training approaches: A comprehensive survey. Artificial Intelligence Review, 52(4): 2263-2293. https://doi.org/10.1007/s10462-017-9610-2

[55] Intel Computer. https://www.intel.com/content/www/us/en/products/sku/126150/intel-nuc-kit-nuc8i3beh/specifications.html, accessed on 2 April 2022.

[56] Keselman, L., Iselin Woodfill, J., Grunnet-Jepsen, A., Bhowmik, A. (2017). Intel realsense stereoscopic depth cameras. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1-10.