A FUSION OF HAND-CRAFTED FEATURES AND DEEP NEURAL NETWORK FOR INDOOR SCENE CLASSIFICATION.

Basavaraj S. Anami[1] and Chetan V. Sagarnal[2*]

[1]Department of CSE, K. L. E. Institute of Technology, Hubballi, Karnataka,580030 India,

[2]Department of ECE, K. L. E. Institute of Technology, Hubballi, Karnataka,580030 India,

Email: anami_basu@hotmail.com[1], chetan.sagarnal@gmail.com[2*] (corresponding author)

DOI: https://doi.org/10.22452/mjcs.vol36no2.5

ABSTRACT

Convolutional neural networks (CNN) have proved to be the best choice left for image classification tasks. However, hand-crafted features cannot be ignored as these are the basic to conventional image processing. Hand-crafted features provide a priori information that often acts as the contemporary solution to CNN in image classification, and hence an attempt is made to fuse the two. This paper gives a feature fusion approach to combine CNN and hand-crafted features. The proposed methodology uses two stages, where the first stage comprises feature encoder that encodes non-normalized features of CNN, which utilizes edge, texture, and local features. The fusion of handcrafted features with CNN features is carried out in the second Hand-crafted crafted features are validated that helped CNN to perform better. Experimental results reveal that the proposed methodology improves over the original Efficient-Net(E) on the MIT-67 dataset and achieved an average accuracy of 93.87%. The results are compared with state-of-the-art methods.

Keywords: Deep Learning, Handcrafted Features, activation functions, Scene Classification.

1.0 INTRODUCTION

Scene recognition remains challenging in computer vision, but the literature reveals different research activities [1–9]. Representation and recognition of the scene is a problem to tackle in scene recognition. Scene representation uses unique features to make the scenes distinct, whereas scene classification takes laborious work in designing the methods to identify various scene categories. Compared to scene classification, representation degrades the performance of scene recognition systems as it uses generalized attributes of the same category of images. Specifically, these characteristics are complex to acquire features due to the spatial layout of scenes often using multiple distinct scene categories with varied objects such that extraction of features is a challenging task.

Inter-class margins and reduced intra-class variations deliver an inspiring problem in scene recognition. Many researchers have proposed, and these are classified into two categories: handcrafted features [2,3,10–13] and machine learning-based features [5,14,15]. Hand-crafted features(HCF) include generalized search trees (GIST) [16], oriented texture curves (OTC) [17], and census transform histograms (CENTRIST) [18], which take complex visual information such as structural and textural. However, these features are not adequate, hence, CNN features help to classify scenes as it uses semantic information.

The HCF features like Haar, HoG, LBP and SIFT were popularly used in image processing in the past, but these features have limited description capabilities and hence given low classification accuracies. In recent times, CNNs in computer vision proved their dominance as they learn by large dataset. However, these fail to have interpretability [1], indicating how features are identified through experience. But, few researchers have broken the myth of network interpretability [8] therefore an attempt is made to explore the influence of HCF with CNN features on the performance of CNN by way of fusing them and taking the study to comprehend the relationship between hand-crafted features and CNN [7-8].

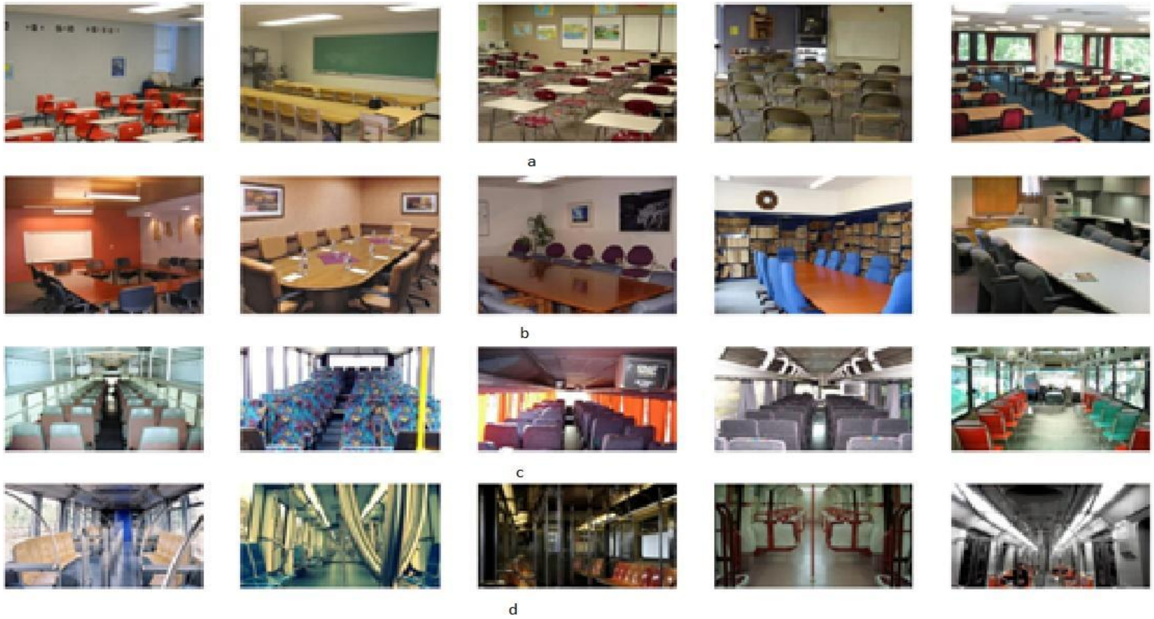


Fig.1: scene categories and sample images from the MIT Indoor-67 dataset. (a) classroom, (b) meeting room, (c) inside a bus and (d) inside a subway [18]

This paper presents a method combining traditional features with CNN's obtained features. It is also necessary to find the best-suited combination to train the architecture. An experiment to concatenate hand-crafted and CNN features prove to have a marginal improvement in the performance. It possesses an encoder that normalizes features in different forms and feature fusion. Experiments show that the proposed method improves classification accuracy for a given CNN. In contrast, the work brings insight into the relationship between hand-crafted features and the given CNN. The work uses the MIT-67 dataset, the combination of HOG, SIFT, and Efficient-Net for evaluation with different activation functions. The contributions of the proposed work are as under,

- A fusion model for combining deep learning and hand crafted features for scene classification.
- A concept of Horizontal Image Division into strips for extracting indoor scene information for restricted computation.
- Efficient use of both global and local features for better obtaining classification accuracy.

The paper is organized with section 2 covering the literature survey of the existing methods Section 3 explains the proposed method. Experimental results are presented and discussed in Section 4, followed by the conclusion given in Section 5.

2.0 RELATED WORK

Designing a simple and efficient model is a challenging task in classifying images. Hand-crafted features are how a human vision perceives an image considering different perspectives. Histograms return the color derivatives of the images. Histogram of Gradient (HoG) defines the shape of the objects, Scale Invariant Feature Transform (SIFT) used to provide local features of the images. Nowadays, CNN is the only choice in the computer vision community. AlexNet [2] achieves an accuracy of 83.6% and stands in one of the top players in Imgenet challenge. Architectures like VGG [3], GoogLeNet [4], and ResNet [5] expanded the depth and width of CNN. The features of CNN hold exhaustive information and learns by experience. Few researchers have explained network interpretability[6]. Some works compare SIFT and CNN features[7]. Use of low-level features help to afford complementary information for CNN.Neural network model using CNN as a feature detector and descriptor to replace SIFT is carried [9].

The NetVLAD architecture uses VGG16 at core, and a trainable and differentiable VLAD layer focuses on aggregating local descriptors. This model is tested on several location identification datasets. Several modifications, including

ESA-VLAD spatial pyramid-enhanced modifications and others, have been added on top of NetVLAD. The Patch-NetVLAD [22] is a remarkable network constructed on top of NetVLAD.

Training deep learning models to recognize various situations requires large-scale datasets. The ImageNet consists of 1000 classes of both objects and places, the SUN dataset [25] comprises of 899 different scene classes and a total of 130,519 images, the MSLS dataset [21] and Places365 dataset have 365 scene categories and covers a range of indoor, outdoor, and urban areas and the MIT-67 Indoor Scene Recognition dataset [23] has 67 classes and 15620 total images describing both public (stores Further extending some of the aforementioned datasets, study [24] demonstrates that hierarchical approaches and image-level descriptors can greatly increase the success rate.

Applying multimodal deep learning algorithms, contribute to the field of indoor scene detection in videos in this work. On the one hand, they used InstaIndoor, a dataset made up of 3,788 videos gathered from the Instagram network and describing nine various inside scenarios[26].

Bai et al. [27] trained a set of scene-specific object models for each scene category by alternating between searching over discriminative regions of images and training a support vector machine (SVM) based on region features for scene recognition.

Basavanna et al. [32] proposed work on scene text binarization using adaptive histogram analysis based on segmenting the words based on the region which works on multi oriented text lines.

Rassem et al. [33] proposed Completed Local Ternary Pattern (CLTP) texture descriptor which is insensitive to noise, the method used to classify texture based image classification considering scene, event and medical image dataset.

Yirui et al. [37-39] proposed edge computing based object detection for low light images which improves detection performance in mobile multimedia and low-light environments.

From the literature survey, It is observed that researchers have worked on scene classification with either handcrafted features or deep learning models. In this work, we first tried combining handcrafted features to achieve the accuracy of ANN. Later, handcrafted features and deep learning model features are combined to improve the accuracy. There are instances of HCF and DLMF being fused with CNN models but scope exists to try different HCF with state-of–the-art CNN models to achieve better classification and hence the work projected in this paper.

3.0 METHODOLOGY

The block diagram of the methodology is shown in Fig.1, which consists of four stages, namely, region extraction(RE), feature extraction(FE), fusion of features(FF) and classification.

Every indoor scene has information spread across the image, Vertically and Horizontally therefore identifying the regions contributing to the process of classification needs region extraction. The HCF for training machine learning (ML) model (ANN) needs extraction and a ML model is tested with different combinations of the extracted features. The best HCF combination is fused with CNN model features for improving the classification accuracy.

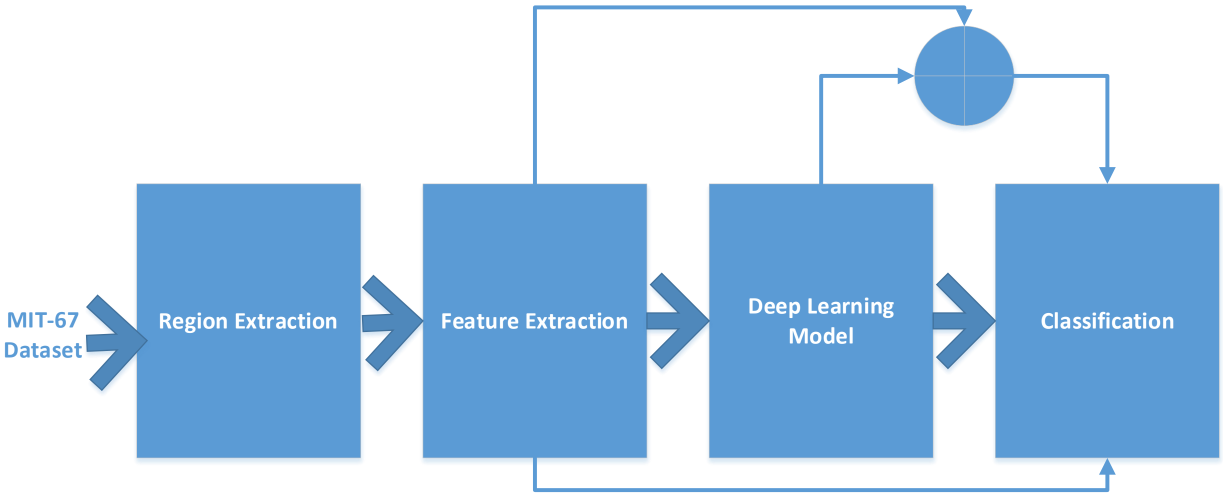


Fig 2. Block diagram of proposed work

3.1 Region Extraction

An indoor scene comprises a spatial layout and connectivity of the objects in a scene, which is observed in the image given in Fig.3. The image of a bedroom is divided initially into two halves and further divided into two more halves. We see more the connectedness of the objects in regions 3 and 4, which gives the maximum information of the scene. The other two regions (namely1 and 2) give information about the ceiling and very little about spatial layout. But more the complexity of spatial layout is seen in layer 3 and 4. Hence, in this work, we have divided images into 4-regions and computed hand-crafted features in each of the regions giving different information of the scene, which is detailed in section-4.

A set of patch descriptors are extracted from the interest points or from densely sampled regions, to summarize them into an image-level feature vector to classify the scene images into classes.Spatial Pyramid Matching (SPM) is different from in which each patch descriptor is associated with its spatial position. This approach augments the patch descriptor '**f**' by dividing the given image into four regions (also called layers) denoted as I_{11}, I_{12}, I_{13} , I_{14},where I denote an image which is divided into four regions, as shown in Fig 3 and 4.

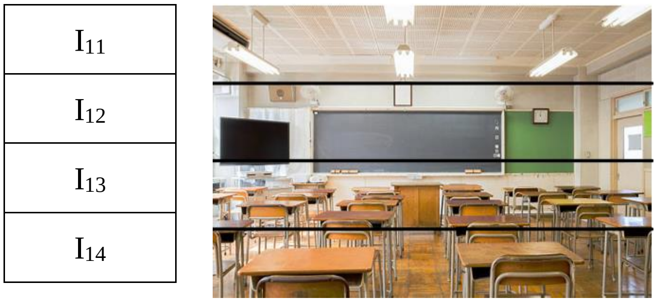


Fig.3 Image divided into number of regions

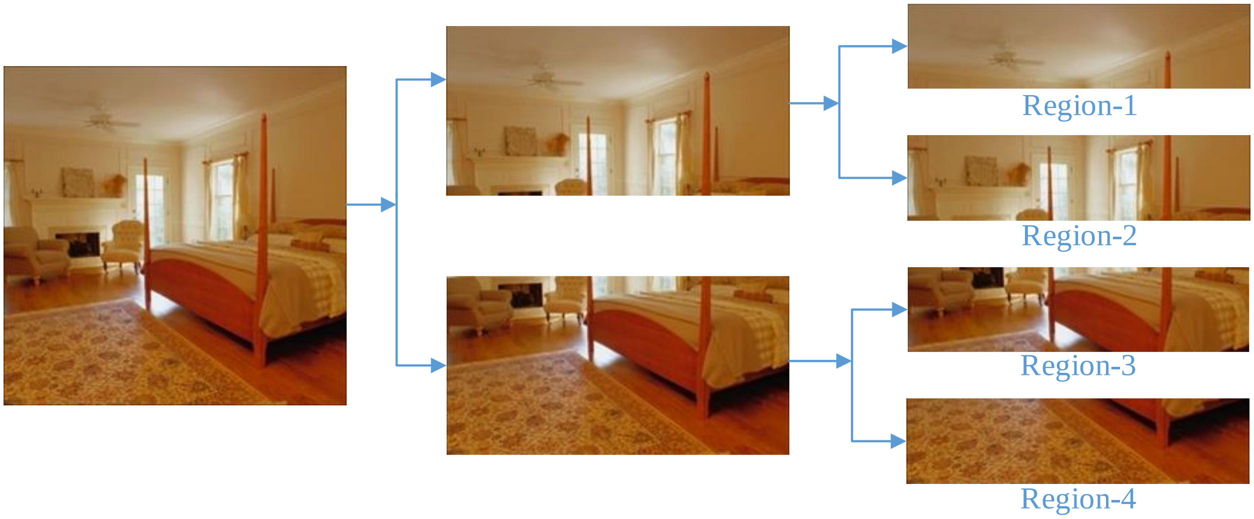


Fig.4 Indoor scene image and its information

Let **S={(f1, I_{11}), (f2,p2), (f3,p3) (f4,p4)}** denote the set of encoded features of each region and the tuple (fi, pi) represent ith feature set and region for i = 1 to 4**.** The indoor scene images consist of orientation features, these features are more discriminative and each region gives the orientational information of an image. In this work, we have considered the MIT-67 data set for experimentation. The modern descriptors SIFT,HoG, SURF, and GIST are used as features.

Consider an indoor scene as a classroom, which comprises of several objects such as Benches, Black Board, Window, Podium etc. In the human vision system, the scene is recognized by scanning the entire scene region by region in quick succession. Hence, the image is divided into four regions, orientational information preserved, a dense feature descriptor applied, concatenated with statistical properties and machine learning algorithms are used to classify the scenes.

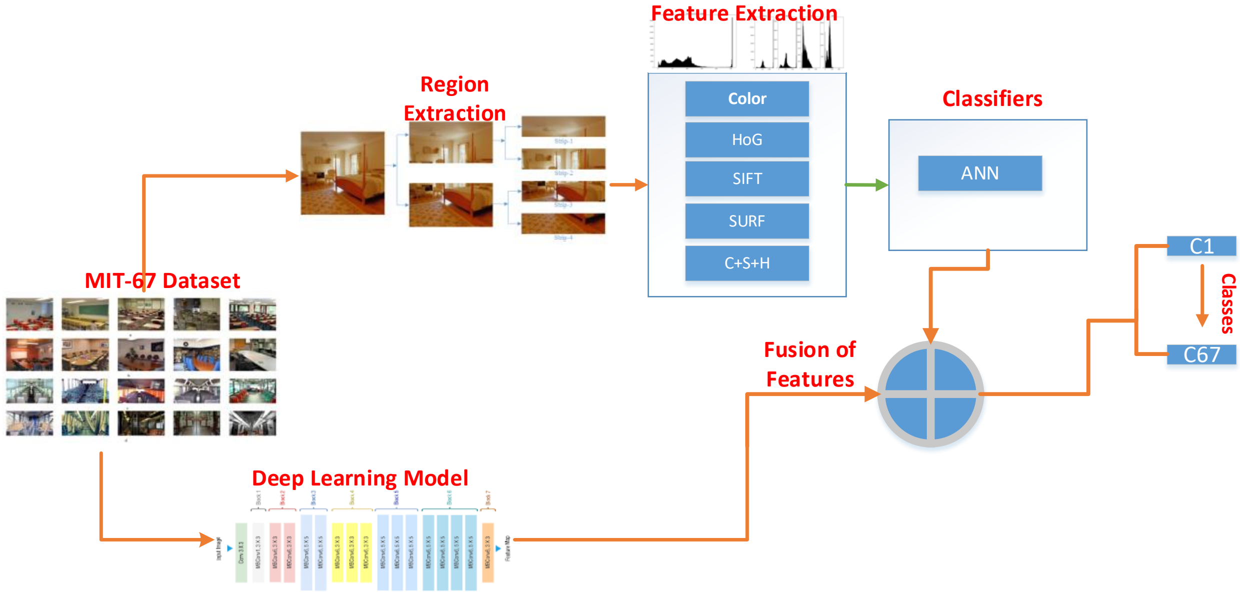


Fig.5 Combination of CNN and handcrafted features.

The considered SURF, HoG and their combination act as feature vectors, and used to train the classifiers SVM and Neural network. Initially, color quantization is applied, as the scene image comprises different colors, which gives the correct and relevant information of an image. The SURF, HoG, SURF+HoG is applied to each region. Mean, standard deviation, variance and kurtosis are computed for each region and concatenated to make a feature vector. As the feature vectors vary for each image, considering the statistical parameters for each region gives better classification results than state- state-of-the-art.

We have also trained the dataset using a combination of hand-crafted features and a deep learning model (Efficient-Net). The Fig.5 shows the block diagram of the proposed work, and at the fully connected layer, the fusion of hand-crafted and Efficient-Net features is made. The results obtained are discussed in section 4.
Efficient Net architecture comprises width,depth,resolution, compound scaling due to these characteristics of architecture features extracted and improvises the accuracy of model compared to other deep learning architectures (ResNet, VGG etc). Efficient Net models vary from B0-B7. In this work we have used B7 because the number of feature maps (subblocks) are more compared to B0-B6 models as shown in Fig.5.

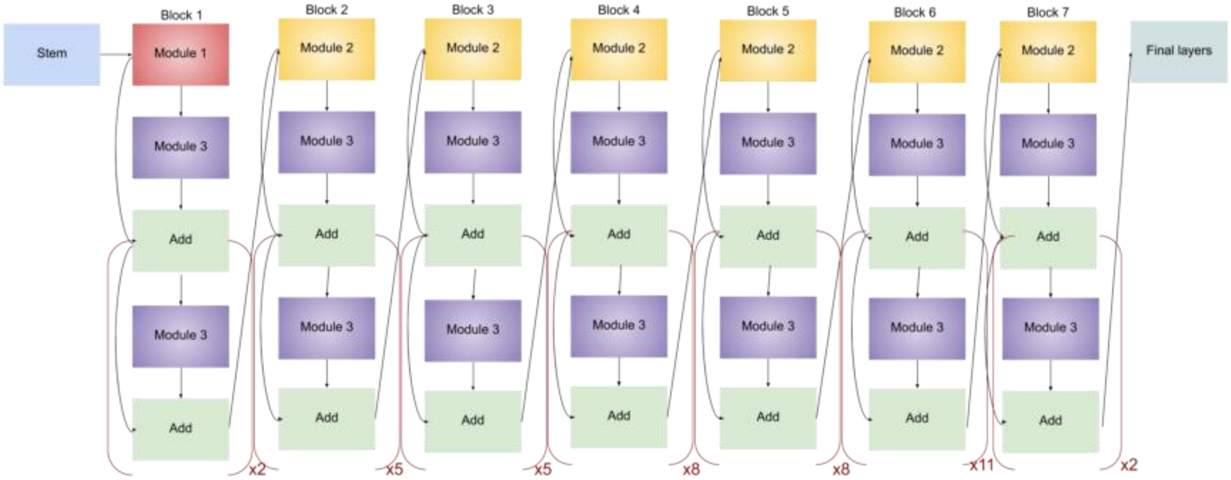


Fig. 6. Architecture of Efficient-Net B7

3.2 Feature Extraction

The spatial and orientational distributions of objects are observed in each image category.The images are divided into. The spatial and orientational histograms are calculated for four regions from the corresponding categories. The appearance and context histograms,128-dimensional SURF descriptor, and covariance plots are shown in Fig. 5(a)and(b). We have used a 576 long feature vector with l2-normalized HoG descriptors, as given by expression (1).

$$f = \frac{v}{\sqrt{v^2+e}} \quad \dots (1)$$

where 'v' is the HoG feature descriptor, which gives the frequency of orientation of an image gradients and 'e' is a constant.

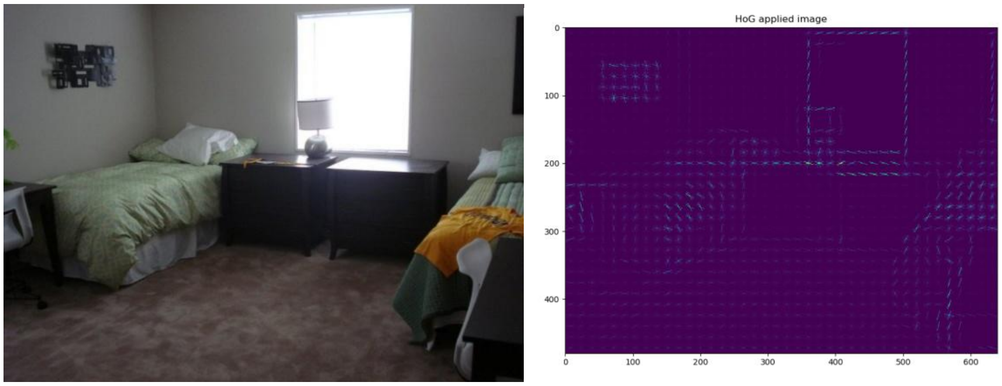


Fig. 7 Bedroom Scene Image and its Histogram of gradient after color quantization

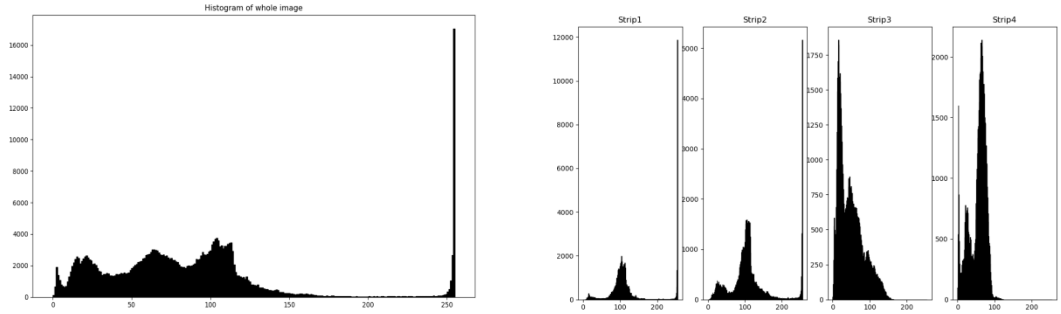


Fig. 7 (a) Histogram of Whole Image and Regions of Bedroom Scene Image

Fig. 7 (b) Image Regions of Bedroom Scene Image

Fig.7 gives the histogram of the whole image and of each of the regions. The Fig.7(a) shows that the information of the images varies with respect to the content of the regions. Hence, we extracted the features of the whole image and of each region to train the CNN.

3.3 Classification

In this stage, we have used the Efficientnet-B7 deep learning model features fused with hand-crafted features to classify the images into 67 classes. EfficientNet-B7 model architecture makes use of scaling which possess wide, deep and high resolution [18]. In turn, the paper uses the same model but still experiments with different activation functions to draw some inferences related to the performance to fine-tune the model parameters that may offer better performance. Local optimum and speed is taken care of by increasing the learning rate to 0.01. Relu, Sigmoid and Tanh activation functions are used in the work. The best classification results are obtained for HCF, namely combination of color and SURF extracted from regions 3 & 4 and fused with CNN model, namely EfficientNet . The classification accuracy obtained is 93.87%.

4.0 EXPERIMENTAL RESULTS

Tables 1 and 2 give different scene categories in the MIT-67 dataset. The dataset contains a total of 15620 images in 67 Indoor scene categories. The number of images varies across scene categories. However, at least 100 images per category is ensured. All images are in .jpg format. The images are used for research purposes only.With available research datasets, the selection is a challenge in the field of computer vision[11]. Even though the dataset seems to have similar visual categories, however, give different performances during the training of the classifier [10].

Table. 1: **MIT-67 Scene Category[18]**

Category	Category	Category	Category
AirportInside	Book Store	Cloister	Deli
Art Studio	Bowling	Closet	Dental Office
Auditorium	Buffet	Clothing Store	Dinning Room
Bakery	Casino	Computer Room	Elevator
Bar	Children Room	Concert Hall	Fastfood
Bathroom	Church Inside	Corridor	Florist
Bedroom	Classroom	Gameroom	Garrage
Green House	Grocery Store	Gym	Hair Salon
Hospital	Inside Bus	Subway	Jewellery shop
Kinder garden	Kitchen	Laboratory	Laundromat
Library	Living Room	Lobby	Locker Room

all	Meeting Room	Movie Theatre	Museum
Nursery	Office	Operating Room	Pantry
Pool Inside	Prison Cell	Restaurant	Rest_kitchen
Shoe Shop	Stair Case	Studio Music	Toy Store
Train Station	Tv Studio	Video Store	Waiting Room
Ware House	Winecellar	Subway	

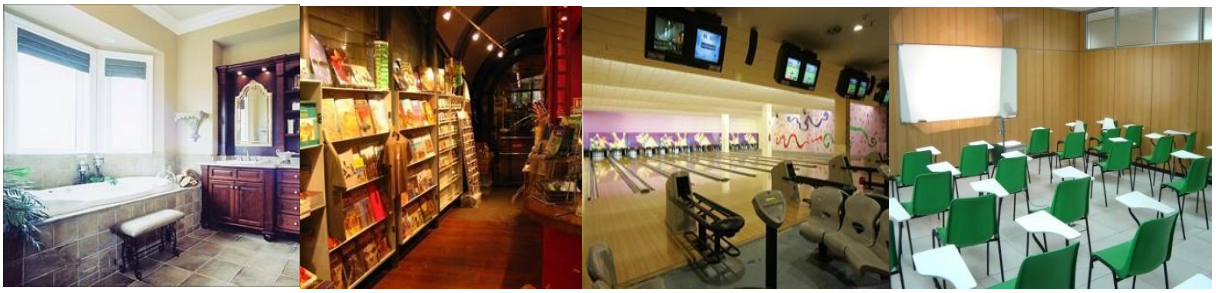
The database necessarily has some qualities like a huge degree of focus and diversified viewpoints. With two images in the dataset, possess similar styles with different frequency components [12-13]. The study uses a dataset comprising around 100,000 bedroom images. This dataset has very high density but a very low diversity, as all the images look very similar. An ideal dataset must have deliberately higher diversity for better generalization. The work is carried in two stages, namely, the first stage classifies the entire dataset in different classes. In the second stage, the scenes are grouped into with and without people, giving emphasis on improvisation of the accuracy with different augmentations.

Table 2: Number of images with and without people of MIT-67

Category	No. of Images WP	No. of Images WiP	Category	No. of Images WP	No. of Images WiP
AirportInside	106	479	Winecellar	246	24
Art Studio	44	88	Cloister	102	21
Auditorium	135	35	Closet	105	3
Bakery	197	200	Clothing Store	105	12
Bar	296	269	Computer Room	115	10
Bathroom	164		Concert Hall	102	19
Bedroom	662	9	Corridor	345	35
Green House	98	4	Gameroom	127	16
Hospital	63	36	Gym	179	47
Kinder garden	127	18	Subway	60	374
Library	107	18	Laboratory	69	52
Mall	9	158	Lobby	101	17
Nursery	144	7	Movie Theatre	145	32
Pool Inside	174	64	Operating Room	86	46
Shoe Shop	72	39	Restaurant	360	134
Train Station	69	98	Studio Music	89	18
Ware House	409	98	Video Store	64	48
Book Store	1069		Subway	198	340
Bowling	42	167	Deli	99	154
Buffet	111	12	Dental Office	60	70
Casino	106	411	Dinning Room	274	
Children Room	91	20	Elevator	64	35
Church Inside	179	41	Fastfood	36	83
Classroom	113	2	Florist	58	45
Grocery Store	124	88	Garrage	103	1
Inside Bus	45	57	Hair Saloon	189	25
Kitchen	734	3	Jewellery shop	120	32
Living Room	706	7	Laund Mat	218	59
Meeting Room	233	4	Locker Room	249	27
Office	109	1	Museum	55	103
Prison Cell	88	8	Pantry	376	13
Stair Case	155		Rest_kitchen	43	62
Tv Studio	56	115	Toy Store	227	128
			Waiting Room	137	13

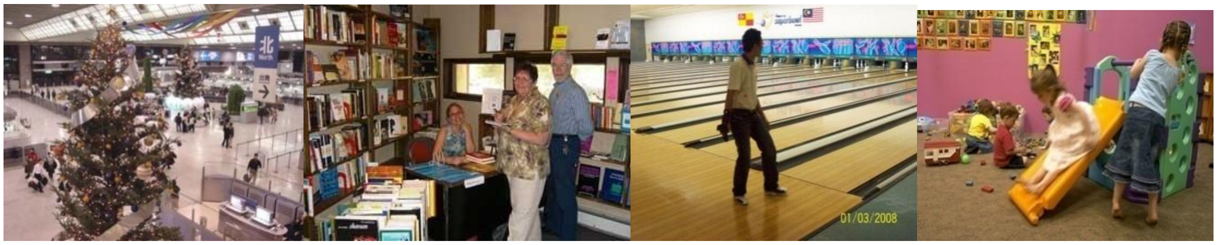
There exists many categories of classess in MIT-67 with a minimum of 50 images from each class [15]. In the dataset considered, it has images with resolution varying from 200x178 to 1024x720. A unique characteristic in the images is observed, that is, presence of with and without people. The classification depends entirely on the spatial features and the orientation in scenes [10]. The indoor scenes are identified based on the scene's global spatial properties and objects. The indoor scene of meeting rooms, Museums, etc are well characterized. But inside the bus and grocery store, the people in the scenes reduce the characterization properties and the connectivity of the objects within the scenes.

Using deep learning, we divided classes of images into two categories: those with people (WP) and those without people (WiP), both manually and automatically. The Fig. 8(a-d) and Figure 8(a-d) show the total number of images in each category. From Table 2, we have observed that the number of images without people varies from 9 (Mall) to 1069 (Book Store) and with people varies from 1 (office) to 480 (Inside Airport).



a b c d

Fig.8: Sample images of bathroom, book store, bowling, classroom (left to right) of MIT-67(without people)[18]



ab c d

Fig.9: Sample images of airport, book store, bowling, children room(left to right) of MIT-67 (with people) [18]

The images of Bookstores, meeting rooms, bowling and Bathrooms have no people, hence considered as images without people in all the 67 categories and with people we have got 64 categories. Since the people are present in the indoor scene images which distracts the spatial orientations and hence the classification of scene classes. To have a large dataset from the available images, we have used certain augmentation techniques to generate more data as the number of images are found to be less in the mall category. After the augmentation process, the total number of images generated without people scene categories are 82460 images and with people are 33579 images.Sample images after augmentation shown in fig.9.

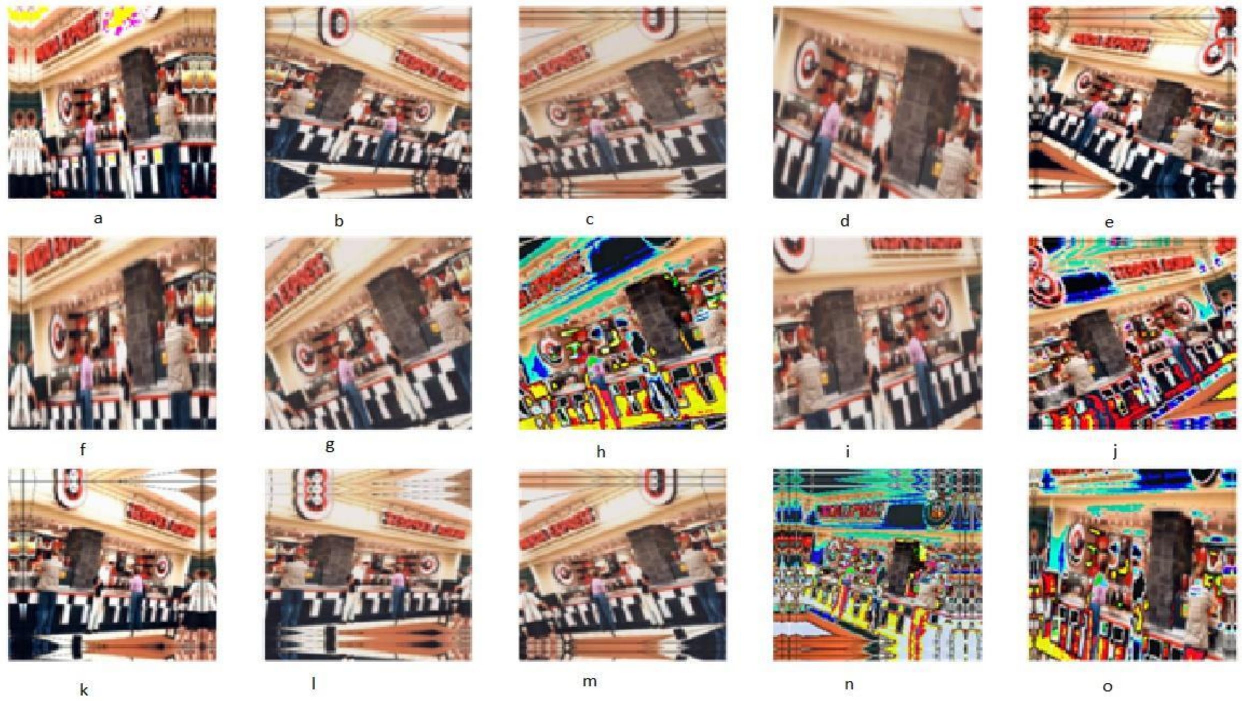


Fig. 9: Sample images after augmentation (as per Table 3)

Accuracy is a metric that generally describes how the model performs across all classes. It is useful when all classes are of equal importance. It is calculated as the ratio between the number of correct predictions to the total number of predictions. The expression for accuracy is given by expression (5).

$$\frac{Tp+Tn}{Tp+Fp+Tn+Fn} \quad \text{-------------} \ (5)$$

Where Tp: True Positive, Tn:True negative, Fp: False positive, Fn False negative.

4.1 Ablation study for different Hand crafted features

In the experiment, we considered only 07 classes, trained them randomly with color quantization, and applied different image descriptors.We have increased the classes up to 15 and the results obtained are shown in Table 3. The data set containing lesser classes gives better accuracy but when the number of classes has increased the accuracy got decreased. In another experiment, we tested for the whole 67 classes by increasing the critical points of the SIFT descriptor and by combining HOG+SURF with SIFT, the classifier used is ANN. The results obtained are shown Table.3

Table 3. Results obtained for MIT-67 Data set by varying number of classes

Feature Extraction	Classes	Accuracy(%)
Color	15	33.00
	13	33.00
	10	44.00
	7	75.00
SIFT	15	66.00
	13	68.00
	10	77.00
	7	82.00
HoG	15	55.00
	13	66.00
	10	78.00
	7	92.30

Color and HoG	15	61.50
	13	69.70
	10	72.80
	7	83.50
SIFT and HoG	15	64.50
	13	71.20
	10	73.40
	7	89.63
Color and SIFT	15	72.00
	13	74.00
	10	82.00
	7	87.80
Color,SIFT and HoG	15	71.40
	13	76.00
	10	81.46
	7	92.00
HoG + SIFT	67	88.56
Color + HoG	67	78.93
Color + SURF+ HoG	67	84.58

The experiments are carried out for different combinations of HCF with variations in the number of regions. It is observed that the classification accuracy increases with the number of regions as shown in Table.4.

Table 4: Experimental results conducted on handcrafted features with number of patches.

No. of Regions (Patches)	Accuracy(%)			
	SIFT	SIFT+SURF	SIFT+HOG	SIFT+SURF +HOG
04	79.7	86.4	78.5	81.58
03	78.6	83.54	72.14	81.08
02	72.1	81.64	73.04	81.00

The handcrafted features have given good results, but the scope exists to improve accuracy and, hence fusion of Efficient-Net with handcrafted features gives improved accuracies. Even we conducted the classifications by changing the activation functions at each region.

4.2 Ablation study on fused Efficient-Net model

The activation functions used in this work are most widely used in different CNN architectures and are given better performanc. Relu is found faster compared to the Tanh function. Mathematically, ReLu is defined as in expression (2).

$$G(t) = \{max(0,t), \quad when\ input\ is + ve\ 0 \qquad when\ input\ is - ve \qquad \ldots(2)$$

Sigmoid activation function is found to be monotonic and continuously differentiable as defined by expression (3), which normalizes the input and outputs values in the range [0,1]. The only constraint of sigmoid is that it possesses vanishing gradients that make the detection of fewer features.

$$G(t) = \frac{1}{1+e^{G(t)}} \qquad \ldots (3)$$

Tanh is centered to zero and falls in the range [-1, 1] and is given by expression (4), when compared to sigmoid and Relu during feature extraction.

$$G(t) = \frac{1-e^{-2t}}{1+e^{-2t}} \qquad \dots (4)$$

Amongst the activation functions sigmoid has performed better for the combination of Regions 3 and 4 when HoG+SIFT HCF are fused with the Efficient-Net model. The sigmoid has not given the same performance for the same features extracted from individual regions and also with images WP and WiP. Tanh has given a maximum of 89.98% for region-2 with HoG and SIFT features and 90.45 for without people. ReLu has given 92.38 for Region 3 and 4. From these statistics we infer that the combination of HoG + SIFT features is found to be effective in the classification of indoor scenes. Amongst the activation functions Sigmoid has given highest accuracy of 93.87% for the same features extracted from the combination of features extracted from Region 3 and 4. Overall, irrespective of scenes WP and WiP sigmoid is found to perform better compared to other activation functions.

The results of the ablation study are tabulated in Table.5 for different regions, activation functions, hand crafted features and images with people and without people.

From the table 5 it is observed that there are instances of misclassification causing reduction in the classification accuracy the images of library, bedroom, dental office, auditorium etc are misclassified as bookstore, living room, operating room, concert hall respectively and vice versa

Table.5 HCF fused efficientnet model performance

	Accuracy(%)								
Experiments	Sigmoid			Tanh			Relu		
	HoG+E	SIFT+E	HoG+SIFT+E	HoG+E	SIFT+E	HoG+SIFT+E	HoG+E	SIFT+E	HoG+SIFT+E
Without People for whole Image	81.51	83.72	89.72	82.00	82.73	89.99	81.12	83.18	88.76
With People for whole image	86.26	85.38	89.98	87.72	88.12	90.45	88.12	87.73	89.64
Region-1+Efficient-Net.	86.5	86.7	86.90	86.4	84.48	87.18	86.03	86.72	88.35
Region-2+Efficient-Net	85.02	86.23	89.71	84.47	87.82	89.98	85.65	88.89	87.76
Region-3+Efficient-Net.	85.59	86.01	86.68	87.81	88.01	89.01	88.01	88.78	89.34
Region-4+Efficient-Net	87.71	86.68	87.81	88.01	86.56	86.71	88.91	89.01	89.37
Region-1+2 Efficient-Net	84.48	84.58	84.98	85.56	85.55	85.08	86.66	85.59	84.48
Region-3+4 Efficient-Net	90.85	91.23	**93.87**	91.23	92.27	92.38	90.87	91.19	92.38

4.3 Experiments for Classification with State-of-the-Art

From the Table.6 it is observed that the existing methods have given the accuracies in the range [43.1% - 81%]. The methods reported in the literature have used individual features such as SIFT color, GISt etc and their combination. The CNN models used are the ResNet model. All the authors have used the entire image but the useful information required for classification for indoor scenes is present in specific regions as objects and their spatial relationship are present in specific regions. This idea is used in the proposed work. It is observed that the regions 3 and 4 with Efficient-Net give better results, as the spatial information of indoor scenes is more evident in these regions. The Table.6 gives the comparison of results on the MIT-67 dataset. The proposed method has achieved 93.87%.

Table 6: Comparison results on MIT-indoor67

Method	Average Precision Accuracy
Deep filter banks [28]	81 %
ResFeats-152 + PCA-SVM [25]	75.6%
SIFT[29]	74.4%
ResFeats-152 + sCNN classifier [30]	73.7%
l2 Normlization + Selective Search + Spatial Pyramid [31]	74.4%
DPM+GIST-color+SP[32-36]	43.1%
Entire Image CNN Features	68.04%
Proposed Method	**93.87%**

5.0 CONCLUSION

This paper discussed the effect of hand-crafted features fusion, at fully connected regions, with EfficineNet model features on the performance of CNN. The performance of hand-crafted features is tested using ANN classifier. It is concluded that by dividing an image into regions,obtaining features from these regions and fusing features with deep learning models (EfficineNet) resulted in good performance. From the results of experiments, it is also evident that the proposed method gives 93.87% precision rate and outperforms the other contemporary methods. The scope exists for a separate study on the effect of the other hyper parameters on the classification accuracy of indoor scenes and also extending the work on other CNN models.

REFERENCES

[1] Zhou T, Miao, Z; Zhang, J, Combining CNN with Hand-Crafted Features for Image Classification, 2018 *14th IEEE International Conference on Signal Processing (ICSP),* 2019, pp. 554-557

[2] McCulloch, WS., and Walter, P. "A logical calculus of the ideas immanent in nervous activity." *The bulletin of mathematical biophysics vol* 5 No 4, 1943, pp.115-133, .

[3] Kuppuswamy, S., & Panchanathan, B . Similar object detection and tracking in H. 264 compressed video using modified local self similarity descriptor and particle filtering. *International Journal of Intelligent Engineering and Systems,* vol *10*, no 5, 2017 pp 95-104.

[4] Girshick, R., Donahue, J., Darrell, T., & Malik, J.. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 2014, pp. 580-587.

[5] Bhatt, M., & Patalia, T. Neural network based Indian folk dance song classification using MFCC and LPC. *International Journal of Intelligent Engineering and Systems*, vol *10*, no 3, 2017, pp 173-183.

[6] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L.. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* , June 2009, pp. 248-255

[7] Herouane, O., Moumoun, L., Gadi, T., & Chahhou, M. A hybrid boosted-SVM classifier for recognizing parts of 3D objects. *International Journal of Intelligent Engineering and Systems*, Vol *11*, No 2 , 2018, pp 102-110.

[8] Patel, H., & Mewada, H.. Analysis of machine learning based scene classification algorithms and quantitative

evaluation. *International Journal of Applied Engineering Research*, Vol *13*, No 10, 2018, pp 7811-7819.

[9] Quattoni, A., & Torralba, A. . Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition* , June 2009, pp. 413-420.

[10] Lazebnik, S., Schmid, C., & Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)* Vol. 2, June 2006, pp. 2169-2178. IEEE.

[11] Wang, Y., & Wu, Y. Scene classification with deep convolutional neural networks. *University of California*, 2014

[12] Wang, P., & Cottrell, G. W. Basic level categorization facilitates visual object recognition. 2015, *arXiv preprint arXiv:1511.04103*.

[13] Krizhevsky, A., Sutskever, I., & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, vol *60*, No 6, 2017, pp 84-90.

[14] Fei-Fei, L., & Perona, P. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* Vol. 2, June 2005, pp. 524-531

[15] Quelhas, P., Monay, F., Odobez, J. M., Gatica-Perez, D., Tuytelaars, T., & Van Gool, L. Modeling scenes with local descriptors and latent aspects. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Vol 1* October 2005, pp. 883-890

[16] Bosch, A., Zisserman, A., & Munoz, X. Scene classification using a hybrid generative/discriminative approach. *IEEE transactions on pattern analysis and machine intelligence*, Vol *30* No 4, 2008, pp 712-727.

[17] Boiman, O., Shechtman, E., & Irani, M. (2008, June). In defense of nearest-neighbor based image classification. In *2008 IEEE conference on computer vision and pattern recognition* . June 2008, pp 1-8.

[18] Elad, M., and Aharon. M. "Image denoising via sparse and redundant representations over learned dictionaries." *IEEE Transactions on Image processing vol* 15 No 12, 2006 pp.3736-3745.

[19] Moosmann, F., Triggs, B., & Jurie, F. Randomized clustering forests for building fast and discriminative visual vocabularies. Neural Inf. Process. Syst. November 2006

[20] Yang, L., Jin, R., Sukthankar, R., & Jurie, F. Unifying discriminative visual codebook generation with classifier training for object category recognition. In *2008 IEEE conference on computer vision and pattern recognition* June 2008, pp. 1-8

[21] Warburg F, Hauberg S, Lopez-Antequera M, Gargallo P, Kuang, Y, Civera, J. Mapillary street-level sequences: A dataset for lifelong place recognition. IEEE Conference on Computer Vision and Pattern Recognition , 2020 pp. 2626–2635

[22] Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T . Patch-netvlad: multi-scale fusion of locally-global descriptors for place recognition. IEEE Conference on Computer Vision and Pattern Recognition , 2021, pp. 14141–14152

[23] Quattoni, A, Torralba, A .Recognizing indoor scenes. IEEE Conference on Computer Vision and Pattern Recognition , 2009, pp. 413–420

[24] Toft, C., Maddern, W., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., ... & Sattler, T. Long-term visual localization revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol *44*, no 4, 2020, 2074-2088.

[25] Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. . Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition* June 2010, pp. 3485-3492.

[26] Glavan, A., Talavera, E. InstaIndoor and multi-modal deep learning for indoor scene recognition. *Neural Comput & Applic* vol 34, 2022 pp 6861–6877.

[27] Bai, S., Tang, H., & An, S. Coordinate CNNs and LSTMs to categorize scene images with multi-views and

multi-levels of abstraction. *Expert Systems with Applications*, vol *120*, 2019, pp 298-309.

[28] Cimpoi, M., Maji, S., & Vedaldi, A. Deep filter banks for texture recognition and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , 2015, pp. 3828-3836).

[29] Hayat, M., Khan, S. H., Bennamoun, M., & An, S. A spatial layout and scale invariant feature representation for indoor scene classification. *IEEE Transactions on Image Processing*, vol *25*, No 10, 2016. pp 4829-4841.

[30] Van Gemert, J. C., Geusebroek, J. M., Veenman, C. J., & Smeulders, A. W. Kernel codebooks for scene categorization. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings*, Springer Berlin Heidelberg, *Part III 10* pp. 696-709. .

[31] Yang, J., Yu, K., Gong, Y., & Huang, T. Linear spatial pyramid matching using sparse coding for image classification. In *2009 IEEE Conference on computer vision and pattern recognition*, June 2009, pp. 1794-1801

[32] Basavanna, M., Shivakumara, P., Srivatsa, S. K., & Kumar, G. H. Adaptive Histogram Analysis for Scene Text Binarization and Recognition. Malaysian Journal of Computer Science, vol 29 no 2, pp 74–85. https://doi.org/10.22452/mjcs.vol29no2.1

[33] Rassem, T. H., Khoo, B. E., Mohammed, M. F., & M.Makbol, N. (2017). Medical, Scene And Event Image Category Recognition Using Completed Local Ternary Patterns (CLTP). Malaysian Journal of Computer Science, vol 30, no 3, pp 200–218.

[34] Nagaraja, B.G. and Jayanna, H.S., Multilingual speaker identification with the constraint of limited data using multitaper MFCC. In International conference on security in computer networks and distributed systems, Springer, Berlin, Heidelberg , October 2012, pp. 127-134 .

[35] Nagaraja, B.G. and Jayanna, H.S., . Feature extraction and modelling techniques for multilingual speaker recognition: a review. International Journal of Signal and Imaging Systems Engineering, vol 9, no 2, 2016, pp.67-78.

[36] Yadava, T. G., Nagaraja, B. G., & Jayanna, H. S. A spatial procedure to spectral subtraction for speech enhancement. *Multimedia Tools and Applications*, vol *81* no 17, 2022, pp 23633-23647.

[37] Y. Wu, H. Guo, C. Chakraborty, M. Khosravi, S. Berretti and S. Wan, "Edge Computing Driven Low-Light Image Dynamic Enhancement for Object Detection," in *IEEE Transactions on Network Science and Engineering*, 2022, doi: 10.1109/TNSE.2022.3151502.

[38] Y. Wu, L. Zhang, S. Berretti and S. Wan, "Medical Image Encryption by Content-Aware DNA Computing for Secure Healthcare," in *IEEE Transactions on Industrial Informatics*, vol 19, no 2, pp. Feb. 2023 pp 2089-2098, , doi: 10.1109/TII.2022.3194590.

[39] Shi, G., Wu, Y., Liu, J., Wan, S., Wang, W., & Lu, T. . Incremental Few-Shot Semantic Segmentation via Embedding Adaptive-Update and Hyper-class Representation. In Proceedings of the 30th ACM International Conference on Multimedia (MM '22). Association for Computing Machinery, New York, NY, USA, 2022, October 2022, pp 5547–5556. https://doi.org/10.1145/3503161.3548218