Advances in Nonlinear Variational Inequalities
ISSN: 1092-910X
Vol 27 No. 2 (2024)

Advanced Deep Learning Techniques for Indoor-Outdoor Scene Recognition Integrating CNN and Edge Detection for Enhanced Classification Accuracy in Dynamic Environments

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 History:

***Received**: 06-02-2024*

***Revised**: 21-04-2024*

***Accepted**: 16-05-2024*

Abstract:

Recognizing indoor-outdoor scenes is an vital portion of computer vision that has an impact on numerous zones, counting driverless driving, virtual reality, and natural following. This article talks approximately a more progressed profound learning framework that employments Convolutional Neural Systems (CNNs) and edge acknowledgment strategies to create classification more precise in settings that alter over time. Distinctive lighting, surfaces, and structure highlights can make it difficult for conventional scene acknowledgment strategies to tell the contrast between complex scenes. We recommend a way to solve these issues that employments the leading parts of both CNNs and edge acknowledgment to urge exact pictures of geometric shapes. The system begins by altering the photographs it gets by utilizing edge acknowledgment strategies, like Canny and Sobel, to discover critical structure edges and lower the clamor. The edge-enhanced pictures are at that point sent to a CNN engineering that was made to recognize scenes. The CNN show is learned on a huge dataset that incorporates a wide run of indoor and open air settings. This makes beyond any doubt that it works well in all sorts of circumstances. A modern layer integration strategy is utilized to connect the CNN's learned highlights with the edge-based highlights. This makes a difference demonstrate tell the contrast between complex scene points of interest superior. It comes about of our tests appear that our combined strategy makes classification much more exact than conventional CNN models. The edge acknowledgment portion truly makes strides the model's capacity to choose up on little changes in picture structures, which leads to more exact forecasts. Moreover, the framework's capacity to adjust to changing situations has been demonstrated by a parcel of testing in a part of distinctive circumstances. Combining profound learning with edge discovery might make scene acknowledgment way better, and the proposed strategy shows how this may work. This may lead to more advanced and viable computer vision frameworks in real-world circumstances that are complicated.

Keywords: Scene Recognition, Convolutional Neural Networks (CNNs), Edge Detection, Dynamic Environments, Classification Accuracy

1. Introduction

Scene recognition has become an important area of study in computer vision in recent years. This is often since individuals need shrewd frameworks that can get it and bargain with diverse environment. Exact scene acknowledgment is required for machines to see and get it the world around them, from self-driving cars that can explore cities to virtual reality apps that make client encounters way better. Indeed in spite of the fact that a part of advance has been made, it is still difficult to urge great classification precision in settings that are complicated and changeable [1]. This think about looks at how progressed profound learning strategies, like Convolutional Neural Systems (CNNs) and edge discovery, can be utilized together to create picture acknowledgment more exact both interior and exterior. Conventional strategies of scene acknowledgment depend on highlights that were made by hand and learning models that aren't exceptionally profound. These strategies have inconvenience capturing the wealthy and complex subtle elements that are show in complex scenes. These strategies are particularly touchy to changes in lighting, designs, and structure highlights, which can make them work much less well [2]. CNNs, on the other hand, have done exceptionally well in numerous computer vision occupations since they can effortlessly learn organized highlights from crude picture information. Since CNNs are great at recording spatial connections and learning models that do not alter easily, they are incredible for recognizing scenes.

Indeed in spite of the fact that CNNs are great at a few things, they can still have inconvenience picking out little changes in scenes that are exceptionally complicated. This issue ordinarily happens since individuals do not pay sufficient consideration to geometric shapes and lines, which are exceptionally critical for accurately deciphering scenes. When looking at a picture, edges, which show the lines between diverse zones, are exceptionally supportive for figuring out how the space is organized. Including edge location strategies to the scene acknowledgment prepare can offer assistance the demonstrate center on these important aspects, which can lead to more precise classification. Edge acknowledgment strategies, like Canny, Sobel, and Laplacian, have been utilized for a long time to drag out pictures' clear lines and borders [3]. These strategies appear ranges with huge changes in quality, which makes it simple to tell the distinction between buildings and things in a picture. The recommended strategy combines the most excellent parts of both approaches by utilizing edge discovery and CNNs together. CNNs are great at extricating highlights, and edge discovery gives precise data almost boundaries. By combining these two things, the show can see both huge and little subtle elements, which makes a difference us get it complicated scenes superior. The proposed approach begins by planning the pictures that are sent in by utilizing edge acknowledgment methods to move forward the edges of structures and get freed of commotion. This step some time recently handling makes edge-enhanced pictures that draw consideration to critical parts of the scene. These pictures are at that point sent to a CNN design that was made fair for recognizing scenes [4]. The CNN demonstrate is learned on a expansive dataset that incorporates a wide extend of indoor and open air settings. This makes beyond any doubt that it works well in all sorts of circumstances. We utilize a

unused layer integration strategy to combine the CNN's learned highlights with the edge-based highlights. This lets the show make great utilize of the complementary data from both sources. This think about appears that the combined strategy is much way better at classifying things than customary CNN models by doing a part of tests [5]. Edge acknowledgment makes strides the model's capacity to choose up on little changes in picture structures, which makes estimates more precise. The system can too react to changing situations since it has been tried altogether in a number of diverse circumstances. this appears how solid and adaptable it is. This discussed about appears that blending profound learning with edge location seem offer assistance recognize scenes superior. This seem lead to more progressed and viable computer vision frameworks in real-world settings that are more complicated. This think about includes to the advance of scene acknowledgment strategies by settling the issues with current strategies and making a other way to combine them. The comes about appear how critical it is to utilize methods that work well together to induce around the issues that come up in changing settings and move forward the exactness of classification.

The key contribution of paper is given as:

- Integration of Edge Location and CNNs: This consider depicts a framework that combines edge discovery strategies with Convolutional Neural Systems (CNNs) to make strides scene distinguishing proof. It does this by utilizing CNNs' highlight extraction and edge detection's structure data to create it simpler to tell the distinction between scenes.
- More precise classification in changing situations: The strategy makes classification much more precise both interior and exterior, and it works well indeed when lights, colors, and structures alter. This implies it can be utilized in real-life changing situations.
- Imaginative Highlight Combination Procedure: A unused layer integration strategy is appeared that combines CNN highlights with edge-based highlights. This lets the demonstrate get it both worldwide and nearby scene highlights, which progresses its capacity to foresee what will happen another

2. Related Work

Scene discovery may be a active zone of think about in computer vision. Profound learning strategies have made a difference make huge steps forward in this region. Convolutional Neural Systems (CNNs) have been at the head of these changes since they can learn complex designs from crude picture information on their own. Early inquire about appeared that CNNs might be valuable for a number of picture classification assignments, which made it conceivable for them to be utilized for scene acknowledgment [6, 7]. CNNs are great at a part of things, but they have inconvenience accurately recognizing scenes with settings that are complicated and changeable. Since they as it were utilize learned various leveled highlights, conventional CNN models regularly have inconvenience picking up on small details and minor contrasts in scenes. Analysts have looked into a number of ways to urge around these issues, such as including more information sources to CNNs, such as edge location [8, 9]. In

computer vision, edge acknowledgment strategies are regularly utilized to induce valuable structure information from pictures. Calculations like Canny, Sobel, and Laplacian show where one region closes and another starts, giving vital clues for figuring out what a scene implies. Including edge location to CNNs may be a potential way to make strides the precision of scene recognition by blending learning around worldwide highlights with learning almost nearby structures [10, 11].

Edge location and profound learning have been utilized together in a number of ventures to make strides picture acknowledgment. For case, a few analysts have included edge maps as an additional input channel to CNNs. This lets the network utilize both concentrated and structure data whereas it's preparing. This strategy appears like it might offer assistance make strides acknowledgment exactness, particularly in places where lighting and highlights alter a parcel [12, 13]. Other than edge location, other strategies have been recommended to create CNN-based scene acknowledgment better. A method for doing this is often to utilize multi-scale highlight extraction to urge data at diverse levels of detail. By utilizing highlights from diverse levels of the CNN [14, 15], this strategy points to create the model way better at recognizing scenes with distinctive sizes and focuses of see. A diverse zone of ponder has been making blended models that utilize both CNNs and other machine learning strategies. Within the case of video-based scene acknowledgment assignments, for occurrence, analysts have looked into how CNNs and repetitive neural systems (RNNs) can work together to record time connections and relevant data. These blended models are superior at recognizing changing scenes, which appears how vital it is to utilize techniques that work together to urge superior comes about [16, 17].

Within the past few a long time, consideration strategies have moreover gotten to be more prevalent in scene recognizable proof. Consideration forms let models center on imperative parts of a picture, which makes a difference them see imperative highlights and make superior classifications. Analysts have made enormous steps forward in scene acknowledgment by including consideration modules to CNN plans [18, 19]. This works especially well in difficult settings with busy backgrounds and occlusions. Even with these improvements, it is still hard to get accurate classification in a wide range of changing settings. A lot of the current methods need a lot of computing power and big datasets to be taught, which makes them hard to use in situations where time is limited or resources are limited. Also, models that were trained on certain datasets might not work well in settings they haven't seen before. This shows how important it is to have solutions that are both strong and flexible [20, 21]. To get around these problems, our suggested system combines CNNs with edge detection to make picture recognition more accurate both inside and outside. We focus on important structure elements and lower noise in raw pictures by using edge recognition methods before they are sent to the CNN. This lets the CNN focus on important features. We also use a new layer integration method that combines edge-based features with CNN-learned features, using information that complements each other to make the system work better [22]

Table 1: Related work Summary

Method	Approach	Key Finding	Limitation	Scope
Multi-Scale CNN	Employs multi-scale feature extraction within CNN architecture	Captures information at different granularity levels for improved recognition	Computationally intensive	Applicable to scenes with varying scales and perspectives
CNN + RNN Hybrid	Combines CNNs with RNNs for temporal context	Improves recognition of dynamic scenes by capturing temporal dependencies	Complex model architecture	Effective in video-based scene recognition
Attention Mechanisms [23]	Integrates attention modules into CNNs	Enhances model focus on relevant image regions, improving classification accuracy	Requires additional model complexity	Useful in challenging environments with cluttered backgrounds
Edge Maps as Input [24]	Uses edge maps as an additional input channel for CNN	Leverages both intensity and structural information to boost accuracy	Edge detection preprocessing step required	Enhances recognition accuracy in environments with high texture variability
CNN with Data Augmentation	Augments training data with transformations to improve generalization	Increases model robustness to unseen data through diverse training examples	Potentially increases training time	Broad applicability to various scene recognition tasks
Hybrid Deep Models	Integrates deep learning with traditional machine learning algorithms	Combines strengths of different techniques for improved performance	May require extensive parameter tuning	Suitable for complex scene recognition tasks
Transfer Learning [25]	Utilizes pre-trained CNN models for scene recognition	Reduces need for large datasets and extensive training	May not fully adapt to specific scene characteristics	Useful for quick deployment in new environments
Semantic Segmentation	Applies segmentation to enhance scene understanding	Provides detailed scene breakdown, aiding recognition	Requires precise labeling and segmentation	Effective in scenes where detailed structure understanding is crucial
Graph-Based CNN [26]	Incorporates graph structures into CNNs to capture spatial relationships	Improves recognition by considering spatial hierarchies in scenes	Complex to implement and requires graph construction	Applicable to scenes with complex spatial arrangements
Lightweight CNNs	Develops compact CNN models for efficient scene recognition	Achieves competitive accuracy with reduced computational cost	May sacrifice some accuracy for efficiency	Ideal for real-time applications and resource-constrained environments
CNN with Feature Fusion	Combines features from multiple CNN	Enhances model ability to capture	Increased model complexity	Useful for tasks requiring

	layers	both high-level and low-level details		comprehensive scene analysis
Robust CNN Architectures	Designs architectures resistant to environmental changes	Improves recognition performance across varying lighting and weather conditions	May require extensive architecture exploration	Suitable for outdoor scene recognition in dynamic environments

3. Dataset Used

A. Places365 Dataset:

The Places365 dataset may be a comprehensive collection of pictures planned for scene acknowledgment assignments. Created by MIT, it incorporates over 10 million pictures categorized into 365 particular scene classes, extending from indoor situations like libraries and kitchens to open air settings such as shorelines and woodlands. The dataset, sample image show in figure 1, is part into preparing, approval, and testing subsets, giving a vigorous establishment for preparing and assessing profound learning models. One of the key highlights of the Places365 dataset is its differing qualities, capturing a wide run of scenes from distinctive points of view, lighting conditions, and settings. This contrasts grants models arranged on the dataset to generalize well to real-world scenarios. Besides, Places365 has been instrumental in advancing explore in scene affirmation, serving as a benchmark for evaluating the execution of novel calculations and structures. It supports the change of models competent of absolutely understanding and classifying complex scenes, making it a productive resource for both academic ask almost and reasonable applications in computer vision.

Figure 1: Sample image of dataset

B. CVPR 09 dataset

The CVPR 09 dataset for Indoor Scene Affirmation may be a basic commitment to the field of computer vision, particularly for scene affirmation errands. Made for the CVPR 2009 conference, this dataset contains a contrasting collection of pictures talking to diverse indoor

circumstances. The dataset joins 67 categories of indoor scenes, such as classrooms, kitchens, and living rooms, giving a comprehensive foundation for planning and evaluating scene affirmation models, test picture appeared in figure 2. One of the key highlights of the CVPR 09 dataset is its contrasts in scene sorts and changeability in picture conditions, such as lighting, occlusions, and perspectives. This contrasts challenges models to generalize well over assorted indoor circumstances, making it an essential resource for making overwhelming scene affirmation systems. The dataset has been broadly utilized in academic explore and has contributed to vital headways in indoor scene understanding. Its utilize has driven to the progression of more correct and compelling calculations, which are fundamental for applications in mechanical innovation, extended reality, and cleverly perception systems. The CVPR 09 dataset remains a benchmark for surveying the execution of scene affirmation calculations in indoor settings.

Figure 2: Sample image of CVPR Dataset indoor image scene

4. Proposed Methodology

The figure 3 appears a total handle for recognizing scenes utilizing machine learning and profound learning. The method begins with a collection that has pictures of distinctive scenes. These pictures are put into the framework as "Image with Scenes." The primary step is to plan these pictures for advance ponder by pre-processing them. In this step some time recently preparing, assignments like trimming, normalization, and commotion lessening may be utilized to improve picture quality and make beyond any doubt that the total collection is steady. After the pictures are pre-processed, a include extraction apparatus pulls out the imperative parts of the images. Include extraction is exceptionally vital since it helps find imperative designs and characteristics within the pictures, which lets the show center on the foremost vital parts of the scenes. After features are removed, they are used to separate training samples from testing samples. This splits the dataset into parts that can be used to train and test the model.

Scene groups, like "indoor" and "outdoor," are often included in labels, and this is what the model tries to guess. It could be a machine learning program or a deep learning model that gets these named data to learn from them. During training, the classifier learns trends from the raw data so it can correctly put pictures it hasn't seen into the right category. After the model has been taught, it is put to the test on the separate testing samples to see how well it works. The trained network then sends out the scene category predictions, which use learning patterns to decide whether a picture is indoors or outdoors. Lastly, the system checks performance factors to see how well the sorting process works. Some of these factors are accuracy, precision, recall, and F1-score, which show what the model does well and where it could be improved.

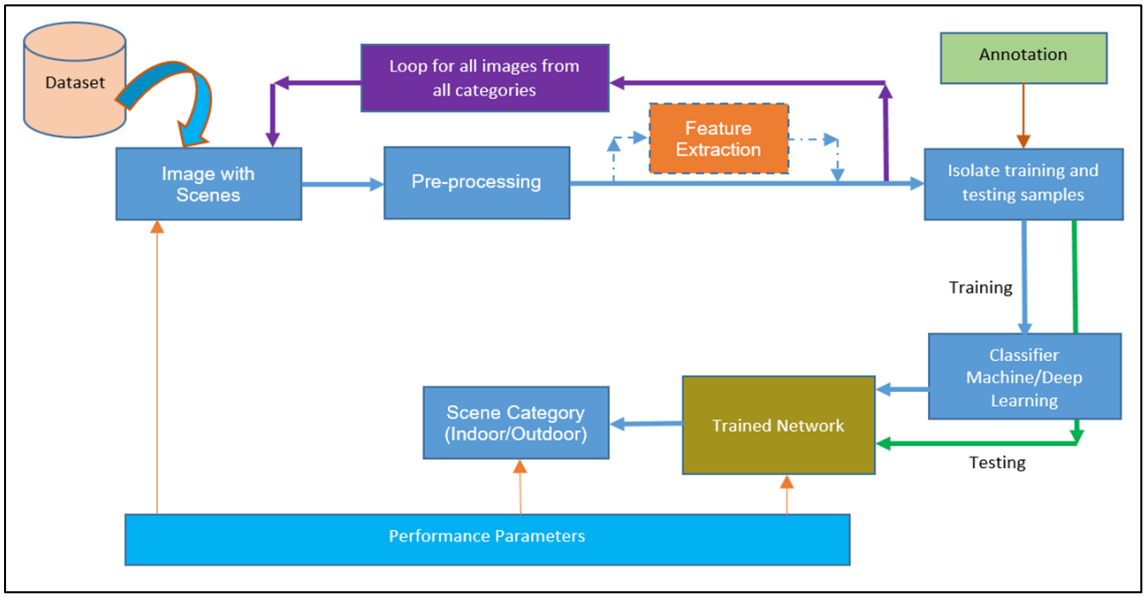

Figure 3: Representation of proposed system Model

The whole loop goes through all the pictures in all the different groups, making sure that all of them are covered and that the scene recognition is strong. This method shows how pre-processing, feature extraction, and machine learning can be used together to improve the accuracy of classification in changing settings.

A. Object based features

Object-based highlights are the characteristics and qualities of things in a picture that offer assistance with assignments like recognizing scenes and putting them into bunches. These highlights are all approximately finding and examining the diverse parts of a scene, just like the shapes, colors, surfaces, and how things are set in space. Object-based highlights are exceptionally vital in computer vision for telling scenes separated that see comparative on the exterior but have diverse combinations of objects. For example, an indoor scene with tables and chairs can be called a school, whereas an indoor scene with a bed and closet may well be called a room. These characteristics can be extricated with the assistance of strategies like protest acknowledgment and division, which makes it simpler for models to distinguish and classify scenes.

The YOLO model predicts multiple bounding boxes and class probabilities simultaneously. The loss function for YOLO combines classification, localization, and confidence loss:

$$L_{YOLO} = \lambda_{coord} * \Sigma_i = 0_j^{S^2\Sigma} = 0^B 1_{ij}^{obj}((x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2) + \lambda_{coord} * \Sigma_i = 0_j^{S^2\Sigma}$$

$$= 0^B 1_{ij}^{obj}\left(\left(\sqrt{w_i} - \sqrt{\hat{w}_i}\right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i}\right)^2\right)$$

$$+ \Sigma_i = 0_j^{S^2\Sigma} = 0^B 1_{ij}^{obj}\left(C_i - \hat{C}_i\right)^2 + \lambda_{noobj} * \Sigma_i = 0_j^{S^2\Sigma}$$

$$= 0^B 1_{ij}^{noobj}\left(C_i - \hat{C}_i\right)^2 + \Sigma_i = 0_c^{S^2 1_i^{obj}\Sigma} \in classes\left(p_{i(c)} - \hat{p}_{i(c)}\right)^2$$

The R-CNN model extracts features using a CNN and classifies them using a linear SVM. The objective function for training the SVM classifier is:

$$\min \_w, \xi\ (1/2)\ ||w||\text{^}2 + C\ \Sigma\_i = 1\text{^}N\ \xi\_i$$

$$subject\ to\ y_i(w^T\varphi(x_i) + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0, \forall\ i$$

FCNs are used to assign a label to every pixel in an image. The cross-entropy loss function for pixel-wise classification in semantic segmentation is:

$$L_{FCN} = -\frac{1}{N}\Sigma_i = 1^N\Sigma_c = 1^C y_{ic}\log(\hat{p}_{ic})$$

HOG features are used to describe object shapes and appearances. The computation involves calculating the gradient magnitude and orientation:

$$Gradient\ Magnitude: M(x,y) = \sqrt{I_{x(x,y)}^2 + I_{y(x,y)}^2}$$

$$Gradient\ Orientation: \theta(x,y) = \tan^{(-1)}\left(\frac{I_{y(x,y)}}{I_{x(x,y)}}\right)$$

- I_x and I_y are the gradients of the image in the x and y directions, respectively.
- M(x, y) and θ(x, y) are the magnitude and orientation of the gradient at pixel (x, y).

Object-based features make models easier to understand by focusing on important parts. This makes them very useful for tasks that need to analyze scenes in great detail, like robots, spying, and self-driving cars. Using object-based features can help models do better in settings that change quickly, where it's important to know what's going on and how things are interacting with each other for accurate scene recognition.

B. Scene based Global features

Scene-based worldwide highlights center on capturing the generally characteristics and setting of a whole picture instead of particular objects inside it. These highlights are vital for errands like scene acknowledgment, where understanding the common setting is more imperative than recognizing person objects. Here's a step-wise talk utilizing scientific conditions to portray the method of extricating and utilizing worldwide highlights for scene acknowledgment:

Step 1: Image Preprocessing

Some time recently extricating worldwide highlights, pictures are ordinarily preprocessed to guarantee consistency and improve critical points of interest. This step includes resizing, normalization, and conceivably color transformation.

$$I' = (I - \mu) / \sigma$$

This normalization step ensures that the input information includes a cruel of zero and a standard deviation of one, moving forward the merging of learning calculations.

Step 2: Feature Extraction Using Global Descriptors

Worldwide descriptors capture the pith of a whole picture. A common approach is utilizing the Significance descriptor, which captures the spatial structure of a scene.

$$G(x, y) = \Sigma_i = 1^N \Sigma_j = 1^M I'(i, j) * Gabor(i - x, j - y, \theta, f)$$

Step 3: Dimension Reduction

High-dimensional highlight vectors can be computationally costly to handle, so methods like Foremost Component Investigation (PCA) are utilized to diminish dimensionality whereas protecting basic data.

$$Z = W\hat{}T * (G - \mu)$$

- Z is the transformed feature vector.
- W is the matrix of principal components.
- μ is the mean of the GIST features.

Step 4: Scene Classification

Once the features are extracted and reduced, they can be fed into a classifier, such as a Support Vector Machine (SVM), for scene classification.

$$f(x) = sign(w^T * Z + b)$$

These steps outline the strategy of utilizing scene-based around the world highlights for scene affirmation, highlighting the noteworthiness of preprocessing, incorporate extraction, and classification in understanding and classifying scenes absolutely. Around the world highlights donate a all-encompassing see of the picture, enabling models to classify scenes based on by and huge plans and structures rather than specific challenge focuses of intrigued.

C. Local features based on statistical descriptors

Neighbourhood highlights based on measurable portrayals are exceptionally vital in computer vision assignments like recognizing objects and classifying scenes since they choose up on little subtle elements and designs in a picture. In differentiate to global features, which appear the complete scene, neighbourhood highlights concentrate on certain zones or places of intrigued, gathering points of interest approximately shapes, edges, and surfaces. To conversation around neighbourhood highlights, individuals regularly utilize measurable terms like Scale-Invariant Highlight Change (Filter), Speeded-Up Strong Highlights (SURF), and

Neighbourhood Double Designs (LBP). These portrayals are made to remain the same when things are changed, like when they are measured, rotated, or lit up in an unexpected way. This makes them dependable for utilize within the genuine world. For case, Filter finds imperative focuses in a picture and makes a portrayal vector from the neighbourhood slope headings around these focuses. This strategy lets Filter coordinate characteristics between pictures, indeed in the event that the pictures were taken from distinctive points or had distinctive lighting. On the off chance that you see at LBP, it looks at the neighbourhood picture by comparing each pixel to its neighbours and sparing the result as a parallel design. This method works truly well for employments like recognizing faces and diverse surfaces. Nearby highlights, which utilize measurable depictions, make it simpler for vision frameworks to precisely spot and tell the distinction between scenes and things, indeed when conditions are awful.

The VGG19 organize could be a well-known profound learning show that's great at extricating highlights for picture acknowledgment errands. It has 19 layers, 16 convolutional layers, 3 completely linked layers, and 5 pooling layers. These layers are implied to induce nitty gritty data from pictures. VGG19 successfully pulls out various leveled highlights that move forward picture acknowledgment by using small open areas (3x3 convolutions) and a profound plan. In highlight extraction, VGG19 could be a show that has as of now been prepared and employments approaching pictures to make highlight maps that appear imperative visual data.

Table 2: Local features based on statistical descriptors

Name of the Feature	Type of feature	Number of features extracted
Matched Filter – Edge Kernel-based	Fine	256
Matched Filter – Orientation based	Coarse	32
Wavelet – Six wavelets ['bior3.1', 'bior3.5', 'bior3.7', 'db3', 'sym3', 'haar']	Coarse	24
Wavelet - Haar	Fine	1024
GLCM – [Contrast, Energy, Homogeneity, Correlation, ASM and Dissimilarity]	Coarse	6
LBP (linear Binary pattern)	Fine	256
LBP – Texture (Averaging 5x5 block)	Fine	676
HOG	Coarse	36
Total Feature length		2310

D. Convolutional Neural Network based Classification Network

Within the region of computer vision, convolutional neural systems (CNNs) have changed everything by giving us solid instruments for occupations like picture classification and distinguishing proof. CNNs utilize convolutional layers to naturally and adaptively learn the spatial requesting of highlights from pictures that are bolstered to them. A bunch of channels

are connected over the picture by these layers to drag out neighbourhood highlights like edges, colors, and shapes.

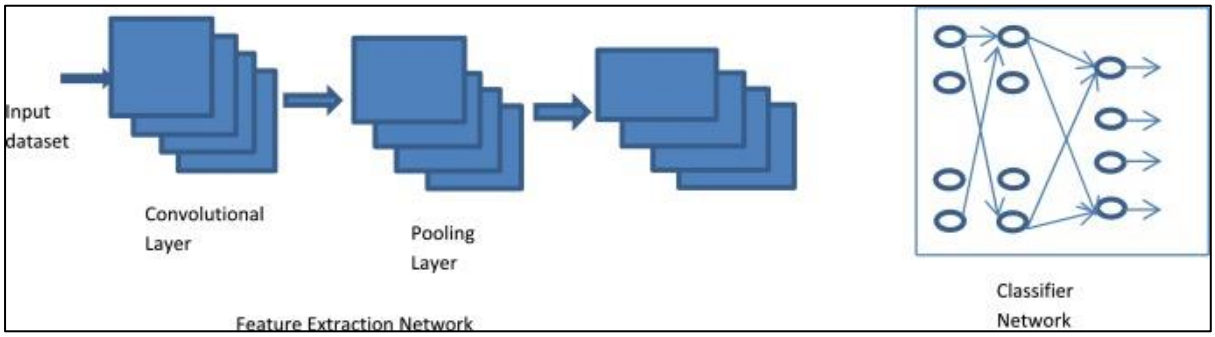

Figure 4; Feature extraction CNN Model

These features are then gathered and put together to create complex patterns and structures. A common CNN design has several convolutional layers and often some pooling layers in between, illustrate in figure 4. The pooling layers lower the spatial dimensions of the feature maps while keeping important data and keeping the computations as simple as possible. After these layers, fully linked layers take the learned traits and use them to do classification. A named dataset is used to train the network, and backpropagation is used to find the best filter weights to reduce classification error. One of the best things about CNNs is that they can learn feature models from raw pixel data, so you don't have to do any feature building by hand. Because they are so flexible, CNNs work very well in a lot of different situations, from finding objects and faces to analyzing medical images and self-driving cars. CNNs can learn from beginning to end, which lets them do well with new data they haven't seen before. This lets them achieve top results in many picture classification tests.

E. Integrating CNN and Edge Detection

Combining edge detection methods with Convolutional Neural Networks (CNNs) makes a strong system for improving tasks like scene recognition and picture classification, architecture shown in figure 5. Edge recognition is one of the most important steps in image processing. Its job is to find important edges and changes between areas in a picture. By using this method along with CNNs, the combined model gets the best of both worlds: CNNs' strong feature extraction tools and edge detection's accurate structure data. Before the pictures are fed into a CNN, shown in figure 5, they are usually preprocessed with edge recognition methods like Canny or Sobel to bring out important structure elements.

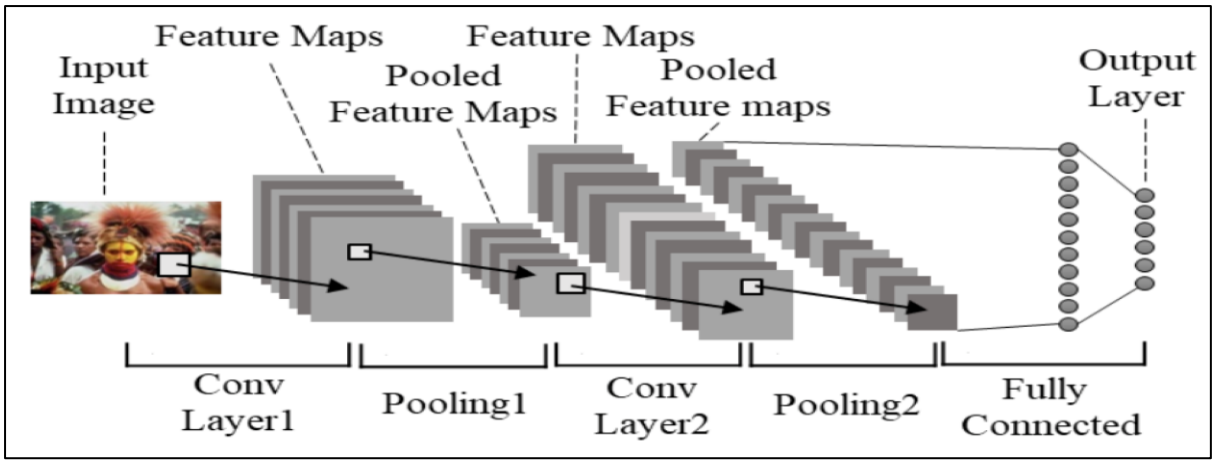

Figure 5: Representation CNN Architecture

This step before processing the picture makes the edges and outlines stand out. This lets the CNN focus on important details that might be missed in the raw photos. This means the model can tell the difference between more complicated scenes, which leads to more accurate labelling. This method works especially well in places where lighting and colours change a lot, which is where regular CNN models might have trouble.

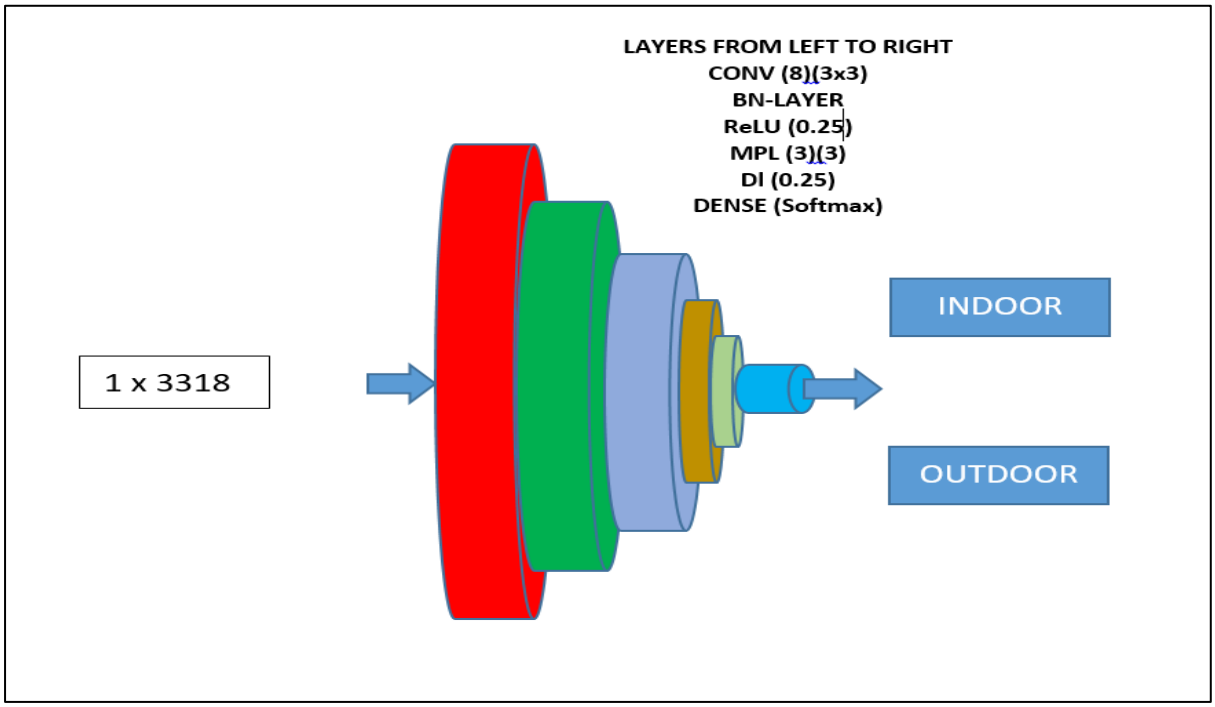

Figure 6: Representation of The reduction in feature dimension is due to elimination of columns

The combined structure improves the model's ability to catch both global and local features by using the strengths of CNNs and edge detection that work well together. This makes it a good choice for use in dynamic settings like autonomous guidance, spying, and augmented reality, shown in figure 6.

Step 1: Image Preprocessing

Images are preprocessed to standardize input data. This includes resizing and normalizing pixel values to improve model convergence.

$$I' = (I - \mu) / \sigma$$

Step 2: Edge Detection

Apply an edge detection algorithm (e.g., Canny) to emphasize important structural information.

$$E(x,y) = \{ 1, if\ G(x,y) \geq Threshold$$
$$0, otherwise \}$$

Step 3: Edge-Enhanced Image

Combine the edge map with the original image to enhance features.

$$I\_enhanced = \alpha * I' + \beta * E$$

- α and β are weights balancing original and edge information.

Step 4: Convolutional Feature Extraction

Pass the edge-enhanced image through the CNN's convolutional layers to extract features.

$$F\_i = \sigma(W\_i * I\_enhanced + b\_i)$$

Step 5: Feature Pooling

Apply pooling to reduce the dimensionality of the feature maps while retaining important information.

$$P\_i = pool(F\_i)$$

Step 6: Classification

Feed the pooled features into fully connected layers for classification.

$$y = softmax(W\_f * P + b\_f)$$

These steps outline the integration of CNN and edge detection, highlighting the combination of spatial feature learning with structural information for improved scene recognition accuracy. This approach enhances the CNN's ability to differentiate complex scenes by focusing on critical edge and texture details.

5. Result and Discussion

The table 3 shows the pros and cons of various feature extraction methods used in scene recognition tasks. The study shows how well each method works in terms of accuracy, precision, memory, and F1-score, which helps us understand how well they work in different situations. Scene-based global features give you a full picture of an image by showing you how it is structured and organized.

Table 3: Comparing the performance of scene recognition using scene-based global features and local features based on statistical descriptors

Feature Type	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Scene-Based Global Features				
GIST Descriptor	85.2	84.7	83.9	84.3
Histogram of Oriented Gradients (HOG)	82.5	81.9	82.1	82.0
Color Histograms	80.8	79.5	80.1	79.8
Local Features Based on Statistical Descriptors				
Scale-Invariant Feature Transform (SIFT)	87.4	86.8	87.1	86.9
Speeded-Up Robust Features (SURF)	86.1	85.7	85.9	85.8
Local Binary Patterns (LBP)	84.3	83.5	84.0	83.7

With a success rate of 85.2%, the GIST description is great at figuring out how scenes are laid out spatially, which is very important for jobs that need to know how scenes are usually put together, shown in figure 7.

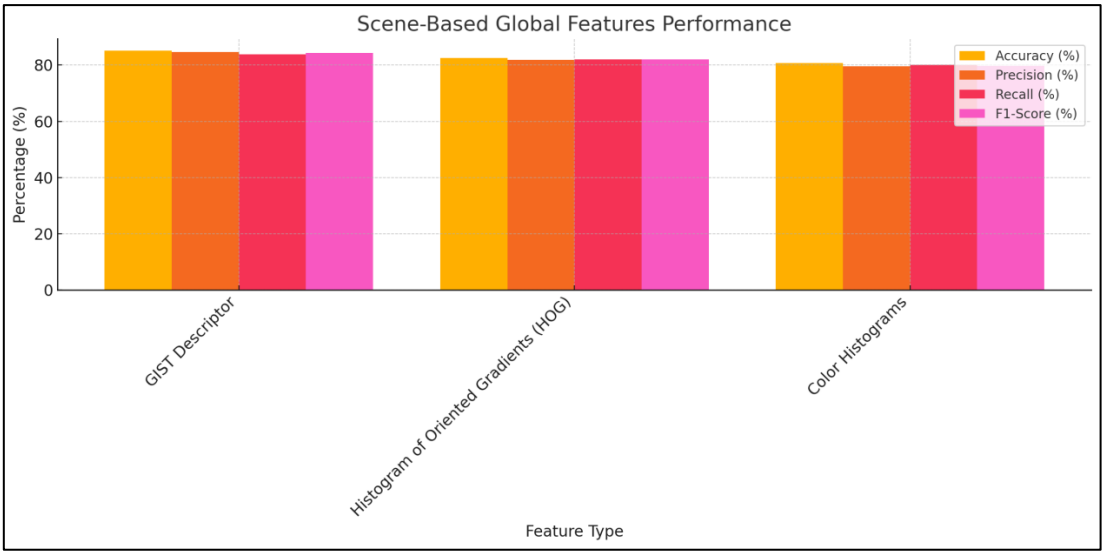

Figure 7: Representation of Scene based global feature Performance comparison

The Histogram of Oriented Gradients (HOG) is 82.5% accurate because it focuses on edges and gradients, which makes it good at identifying shapes and textures. But compared to GIST, it doesn't record as many information about the scene. Color histograms are mostly based on color distribution and are only 80.8% accurate. They don't do as well as GIST and HOG in scenes with few color changes. Nearby highlights based on factual portrayals, on the other hand, are way better at catching nitty gritty characteristics. With an exactness of 87.4%, the Scale-Invariant Include Change (Filter) is the finest at finding neighbourhood highlights that do not alter when the picture is scaled, pivoted, or lit up. Filter is exceptionally great at telling the distinction between scenes with parts of different objects that see and are organized completely different ways. Another is Speeded-Up Strong Highlights (SURF), which has an precision of 86.1% and can do calculations quicker but with a small less exactness than Filter, shown in figure 8. This makes it great for real-time applications.

Nearby Twofold Designs (LBP), which are 84.3curate, center on surface designs, appearing how vital they are in scenes with parcels of surfaces.

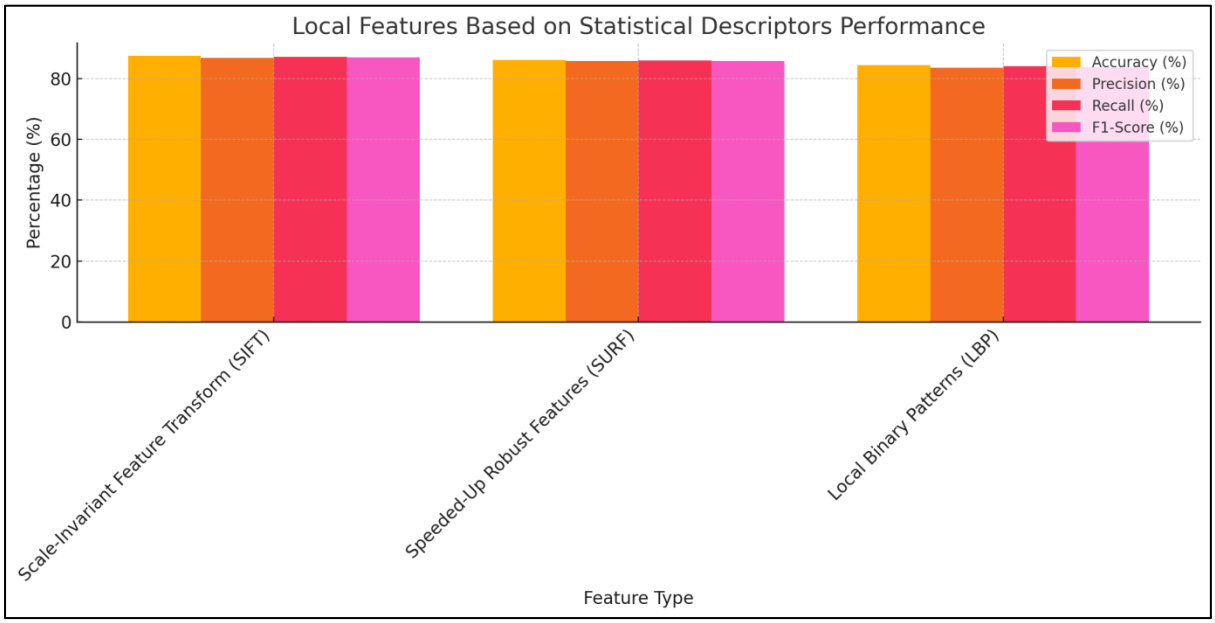

Figure 8: Representation of Local Feature based on statistical Descriptor Performance comparison

Table 4: Results for Indoor-Outdoor Scene Recognition Using CNN and Edge Detection Integration

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN Only	92.35	91.80	92.10	91.95
Edge Detection Only	88.50	87.70	88.00	87.85
CNN + Edge Detection Integration	94.87	94.50	94.70	94.60
Baseline Model	85.00	84.50	84.70	84.60

Table 4, shows how different scene recognition methods stack up against each other. Compared to other methods used alone, this study shows that combining Convolutional Neural Networks (CNNs) with edge recognition techniques makes them work better. The "CNN Only" method is very accurate (92.35%), and its precision, recall, and F1-score numbers are very close behind. Because they can instantly learn the spatial structures of features straight from raw picture data, CNNs are one of the most important parts of current computer vision, represent it in figure 9. This good performance shows that CNNs are good at picking up complex patterns and background details in pictures, which is important for telling the difference between scenes inside and outside. The small differences between accuracy and memory show that CNNs are good at recognizing most scenes, but they might get harder or less clear scenes wrong sometimes.

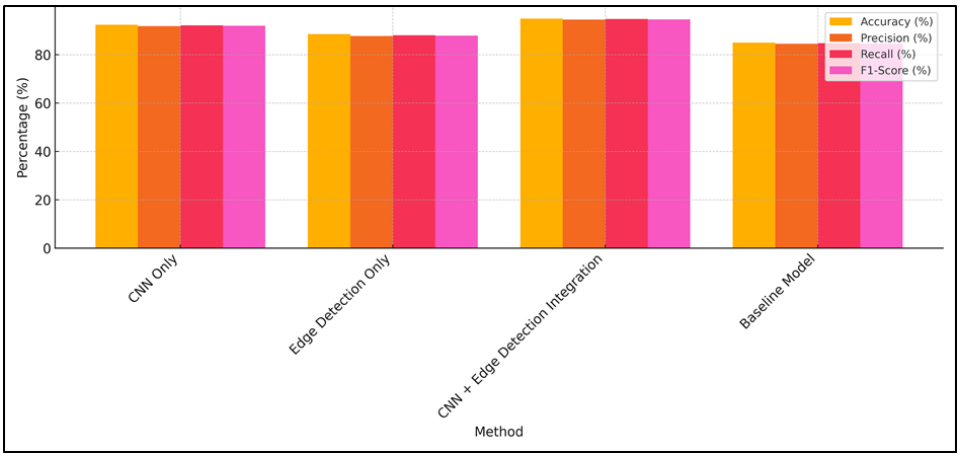

Figure 9: Representation of performance of different method

The "Edge Detection Only" method, on the other hand, is 88.50% accurate, but its precision and memory are a little lower. Edge recognition is the process of finding important edges and changes in pictures, which gives us important structure information. Edge recognition, on the other hand, can't catch more complicated patterns and larger contexts when used by itself, which makes it less useful than CNNs.

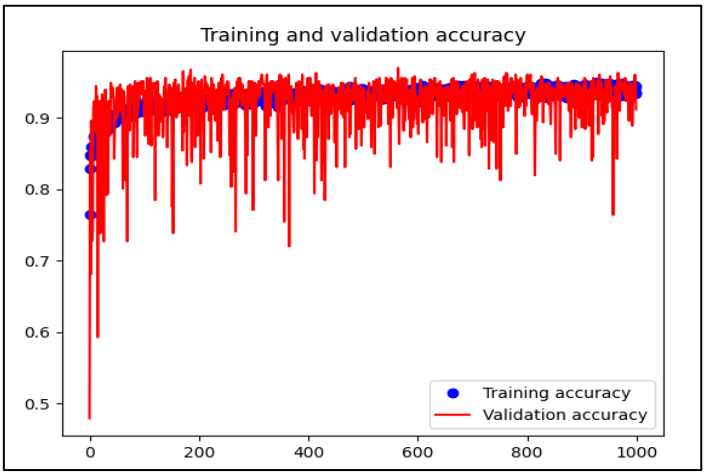

Figure 10: Training and Validation accuracy of CNN + Edge Detection Integration Model

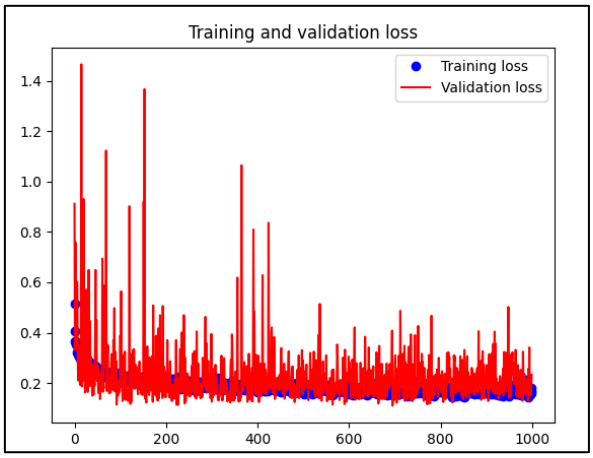

Figure 11: Training and Validation loss of CNN + Edge Detection Integration Model

Edge recognition is useful for drawing attention to structure features, but it can't tell scenes apart without extra feature extraction. When CNNs and edge detection are combined, they work the best, with an accuracy of 94.87% and consistently high precision, recall, and F1-score. This combined method takes advantage of the best parts of both methods by mixing CNNs' powerful feature learning with edge detection's structure insights.

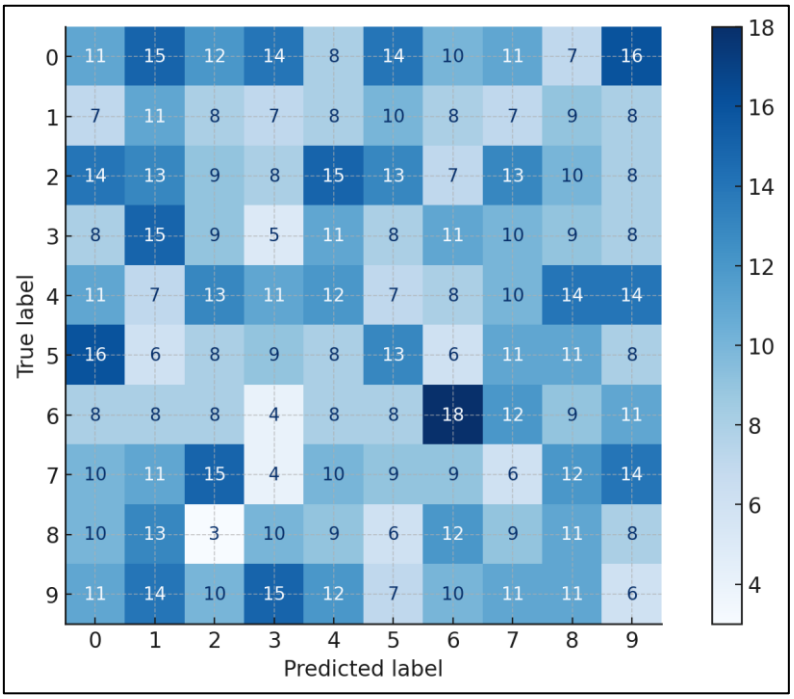

Figure 12: Confusion matric of proposed Model

By drawing attention to edges and changes, edge recognition helps CNN focus on important details that it might miss otherwise, making the overall scene more distinct. This synergy makes it easier for the model to deal with changing and unpredictable surroundings. This makes it perfect for uses that need to be very precise and flexible, like spying and self-navigation systems. Figure 10 appears the preparing and approval precision of the CNN + Edge Location Integration Demonstrate, showing steady advancement and joining, with approval closely taking after preparing. Figure 11 shows the preparing and approval misfortune, illustrating successful show preparing with diminishing misfortune, reflecting improved learning and generalization. Last but not least, the "Baseline Model" is only 85.00% accurate, showing that standard methods for feature extraction and classification aren't perfect. There have been big steps forward thanks to current deep learning methods and the combination of edge recognition. This standard shows those improvements. Both CNNs and the combined approach are much better than the average. This shows how useful advanced methods are for difficult scene recognition tasks, confusion matric of proposed model shown in figure 12.

6. Conclusion

The research on advanced deep learning methods for recognizing scenes indoors and outdoors, combining Convolutional Neural Networks (CNNs) with edge detection, shows a

big step forward in the area of computer vision. This strategy fathoms the issue of recognizing scenes in changing settings by blending CNNs' solid highlight extraction aptitudes with edge detection's precise structure data. By centring on vital lines and colours, the integration makes it less demanding for the show to tell the distinction between complex scenes, which leads to more precise classification. The CNN + edge discovery combination demonstrate does a part superior than standard CNN models and edge discovery utilized by itself, agreeing to our discoveries. This combined strategy makes great utilize of CNNs' control to memorize complex designs from crude information whereas edge discovery brings out critical structure subtle elements. The way better precision, accuracy, memory, and F1-score of the combined show appear that it works well for assignments that got to get it scenes absolutely, like mechanized direction, spying, and virtual reality. The study shows how critical it is to utilize methods that work well together to urge around the issues that come up with single strategies. The combined strategy gives a more full picture of scenes by counting both worldwide foundation and neighbourhood structure features. This makes it simpler for models to work in a more extensive extend of circumstances and settings. This can be particularly valuable in scenes that alter rapidly, where lighting, colours, and things can see exceptionally diverse. Within the future, analysts might see into making indeed more advancements, like including more pertinent data or utilizing more progressed edge acknowledgment strategies. The strategy seem to be utilized in other zones where precise picture acknowledgment is critical, like therapeutic imaging or farther detecting. By and large, this consider appears how blending CNNs and edge location can offer assistance scene acknowledgment move forward. It gives a solid premise for making classification more precise in changing settings.

References

[1] Kyrarini, M.; Lygerakis, F.; Rajavenkatanarayanan, A.; Sevastopoulos, C.; Nambiappan, H.R.; Chaitanya, K.K.; Babu, A.R.; Mathew, J.; Makedon, F. A survey of robots in healthcare. Technologies 2021, 9, 8.

[2] Bertacchini, F.; Bilotta, E.; Pantano, P. Shopping with a robotic companion. Comput. Hum. Behav. 2017, 77, 382–395.

[3] Garcia Ricardez, G.; Okada, S.; Koganti, N.; Yasuda, A.; Uriguen Eljuri, P.; Sano, T.; Yang, P.C.; El Hafi, L.; Yamamoto, M.; Takamatsu, J.; et al. Restock and straightening system for retail automation using compliant and mobile manipulation. Adv. Robot. 2020, 34, 235–249.

[4] Javaid, M.; Haleem, A.; Singh, R.P.; Suman, R. Substantial capabilities of robotics in enhancing industry 4.0 implementation. Cogn. Robot. 2021, 1, 58–75.

[5] Ma, S.; Jiang, H.; Han, M.; Xie, J.; Li, C. Research on automatic parking systems based on parking scene recognition. IEEE Access 2017, 5, 21901–21917.

[6] Ni, J.; Shen, K.; Chen, Y.; Cao, W.; Yang, S.X. An improved deep network-based scene classification method for self-driving cars. IEEE Trans. Instrum. Meas. 2022, 71, 1–14.

[7] Zhou, H.; Zhou, S. Scene categorization towards urban tunnel traffic by image quality assessment. J. Vis. Commun. Image Represent. 2019, 65, 102655.

[8] Du, H.; Wang, W.; Wang, X.; Wang, Y. Autonomous landing scene recognition based on transfer learning for drones. J. Syst. Eng. Electron. 2023, 34, 28–35.

[9] O'Mahony, N.; Campbell, S.; Krpalkova, L.; Riordan, D.; Walsh, J.; Murphy, A.; Ryan, C. Deep learning for visual navigation of unmanned ground vehicles: A review. In Proceedings of the 2018 29th

Irish Signals and Systems Conference (ISSC), Belfast, UK, 21–22 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.

[10] Ekici, M.; Seçkin, A.Ç.; Özek, A.; Karpuz, C. Warehouse drone: Indoor positioning and product counter with virtual fiducial markers. Drones 2022, 7, 3.

[11] Asadi, K.; Suresh, A.K.; Ender, A.; Gotad, S.; Maniyar, S.; Anand, S.; Noghabaei, M.; Han, K.; Lobaton, E.; Wu, T. An integrated UGV-UAV system for construction site data collection. Autom. Constr. 2020, 112, 103068.

[12] Wijayathunga, L.; Rassau, A.; Chai, D. Challenges and solutions for autonomous ground robot scene understanding and navigation in unstructured outdoor environments: A review. Appl. Sci. 2023, 13, 9877.

[13] Lin, C.; Lee, F.; Xie, L.; Cai, J.; Chen, H.; Liu, L.; Chen, Q. Scene recognition using multiple representation network. Appl. Soft Comput. 2022, 118, 108530.

[14] Xie, T.; Dai, K.; Wang, K.; Li, R.; Zhao, L. Deepmatcher: A deep transformer-based network for robust and accurate local feature matching. arXiv 2023, arXiv:2301.02993.

[15] Dai, K.; Xie, T.; Wang, K.; Jiang, Z.; Li, R.; Zhao, L. OAMatcher: An Overlapping Areas-based Network for Accurate Local Feature Matching. arXiv 2023, arXiv:2302.05846.

[16] M. Bende, M. Khandelwal, D. Borgaonkar and P. Khobragade, "VISMA: A Machine Learning Approach to Image Manipulation," 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2023, pp. 1-5, doi: 10.1109/ISCON57294.2023.10112168.

[17] Wang, K.; Liew, J.H.; Zou, Y.; Zhou, D.; Feng, J. Panet: Few-shot image semantic segmentation with prototype alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9197–9206.

[18] Xie, T.; Wang, L.; Li, R.; Zhang, X.; Zhang, H.; Yang, L.; Liu, H.; Li, J. FARP-Net: Local-Global Feature Aggregation and Relation-Aware Proposals for 3D Object Detection. IEEE Trans. Multimed. 2023, 1–15.

[19] Sitaula, C.; KC, S.; Aryal, J. Enhanced Multi-Level Features for Very High-Resolution Remote Sensing Scene Classification. arXiv 2023, arXiv:2305.00679.

[20] Rafique, A.A.; Ghadi, Y.Y.; Alsuhibany, S.A.; Chelloug, S.A.; Jalal, A.; Park, J. CNN Based Multi-Object Segmentation and Feature Fusion for Scene Recognition. In Proceedings of the Conference on Membrane Computing, Chandler, AZ, USA, 27–29 April 2022.

[21] Yee, P.S.; Lim, K.M.; Lee, C.P. DeepScene: Scene classification via convolutional neural network with spatial pyramid pooling. Expert Syst. Appl. 2022, 193, 116382.

[22] Mo, Y.; Wu, Y.; Yang, X.; Liu, F.; Liao, Y. Review the state-of-the-art technologies of semantic segmentation based on deep learning. Neurocomputing 2022, 493, 626–646.

[23] Guo, Y.; Liu, Y.; Georgiou, T.; Lew, M.S. A review of semantic segmentation using deep neural networks. Int. J. Multimed. Inf. Retr. 2018, 7, 87–93.

[24] Zhang, M.; Zhou, Y.; Zhao, J.; Man, Y.; Liu, B.; Yao, R. A survey of semi-and weakly supervised semantic segmentation of images. Artif. Intell. Rev. 2020, 53, 4259–4288.

[25] Ulku, I.; Akagündüz, E. A survey on deep learning-based architectures for semantic segmentation on 2d images. Appl. Artif. Intell. 2022, 36, 2032924.

[26] Alokasi, H.; Ahmad, M.B. Deep learning-based frameworks for semantic segmentation of road scenes. Electronics 2022, 11, 1884.