JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 08 Issue 01
2024

Research Article
https://doi.org/10.56979/801/2024

Unveiling Complex Scenes: A Deep Belief Network and Semantic Segmentation Approach

Adnan Ahmed Rafique[1*], and Yasir Javaid[2]

[1]Department of Computer Sciences & Information Technolgy, University of Poonch Rawalakot, 12350, Pakistan.
[2]Department of Informaiton and Communcation Technolgy, Government College of Technology, Rawalakot, 12350, Pakistan.
[*]Corresponding Author: Adnan Ahmed Rafique. Email:adnanrafique@upr.edu.pk

Received: April 21, 2024 **Accepted:** September 26, 2024

Abstract: Scene classification is a meaningful and challenging research field in computer vision due to the wide variety of objects present in a scene, their internal relationships, and inter-class similarities**.** Thus, complex scene recognition and understanding are needed in various applications, including virtual reality-based scene integration, robotics, autonomous driving, and tourist guide systems. Therefore, a novel scene recognition system that integrates various components to recognize the scenes in complex imagery is developed. Compared to the state-of-art systems, our system combines many significant features for improving the classification accuracy. Initially, the images are acquired and preprocessed. It is worth mentioning that semantic segmentation approaches are powerful as they not only detect objects present in an image but recognize the boundaries of each object. To leverage the effectiveness of semantic segmentation, we propose a modified fuzzy C-means (MFCM) segmentation method that partitions the image into various objects to label the pixels according to different segmented objects. Then, convolutional neural network (CNN) features, the dynamic geometrical (GF), and, blob features (BF) are extracted and fused for further analysis to recognize the scene through a deep belief network (DBN). The latter incorporates a genetic algorithm that optimizes the number of hidden units based on the error rate and the training time. The effectiveness of the proposed system is validated over Pattern Analysis, Statistical Modeling, and Computational Learning Visual Object Classes (PASCAL VOC 2012) and the Microsoft Research Cambridge (MSRC) datasets by achieving 93.30% and 92.53% recognition accuracies respectively.

Keywords: Blob Features; Deep Belief Network; Feature Fusion; Fuzzy C-Means; Scene Recognition.

1. Introduction

Deep learning has become a powerful tool for computer vision processes including scene recognition. However, scene recognition still faces many challenges due to the complexity and diversity of real-world scenes, including variations in light intensity, viewpoint, and object appearance. Semantic segmentation is an effective technique for understanding the scene, which labels each pixel in an image according to its semantic meaning. By segmenting an image into semantic regions, the model can capture the spatial relationships between different objects and their context, which is essential for scene recognition. Moreover, feature fusion is an important step for integrating different sources of information, such as visual and semantic features, to improve the recognition accuracy. The ability to accurately recognize scenes in images and videos is a crucial responsibility with a multitude of practical applications in the real world within the field of computer vision. To achieve this goal, researchers have proposed various approaches, including deep learning-based methods that can automatically learn useful representations of visual features from large-scale datasets. Among these methods, DBNs [1] have gained increasing attention due to their capacity to output complex hierarchical representations from raw input data.

This paper introduces a DBN architecture that combines semantic segmentation and feature fusion to outperform the state-of-the-art performance on scene recognition. Specifically, our approach first performs

semantic segmentation on the input image to obtain a set of semantic regions. Then, the model extracts both visual and semantic features from each region and fuses them using a novel feature fusion strategy. Finally, the fused features are handled by a fully connected layer for scene classification. We have evaluated our approach on several benchmark datasets and compared it with other state-of-the-art methods. The obtained results have demonstrated that our DBN-based approach achieves superior performance relating to accuracy and robustness. Overall, this work provides a promising direction for scene recognition research by leveraging the power of deep learning, semantic segmentation, and feature fusion.

Our key contributions are as follows:

- A novel approach namely, modified fuzzy C-means (MFCM) segmentation is presented and implemented for semantic segmentation of complex images with dynamic environments.
- The proposed algorithm aims to obtain dynamic geometrical features by identifying the object's key points and linking them to generate various geometric shapes for the purpose of feature extraction.
- CNN and machine learning features including blob extraction are extracted and fused to effectively recognize the scenes with a comprehensive feature set.
- To recognize a scene using feature fusion, a DBN is employed.

The rest of the article has been structured as follows: Section II presents the methodology of the proposed system and the architecture of the model in detail. Section III illustrates experimental results and datasets utilized and compares existing techniques. Finally, in section IV, the conclusion and future work are discussed.

2. Literature Review

Several researchers have employed conventional systems for investigating scene comprehension and categorization. These conventional systems compute various features to identify objects and categorize scenes. L. Zhang et al. [2] proposed a novel approach for scene classification based on learning object-to-class kernels. The authors have addressed the limitations of existing methods which either rely on handcrafted features or involve a large number of parameters in kernel-based methods. The proposed approach learns the object-to-class kernels from the data and uses them for scene classification. The method involves a joint optimization of the kernel weights and the classifier parameters. The authors have conducted experiments on several benchmark datasets, including Caltech-101 and Scene-15, and have demonstrated that the proposed method outperforms the state-of-the-art performance in scene classification. Thus, the paper presents a promising approach for scene classification, which reduces the reliance on handcrafted features and achieves high accuracy using a relatively small number of parameters. Their framework has achieved an accuracy of 88.80% for the LS dataset. The proposed method has potential applications in areas such as image retrieval and object recognition.

X. Song et al. [3] presented a method for scene recognition that combines multiple visual features and spatial information. The proposed method constructs a semantic manifold by embedding the visual features into a common space, which allows for effective fusion of different features. Spatial context is also considered in the model by incorporating local and global spatial information. The proposed method outstands the performance of the state-of-the-art techniques on three benchmark datasets for scene recognition. The results presented in [3] demonstrate the effectiveness of combining multiple features and spatial information for scene recognition tasks. In [4], R. Kachouri et al. proposed an unsupervised image segmentation method that combines local pixel clustering and low-level region merging techniques. The proposed method clusters pixels based on local similarity and then groups cluster into coherent segments using low-level feature similarity between adjacent clusters. The method achieves competitive performance with other unsupervised segmentation methods while being computationally efficient, as shown by experiments on several benchmark datasets. The paper provides a comprehensive overview of unsupervised image segmentation techniques, highlighting the strengths and limitations of different approaches. More particularly, the authors [4] have used visual accuracy and Liu factors to evaluate their technique.

In [5] H. Zhao et al. proposed a Pyramid Scene Parsing Network (PSPNet) for semantic scene segmentation, which utilizes global context information effectively. The PSPNet adopts a pyramid pooling module that generates feature maps of different scales, which are fused with the original feature map through concatenation. The performance of the proposed network is superior to the performance of the

state-of-the-art methods on four challenging scene-parsing benchmark datasets, and extensive experiments show the effectiveness of the proposed pyramid pooling module in capturing global context information for semantic segmentation. The paper provides a comprehensive review of related work on semantic segmentation and global context modeling, highlighting the limitations of previous methods and the advantages of the proposed PSPNet. The technique described in [5] has provided an accuracy of 85.4% for the Pascal VOC 2012 dataset. In [6] N. Hussain et al. presented a novel deep learning and classical features-based scheme for object recognition, which is has been applied to machine inspection. The proposed scheme uses a DNN to extract high-level features and a set of classical features, including texture, color, and shape, to capture low-level features. The classical features are fused with the DNN features to improve the recognition accuracy. The proposed scheme is evaluated on a machine inspection dataset, and the results show that it outperforms several state-of-the-art object recognition methods. Paper [6] provides a comprehensive literature review on object recognition and machine inspection, highlighting the limitations of previous methods and the advantages of the proposed scheme. The proposed scheme has the potential to be applied in various applications, such as quality control and defect detection in manufacturing processes.

S. Xia et al. [7] discussed a weakly supervised attention map (WS-AM) for scene recognition, which can localize the discriminative regions in an image without the need for pixel-level annotations. The proposed method first generates a coarse attention map by training a CNN with image-level labels and then refines the attention map by iteratively training the network with attention-guided feature maps. The refined attention map is used to localize the discriminative regions in the image and improve the scene recognition accuracy. The proposed method has been evaluated on various benchmark datasets, and the results show that it outperforms several state-of-the-art weakly supervised methods while being computationally efficient. Paper [7] provides a comprehensive literature review on weakly-supervised methods for scene recognition, highlighting the limitations of previous methods and the advantages of the proposed WS-AM. The proposed method has the potential to be applied in various applications, such as image and video retrieval. The work presented in [7] has considered three benchmark datasets including MIT Indoor 67, Scene 15, and UIUC Sports. The results obtained in [7] demonstrate that the proposed method selects fewer local regions.

In [8], Jun Chu et al. have introduced a unique object detection framework for small and occluded objects. The framework is a combination of two techniques: multi-layer convolution feature fusion (MCFF) and online hard example mining (OHEM). MCFF is a technique for fusing features from different layers of a CNN. This helps to improve the representation of objects in the network, which can be helpful for detecting small and occluded objects using KITTI dataset. In [9], Y. Zhand et al. have presented the instance segmentation task in the context of flexible vision sensors and visual sensor networks. The authors have proposed a new method called Mask-Refined R-CNN (MR R-CNN) to overcome the limitations of Mask R-CNN. They have identified that the scale-invariant fully convolutional network structure of Mask R-CNN did not handle spatial differences among receptive fields having different sizes. This leads to the misclassification of pixels at object edges, affecting instance detail prediction. To address this issue, the stride of ROIAlign (region of interest align) is modified, and MR R CNN applies feature fusion by replacing the original fully convolutional layer with a new semantic segmentation layer. The authors of [9] have used MS COCO for experiments.

Furthermore, the idea of feature extraction and fusion has been adopted for palm print recognition. The authors of [10] have fused left and right palm prints while considering features with high discrimination. Another interesting study [11] has also applied adaptive selection and weighting of features for palm print protection. Thus, the scheme described in [11] is based on correlation from 2D feature representations to improve the verification performance. Various authors have proposed methods and techniques for scene understanding and recognition. These techniques have been evaluated on various benchmark datasets, highlighting their potential in different applications.

Table 1. Overview of the various studies with datasets, methods used and their limitations.

Reference	Author	Dataset used	Contribution/Method used	Limitations

[2]	L. Zheng et al.	Caltech-101, Scene-15	Learning object-to-class kernels, joint optimization	Reduces reliance on handcrafted features, achieves high accuracy with fewer parameters
[3]	X. Song et al.	Scene-15, Scene-12, Caltech-256	Fusion of multiple visual features, spatial information	Effective combination of features and spatial information for solving scene recognition
[4]	R. Kachouri et al.	Corel, Barkeley datasets	Unsupervised image segmentation using local pixel clustering	Competitive performance, computational efficiency
[5]	H. Zhao et al.	Pascal VOC 2012, Cityscapes	Pyramid Scene Parsing Network (PSPNet) with global context modeling	State-of-the-art semantic scene segmentation, effective global context modeling
[6]	N. Hussain et al.	MNIST, CIFAR-10, ImageNet ILSVRC 2012	Deep learning and classical features-based object recognition	Improved recognition accuracy by fusing classical and DNN features
[7]	S. Xia et al.	MIT Indoor 67, Scene 15, UIUC Sports	Weakly supervised attention map (WS-AM) for scene recognition	Discriminative region localization without pixel-level

annotations

outperforms weakly

supervised methods

3. Materials and Methods

This section briefly introduces a novel framework for scene recognition in complex scenarios. Figure 1 depicts a high-level overview of the proposed model. The system architecture is divided into multiple phases, including (i) segmentation, (ii) feature extraction, (iii) feature Fusion, and (iv) scene recognition.

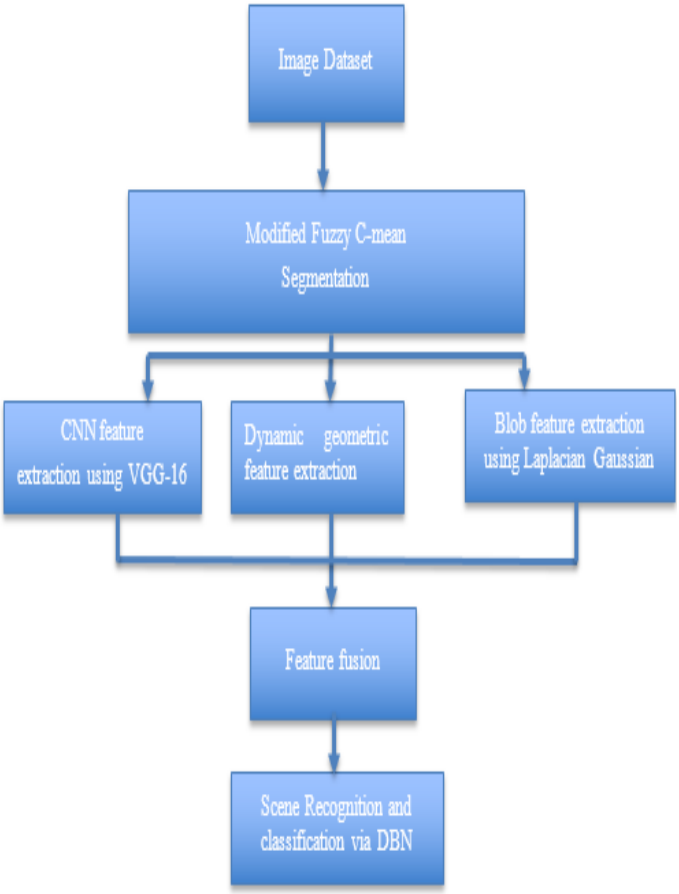

Figure 1. System Architecture of Proposed Model

3.1. Pre-processing

After data acquisition, the first step is to prepare the data for analysis. The information gathered for this study has been manipulated statistically and geometrically. When saving an image, the RGB format is used in order to process and remove any unwanted particles in the image. An image's artifacts and noise can be filtered out using digital filters.

3.2. Modified Fuzzy C-Means Segmentation

MFCM is a variant of FCM that introduces modifications to better handle complex images with dynamic environments, particularly for semantic segmentation tasks. The modification in MFCM involves adapting the cluster weight reduction strategy based on minimizing the error in the objective function, allowing for more effective segmentation of spatially structured data. During segmentation, similar areas are separated into their constituent parts. Figure 2 illustrates the MFCM segmentation results over PASCAL VOC 2012 dataset.

MFCM is a fuzzy-based [12] segmentation method. In this method, pixels are used as data points to find pairs of similar parts. In fuzzy logic, pixels do not belong to one particular group but are instead categorized into multiple different categories. By iteratively minimizing the objective function, the FCM has an effect on the image. Using a cluster weight reduction strategy based on minimizing the square of

the error in the objective function $A_N(P, Q)$, useful clusters can be created. An expression for the objective function is as follows:

(a) (b) (c)

Figure 2. Semantic segmentation (a) original image, (b) grayscale image, and (c) segmented image

$$A_N(P,Q) = \sum_{i=1}^{c} \sum_{j=1}^{n} p_{ij}^{r} \left| x_j - q_i \right|^2 \quad (1)$$

where n, r are data points to represent the i^{th} cluster in real numbers; c is to represent clusters, and p_{ij}^{r} describes the membership of x_j pixels in the i^{th} cluster, the centroid of the cluster is symbolized as q_i:

$$p_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{|x_j - q_i|}{|x_j - q_k|^2} \right)^{\frac{1}{r-1}}} \quad (2)$$

$$p_{ij} \in [0,1], for\ i = [1, \ldots, c] \quad (3)$$

$$q_i = \frac{\sum_{j=1}^{n} p_{ij}^{r} x_j}{\sum_{j=1}^{n} p_{ij}^{r}} \quad (4)$$

where the distance between the pixel to the cluster's center is computed via P and Q $A_N(P,Q)$. Usually, the MFCM technique is computationally complex as it analyses the spatial values along with each iteration. These spatial values determine the distance between the cluster center and each pixel. The distance and the membership value are indirectly proportional to each other. For instance, if the distance between the cluster center and the pixel is large, the membership value will be small, and vice versa. Consequently, the computation for all the adjacent pixels significantly increases.

The modification in MFCM aims to enhance the original FCM algorithm by considering spatial relationships and dynamic environments in image segmentation tasks.

Unlike standard FCM, which may not effectively capture spatial information, MFCM incorporates spatially structured data into the clustering process, leading to improved segmentation results.

Alternatively, we can also compare it Mathematically, MFCM adjusts the objective function to account for spatial relationships among data points and cluster centroids, resulting in more accurate segmentation of complex images.

The Modified Fuzzy C-Means (MFCM) algorithm builds upon the classical Fuzzy C-Means (FCM) by introducing modifications to the objective function to enhance the clustering process, particularly for spatially structured data. Here's how the mathematical foundation for MFCM could be expressed, incorporating the details from our earlier discussion:

In classical FCM, the objective function JFCMJFCMis given by:

$$J_{FCM} = \sum_{i=1}^{n} \sum_{j=1}^{c} (W_{ij})^m \cdot \left\| x_i - v_j \right\|^2 \quad (5)$$

where n is the number of data points, c is the number of clusters, (W_{ij}) is the membership degree of the data point x_i to the cluster v_j, m is a fuzziness exponent, controlling the degree of cluster fuzziness, $\left\| x_i - v_j \right\|^2$ represents the squared Euclidean distance between the data point x_i and the cluster centroid v_j.

The membership degrees are updated by using the following formula:

$$W_{ij} = \left(\sum_{k=1}^{c} \left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{\frac{2}{m-1}} \right)^{-1} \quad (6)$$

By using the above mathematical formula, the membership degree based on the relative distances from the data point to all cluster centroids are adjusted, emphasizing the data points' closer centroids.

MFCM extends the objective function of FCM by adding a regularization term that incorporates an additional error term (,). The modified objective function, MFCM, is:

$$J_{MFCM} = \sum_{i=1}^{n} \sum_{j=1}^{c} (W_{ij})^m . \left\| x_i - v_j \right\|^2 + \lambda . A N (P, Q) \quad (7)$$

where λ is a regularization parameter that adjusts the impact of the $A N (P, Q)$ term on the overall objective function, $A N (P, Q)$ is an error term designed to address specific clustering challenges, such as minimizing errors due to overlapping clusters or spatial separations that are not well-captured by traditional methods.

The introduction of $A N (P, Q)$ allows MFCM to handle more complex data distributions and structures, making it particularly useful for datasets where spatial relationships or other contextual factors are important. By adjusting cluster weights and taking into account additional error considerations, MFCM can achieve more accurate and contextually relevant clustering than the standard FCM algorithm.

Here are some key differences between FCM and MFCM for understanding. FCM is a more general-purpose clustering algorithm, while MFCM is a more specialized clustering algorithm that is designed for data that is spatially structured. MFCM is also less computationally expensive than FCM, which makes it a more attractive option for large datasets. The modification in MFCM aims to enhance the original FCM algorithm by considering spatial relationships and dynamic environments in image segmentation tasks. Unlike standard FCM, which may not effectively capture spatial information, MFCM incorporates spatially structured data into the clustering process, leading to improved segmentation results.

By integrating spatial information, MFCM minimizes the effects of noise and improves the coherence of the segmented regions, which is crucial for accurate scene recognition. Additionally, MFCM adjusts the objective function to account for spatial relationships among data points and cluster centroids, resulting in more accurate segmentation of complex images. The theoretical background and comparative analysis between FCM and MFCM provide a clear understanding of how the modification enhances the original algorithm and its impact on semantic segmentation. By incorporating spatial information and adapting the clustering strategy, MFCM offers improved segmentation results, particularly for complex images with dynamic environments. Table 2 demonstrates the key modifications of both clustering algorithms.

Table 2. Comparison of FCM and MFCM in terms of features

Feature	FCM	MFCM
Distance metric	Euclidean	Spatial
Optimization approach	Global	Local
Computational complexity	More computationally expensive	Less computationally expensive
Specialization	General-purpose	Spatially structured data

3.3. Features Extraction and Fusion

In this section, the unique features from various segmented objects are extracted. Different deep and machine learning feature extraction approaches are discussed and elaborated. These features are then fused to accurately recognize the scenes in the imagery.

3.3.1. CNN Features Extraction

CNN features [13] are extracted through a pre-trained CNN model i.e. VGG-16 [14]. This architecture normally uses ImageNet as a training dataset. The VGG-16 architecture has an input as well as an output layer while sixteen convolution layers are embedded to process the input data accordingly. All the images are converted to a size of 224x224x3 before being taken as input for the CNN model. Additionally, five max-pooling and three fully-connected layers are incorporated. a window size of 2 x 2 is used to maximize the results of pooling. When it comes to activation functions, ReLU is unquestionably the most prevalent, and as such, it is employed in the layers. To successfully extract useful CNN features, we employ a transfer learning approach. In order to make the model more useful than a newly designed

model, this approach makes use of the learned features. Figure 3 illustrates the architecture of CNN features extraction by using VGG-16.

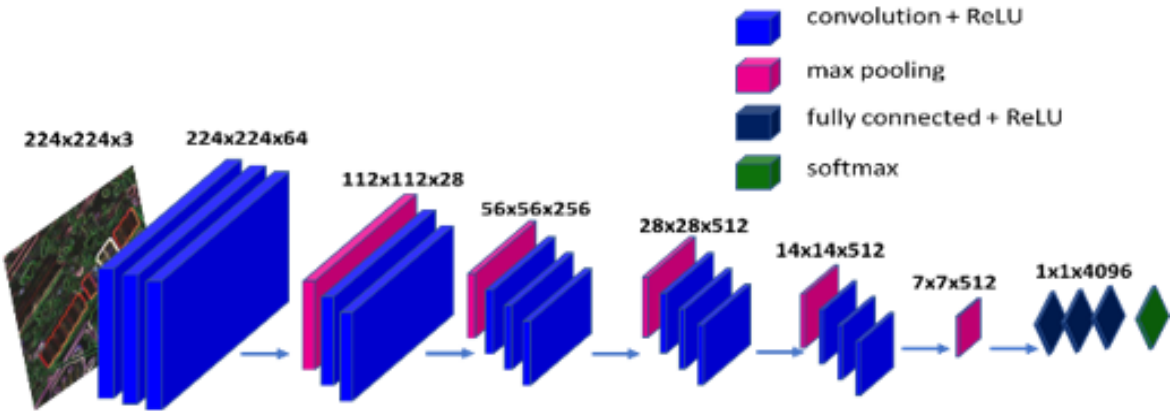

Figure 3. The architecture of VGG-16 for CNN features extraction over segmented objects

3.3.2. Dynamic Geometrical Features

The segmented objects undergo a local computation of geometrical attributes. At first, four corners are determined for each object that has been recognized and segmented. In addition, the upper and lower bounds, as well as the left and right extremes have been considered. The four local geometric aspects, including the Euclidean distance, the area of triangles generated by connecting extreme points, the perimeter of these triangles, and measuring the cosine for the angles estimated between the sides of a triangle, may be determined using these extrema. Algorithm 1 illustrates the process of dynamic geometrical feature extractions. This method significantly extends the feature extraction capabilities of typical scene recognition systems by focusing on the geometrical relationships within the image. By identifying key points and constructing geometric shapes (like triangles), the system can capture more nuanced information about the structure and layout of the scene. The step by step procedure for the dynamic geometrical feature extraction are described in Algorithm 1.

Algorithm 1: Dynamic geometric features computation

Input: Segmented objects

Output: Dynamic geometric features of objects.

```
[ROW COL] = size (seg_object) /* returns number of rows and number of col. in the object*/
/* Find these extrempoints */
[extreme_top , extreme_right, extreme_bottom, extreme_left, mid_pixel]=extremes(seg_object)
for 1: ROW    /* read all pixel's pix in the scene/image */ and mark spot points
    for 1: COL    /* Extract points*/
    if pix (u,v) = mid_pixel Then spot point on pix(u,v) end
  else-if pix (u,v) = ext_left Then spot point on pix (u,v) end
  else-if pix (u,v) = ext_right Then spot point on pix (u,v) end
  else-if pix (u,v) = ext_top Then spot point on pix (u,v) end
  else-if pix (u,v) = ext_bottom Then spot point on pix (u,v) end
  else Then pix (u,v) is other than extreme point pixels end
     end
  end
/* compute distance and draw lines between these extreme points by using an array of points
for all images */
ArrayPoints = [ext_top, ext_right, ext_bottom, ext_left]

mid_pixel;

for j = 1 : ArrayPoints []
     if (ArrayPoints [j] => 4) Then ArrayPoints [j+1]=0; end
    dist (ArrayPoints [j], ArrayPoints [j+1])   /*1st iteration show distance between extreme top
and right points */
  draw_line(ArrayPoints [j], ArrayPoints [j+1]) end
for j = 1 : ArrayPoints []
```

dist (ArrayPoints [j], mid_ pixel) /* 1st iteration show distance between extreme top and mid_ pixel */

draw_line(ArrayPoints [j], mid_ pixel)

end

return ext_left, ext_right, ext_top, ext_bottom, mid_ pixel, and distance between these points

Capturing the spatial relationships between objects through geometrical shapes allows for a deeper understanding of the scene, which is critical for distinguishing complex scenes where simple feature extraction methods might struggle. The use of extreme points (corners, edges) to generate geometric features adds a layer of precision in representing the structural attributes of objects within the scene.

3.3.3. Blob Features Extraction

A blob is defined a cluster of connected pixels belonging to a specific shape [15-16]. Given an image, the morphological closing operation is utilized for pixel clustering. This operation allows to detect an object as a set of blobs. Moe importantly, the Laplacian of Gaussian (LoG) is investigated by the most popular technique for blob detection. Hence, the convolution of an image I(u,v) with LoG is represented by the following equation:

$$g(u,v,t) = \frac{1}{2\pi t} e^{-\frac{u^2+v^2}{2t}} \quad (8)$$

to provide a scale space representation at a specific scale t $S(u,v;t)=g(u,v,t)*I(u,v)$. So, after the computation of the Laplacian operator

$$\nabla^2 S = S_{uu} + S_{vv} \quad (9)$$

Consequently, greatly positive responses will be provided for dark blobs with radii r=√2t. However, greatly negative responses will be provided for bright blobs with similar radii. The most important concern of the response, while applying at a single scale, is its dependency. Both the Gaussian kernel and the scale of the blobs are strongly correlated and dependent on each other. Therefore, to maintain the rationale for the blob feature extraction, this paper incorporates a multi-scale technique. Hence, a scale-normalized Laplacian operator is embedded as follows:

$$\nabla^2_{norm} S = t(S_{uu} + S_{vv}) \quad (10)$$

3.3.4. Features Fusion

In this section, separately computed features, i.e., CNN features(F_{CNN}), Blob features (F_{Blob}), and Geometrical features (F_{Geo}) are fused. The feature vectors are normalized before fusion to ensure the consistency of the fused feature vector. A completely fused feature vector is created by combining the CNN, blob features, and geometrical features together after normalization as follows:

$$F_{Fused} = [F_{CNN}\ F_{Blob}\ F_{Geo}\ ] \quad (11)$$

3.4. Scene Recognition via DBN

This section describes the details of DBN as a classifier. The DBN architecture [17-19] is simple and composed of various layers, including input, hidden, and an output layer, as shown in Figure 4. The input layer takes the features set as an input; hidden layers are the actual processing units where a genetic algorithm (GA) is incorporated to optimize hidden units and compute the total epochs. While the output layer predicts the class label of each category included in the training process. The training process continues until one of the following conditions is achieved: a minimum gradient is approached, performance standard in terms of mean squared error is reached, or epoch upper bound is approached.

In order to optimize the hidden units, the fitness function plays an important role. As a result, a fitness function is developed to decrease the DBN's training time, and accordingly raise its accuracy. This paper considers the following fitness function:

$$Fitness_{Fun} = 1000 \times E + (T_T + T_B)/40 \quad (12)$$

where the misclassification rate is denoted by E, the training time of DNB before backpropagation T_T, and parameter tuning time by T_B during backpropagation. The smaller the error and time, the lower the level of fitness. The results of object detection by using DBN are presented in Figure 5.

The dimensions of a Deep Belief Network (DBN) when applied to the MSRC (Microsoft Research Cambridge) dataset are as follows:

The MSRC dataset consists of images with pixel-wise annotations for different semantic classes. Each image typically has dimensions of width x height, where width and height can vary across images. The number of semantic classes in the MSRC dataset is fixed and typically ranges from around 20 to 23 while we used only 15 classes for experiments

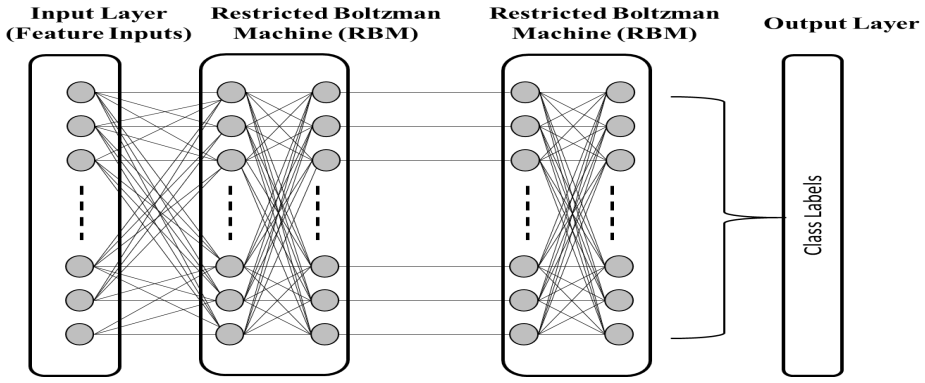

Figure 4. Classifier network architecture using DBN

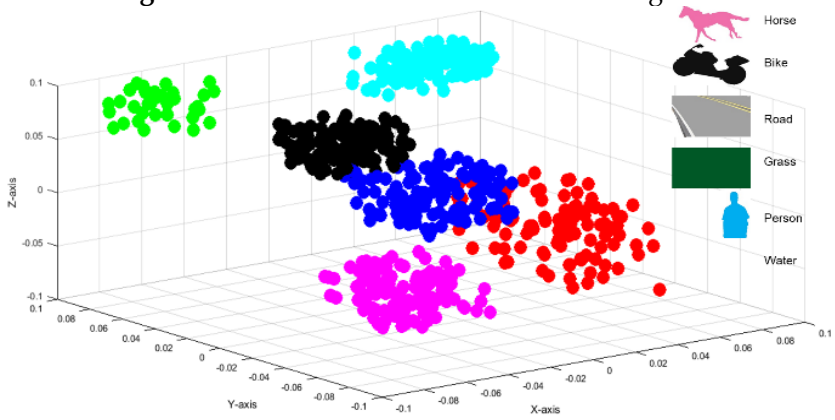

Figure 5. Scene recognition over PASCAL VOC 2012 dataset by applying the DBN classifier

Our model has used the same dimensions for the input layer as the dimensions of input images in the MSRC dataset, i.e., width x height which is resized as 320x320. The output layer of the DBN has dimensions corresponding to the number of semantic classes in the MSRC dataset. In our case, 15 classes have been evaluated, so the output layer has a dimension of 15. Similarly, when applied to PASCAL VOC 2012, the dimensions are used accordingly.

3.5. Impact of Fusion Technique over Scene Recognition

In the pursuit of advancing scene recognition capabilities, the integration of CNNs with distinctive machine learning techniques—specifically blob detection algorithms—culminates in a remarkable methodological breakthrough when fused with DBNs. The strategic incorporation of CNNs, renowned for their adeptness in hierarchical feature extraction from visual data, aligns symbiotically with blob detection methods, which are exemplary in identifying regions within an image that differ in characteristics, such as brightness or color, compared to surrounding regions.

The innovation lies in combining the strengths of both approaches. CNNs excel in capturing and learning features from raw images that are vital for identifying and classifying various elements within a scene. When combined with the localized feature detection capabilities of blob detection, the resulting hybrid feature set is both rich and nuanced, containing both high-level abstractions and specific, localized information.

DBNs act as the integrative core in this methodological fusion, adeptly combining the detailed textural and shape information gleaned from the blob detection with the abstracted representations learned by CNNs. DBNs, with their multiple layers of stochastic, latent variables, serve as a sophisticated architecture for feature combination and transformation. They provide a probabilistic framework that can discern and represent complex data distributions, making them particularly suited for integrating and refining the feature sets into a composite representation that enhances scene recognition performance.

By orchestrating the fusion of CNN and machine learning features with the robust structure of DBNs, the proposed approach capitalizes on the complementary strengths of each component. This innovative fusion not only intensifies the discriminatory power of the scene recognition system but also offers new pathways for research in machine learning and computer vision. The capacity to combine various types of features and to learn from these combinations in a deep architecture paves the way for developing more nuanced and powerful algorithms for scene understanding and image analysis in scientific and real-world applications.

4. Experiments and Results

This section describes the details of the experiments performed during this study and the results of these experiments in order to demonstrate the significance of the proposed model.

4.1. Datasets Description

The performance of our system is evaluated using the PASCAL VOC 12 and MSRC datasets. The details of these datasets are described as follows:

4.1.1.PASCAL VOC 2012 dataset

The PASCAL VOC dataset [20] has been developed for the recognition of a number of real scenes. The dataset consists of 11,530 images spanning 20 object classes with annotations for object detection and segmentation tasks. Images were randomly split into 70% for training, 15% for validation, and 15% for testing. Images including boat, airplane, bicycle, bottle, bus, bird, car, cat, cow, chair, dining table, horse, dog, motorbike, potted plant, person, sheep, sofa, train, and TV/monitor were resized to 320x320 pixels. More precisely, the data of PASCAL VOC belong to four categories namely, Animal, Person, Indoor, and Vehicle. Moreover, PASCAL VOC organizes samples into three categories that are comprised of regions of interest annotated objects, segmentation, and training/validation images. Some samples of PASCAL VOC 2012 are shown in Figure 6.

Figure 6. A few examples of the PASCAL VOC 2012

4.1.2. MSRC dataset

The MSRC dataset [21] comprised an object of 591 various categories with dynamic environments including street buildings, landscapes with hills, traffic signs, seaside, etc. The dataset is a collection of 15 different classes including bench, duck, boat, flower, cow, sky, cat, sign, water, tree, bird, dog, grass, car, and building. The resolution of the images in the dataset is 213x320 and every image has multiple objects in the scene image. Figure 7 illustrates the example images of the MSRC dataset.

Figure 7. Example images of the MSRC dataset

4.2. Performance Measurement and Result Analysis

This section presents the recognition accuracies of the proposed model using DBN over the PASCAL VOC 2012 and MSRC benchmarks. The proposed system is evaluated against the LOSO cross-validation

technique. Thus, the confusion matrix of the PASCAL VOC 2012 and MSRC datasets for scene recognition have been computed to present the results, as shown in Tab. 3 and Tab.4. Tab. 3 describes the confusion matrix of the obtained recognition accuracies that correspond to twenty classes over PASCAL VOC 2012 with an average accuracy of 93.30%. The accuracy of PASCAL VOC 2012 is slightly higher than MSRC due to the feature fusion technique. However, the higher recognition accuracies over other benchmark techniques using PASCAL VOC 2012 and MSRC where an extensive evaluation is performed, with the proposed system being compared against state-of-the-art methods. Tab. 6 provides the comparison of the proposed model's average recognition accuracy to that of previously existing approaches.

Experiments were conducted on a system equipped with an NVIDIA Tesla V100 GPU. Implemented using Python 3.7 and TensorFlow 2.3. All code was run under Linux Ubuntu 20.04.

The DBN consisted of three hidden layers with 500, 500, and 1000 neurons respectively. The activation function used was ReLU for hidden layers and softmax for the output layer. Learning Rate: 0.001, adjusted dynamically with a decay rate of 0.95 every 5 epochs. The batch size of 64 is used with 50 Epochs. Adam optimizer with $\beta1 = 0.9$ and $\beta2 = 0.999$ is used for optimization.

Initially, only CNN features are used to recognize the scene by applying DBN and achieved an accuracy of 85.75% and 86.11% over PASCAL VOC 2012 and MSRC datasets respectively. Next, an addition of blob features improved the recognition accuracy to 87.97% and 89.25% over PASCAL VOC 2012 and MSRC datasets respectively. Finally, the complete set of features (CNN+BF+GF) after fusion is provided to DBN for scene recognition. The results demonstrate that after the fusion of the features, our method has produced the highest accuracies of 93.30% and 92.53% over PASCAL VOC 2012 and MSRC datasets respectively during the experiments, as shown in Table 3 and Table 4. Subsection

Table 3. Recognition accuracies across the Pascal VOC 2012 dataset using DBN in terms of the confusion matrix

	HS	**BD**	**PR**	**CW**	**SH**	**AP**	**CT**	**DG**	**BT**	**BS**
HS	**0.94**	0	0	0.05	0	0	0	0.01	0	0
BD	0	**0.90**	0	0	0	0	0.06	0	0	0
PR	0	0.00	**0.96**	0	0	0	0	0	0	0
CW	0.04	0	0	**0.95**	0	0	0	0.01	0	0
SH	0.01	0	0	0	**0.94**	0	0	0.05	0	0
AP	0	0	0	0	0	**0.96**	0	0	0.01	0.03
CT	0	0	0	0	0	0	**0.95**	0.05	0	0
DG	0	0	0	0	0	0	0.04	**0.96**	0	0
BT	0	0	0	0	0	0.01	0	0	**0.92**	0.07
BS	0	0	0	0	0	0	0	0	0.04	**0.95**
TN	0	0	0	0	0	0	0	0	0.07	0.02
CR	0	0	0	0	0	0	0	0	0.03	0.02
MB	0	0	0	0	0	0	0	0	0	0

BC	0	0	0	0	0	0	0	0.04	0	0
BL	0	0	0	0	0	0	0	0	0	0
CH	0	0	0	0	0	0	0	0	0	0
DT	0	0	0	0	0	0	0	0	0	0
PP	0	0	0	0	0	0	0	0	0	0
SF	0	0	0	0	0	0	0	0	0	0
TV	0	0	0	0	0	0	0	0	0	0

TN	**CR**	**MB**	**BC**	**BL**	**CH**	**DT**	**PP**	**SF**	**TV**
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0.04	0	0
0	0	0	0	0	0	0	0.04	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0.01	0	0	0	0	0	0	0	0
0.91	0	0	0	0	0	0	0	0	0
0	**0.95**	0	0	0	0	0	0	0	0
0	0	**0.92**	0.08	0	0	0	0	0	0
0	0	0.03	**0.93**	0	0	0	0	0	0
0	0	0	0	**0.91**	0	0	0.09	0	0
0	0	0	0	0	**0.93**	0	0	0.07	0.00
0	0	0	0	0	0	**0.92**	0.04	0	0.04

0	0	0	0	0.05	0	0	**0.91**	0	0.04
0	0	0	0	0	0.01	0	0	**0.93**	0.06
0	0	0	0	0	0.05	0.03	0	0	**0.92**

Mean recognition accuracy = 93.30%

HS = horse; BD = bird; PR = person; CW = cow; SH = sheep; AP = aeroplane; CT = cat; DG = dog; BT = boat; BS = bus; TN = train; CR = car; MB = motorbike; BC = bicycle; BL = bottle; CH = chair; DT = dining table; PP = potted plant; SF = sofa; TV = television.

Table 4. Recognition accuracies across the MSRC dataset using DBN in terms of the confusion matrix

	CW	**DG**	**GR**	**BN**	**DK**	**FW**	**WT**	**SK**
CW	**0.95**	0.05	0	0	0	0	0	0
DG	0.04	**0.91**	0	0	0	0	0	0
GR	0.03	0	**0.89**	0	0	0.03	0	0
BN	0	0	0.01	**0.93**	0	0.01	0	0.01
DK	0	0.03	0.02	0	**0.91**	0	0	0.02
FW	0	0	0.02	0	0	**0.94**	0	0.02
WT	0	0	0	0	0	0	**0.96**	0.03
SK	0	0	0.03	0	0	0	0.03	**0.90**
CT	0.02	0.04	0	0	0	0	0	0
SN	0	0	0	0.06	0	0	0	0
TR	0	0	0.05	0	0	0.01	0	0.01
BT	0	0	0	0	0.03	0	0	0
CR	0	0	0.02	0	0	0	0	0
BD	0	0	0	0	0	0	0	0.08
BG	0	0	0.02	0.01	0	0	0	0.02

CT	**SN**	**TR**	**BT**	**CR**	**BD**	**BG**
0	0	0	0	0	0	0
0	0	0	0	0.05	0	0
0	0	0.03	0	0.01	0	0

0	0.02	0	0.02	0	0	0
0	0	0	0.02	0	0	0
0	0	0.02	0	0	0	0
0	0	0	0.01	0	0	0
0	0	0.02	0	0	0.02	0
0.93	0	0	0	0	0.01	0
0	**0.92**	0	0	0	0	0.02
0	0	**0.93**	0	0	0	0
0	0	0	**0.91**	0.06	0	0
0	0	0	0	**0.96**	0	0.02
0	0	0	0	0	**0.92**	0
0	0.02	0	0	0.01	0	**0.92**

Mean recognition accuracy = 92.53%

CW=cow; DG =dog; GR =grass; BN = bench; DK =duck; FW =flower; WT =water; SK =sky; CT =cat; SN =sign; TR =tree; BT =boat; CR =car; BD =bird; BG =building.

Table 5. Scene recognition accuracies of the proposed system by incorporating various feature combinations over benchmark datasets

Features	**Accuracy (%)**	
	MSRC	**PASCAL VOC 2012**
CNN	86.11	85.75
CNN + BF	89.25	87.97
CNN + BF + GF	92.53	93.3

Table 6. Comparison of recognition accuracies with other state-of-the-art methods over the PASCAL VOC 2012 dataset.

Author/Method	**Mean Recognition Accuracy %**
	PASCAL VOC 2012 Dataset
H. Zhao et al. [5]	85.40
S. Shetty et al. [22]	85.60

Y. Wei et al. [23]	81.80
T. Thanh-Dat et al. [28]	81.20
Z. Chang-Bin [29]	78.80
Proposed	**93.30**

Table 7. Comparison of recognition accuracies with other state-of-the-art methods over the MSRC dataset.

Author/Method	**Mean Recognition Accuracy %**
	MSRC Dataset
Z. Ye et al. [25]	77.00
C. Wu et al. [26]	90.30
D. Xie et al. [27]	92.50
Proposed	**92.53**

The proposed method significantly outperforms other methods, with a mean recognition accuracy of 93.30% over PASCAL VOC 2012 as shown in Table 6. However, the results are comparable over MSRC with equivalent recognition accuracy using the proposed technique as shown in Table 7. It reflects that the proposed approach, potentially utilizing a combination of DBNs for deep feature extraction and sophisticated semantic segmentation techniques, is highly effective for the PASCAL VOC 2012 dataset. The substantial lead over other methods indicates possibly superior handling of the dataset's complexity, possibly through more effective feature representation and generalization capabilities.

5. Discussion

The study presents an approach that combines semantic segmentation and feature fusion to improve scene recognition accuracy. The results obtained by the study on both the MSRC and PASCAL VOC 2012 datasets show that the proposed approach outperforms existing state-of-the-art methods. On the MSRC dataset, the proposed approach achieves an accuracy of 92.53%, which is significantly higher than the accuracy obtained by the benchmark methods. On the PASCAL VOC 2012 dataset, the proposed approach achieves an accuracy of 93.30%, which is also significantly higher than the state-of-the-art method.

While the overall performance is commendable, certain classes within the datasets presented challenges. For instance, in the PASCAL VOC 2012 dataset, some misclassifications were noted between similar object classes, such as between 'bus' and 'train', likely due to their structural similarities. In the MSRC dataset, differentiation between natural elements like 'grass' and 'tree' proved difficult, potentially due to overlapping features in the greenery.

The study demonstrates the effectiveness of combining semantic segmentation and feature fusion for scene recognition tasks. The proposed approach is able to extract and fuse multi-level features from both the original images and their corresponding semantic segmentation maps, which leads to more discriminative features and improved recognition accuracy. The study also shows that using a DBN for feature learning is a powerful tool for scene recognition, as it is able to automatically learn hierarchical representations of the input data.

The evaluation of our system across the PASCAL VOC 2012 and MSRC datasets has demonstrated the efficacy of the proposed model in recognizing a diverse array of scenes. With an achieved mean

accuracy of 93.30% on PASCAL VOC 2012 and 92.53% on MSRC, our system not only shows high performance but also indicates good generalizability across different types of datasets.

The PASCAL VOC 2012 dataset, characterized by its 20 distinct classes categorized into animals, person, indoor, and vehicle, poses a complex challenge due to the variability of object scale, pose, and occlusion. Despite these challenges, our model's fusion of CNN, blob, and geometrical features (CNN+BF+GF) yielded a high recognition accuracy, suggesting robust feature extraction and an effective strategy for handling diverse scene complexities.

On the other hand, the MSRC dataset includes dynamic environments like street buildings and landscapes with hills, amongst its 15 classes. The slightly lower accuracy on this dataset could be attributed to the smaller number of classes and possibly more challenging scene compositions, such as outdoor and landscape scenes that may present less-defined object boundaries and more varied textures.

By combining both deep learning features (from CNNs) and traditional machine learning features, the model achieves a more robust and discriminative feature set. Employing DBN allows for the optimization of feature integration, leveraging the model's ability to learn complex hierarchical representations, which significantly boosts the accuracy and robustness of scene classification.

Individually, each of these techniques contributes to the enhanced capability of the system to recognize and interpret complex scenes more accurately than existing methods. Collectively, they integrate to form a system that is not only capable of high accuracy in variable conditions but also robust against the common challenges in scene recognition such as lighting variations, occlusions, and diverse object presentations.

Overall, the results obtained by the study suggest that the proposed approach is a promising method for scene recognition tasks and has the potential to be applied to other computer vision tasks as well. The study provides insights into the importance of semantic segmentation and feature fusion for scene recognition and highlights the effectiveness of DBNs for feature learning in such tasks.

Our findings suggest that while the proposed method is generally robust and adaptable, performance can vary depending on the scene complexity and the distinctiveness of object features within the datasets. The fusion of CNN with blob and geometrical features has proven beneficial in capturing a comprehensive feature set that aids in the generalizability of the model. These contributions collectively push the boundaries of what's possible with scene recognition systems, providing advancements that could be beneficial for applications in autonomous driving, surveillance, and augmented reality, where understanding the complex dynamics of real-world scenes is paramount.

6. Conclusion and Future Work

This paper presents an approach for scene recognition over various complex images. The proposed system performs segmentation and extracts various features including deep and machine learning features. After feature fusion, scene recognition is performed by applying the DBN. The features fusion process is the key to improving the recognition rate of scenes over benchmark datasets. The proposed recognition system has a number of real-time applications like sports activity recognition, autonomous driving, robotics, and surveillance systems. When compared to other recognition systems, the methodology of our proposed system outperformed the other systems in terms of recognition accuracy. The feature fusion process is integral to this success, enabling the system to capture a rich set of discriminative features that enhance recognition accuracy.

We are committed to extending our work with different CNN-based semantic segmentation techniques along with multiple feature extraction and fusion for scene recognition in general as well as aerial images. Moreover, future work will explore further refinement of CNN-based semantic segmentation techniques and the integration of additional features to improve recognition in even more generalized scenes, including aerial imagery. Our commitment to advancing the field will focus on enhancing the adaptability and robustness of scene recognition systems to a wider range of applications, such as autonomous driving and robotics.

This analysis reflects on the provided data, emphasizing the high generalizability of the proposed model while also noting areas where performance varies, possibly due to the inherent challenges of certain scene types or dataset specifics. Future work is outlined with a focus on continuing to improve the system's generalizability.

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Y. Hua, J.Guo and H. Zhao (2015). Deep belief networks and deep learning, In Proceedings of 2015 International Conference on Intelligent Computing and Internet of Things (pp. 1-4). IEEE.
2. L. Zhang, X .Zhen and L. Shao (2014). Learning object-to-class kernels for scene classification, IEEE Transactions on image processing, vol. 23, no.8, pp. 3241-3253.
3. X. Song, S. Jiang and L. Herranz, "Joint multi-feature spatial context for scene recognition on the semantic manifold," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Bostom, USA, pp. 1312-1320, 2015.
4. R. Kachouri, M. Soua, and M. Akil (2016). Unsupervised image segmentation based on local pixel clustering and Low-Level region merging, In 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Monastir, Tunisia, pp. 177-182, IEEE.
5. H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia (2017). Pyramid scene parsing network, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, pp. 2881-2890.
6. N. Hussain, M. A. Khan, M. Sharif, S. A. Khan and A. A. Albesher et al (2020). A deep neural network and classical features based scheme for objects recognition: an application for machine inspection, Multimedia Tools and Applications, pp. 1-23.
7. S. Xia, J. Zeng, L. Leng and X. F (2019). WS-AM: weakly supervised attention map for scene recognition, Electronics, vol. 8, no.10, pp. 1072.
8. Chu, J., Guo, Z., and Leng, L. (2018). Object detection based on multi-layer convolution feature fusion and online hard example mining. IEEE access, 6, 19959-19967.
9. Y.Zhang, J.Chu, L.Leng and J.Miao, (2020). Mask-refined R-CNN: A network for refining object details in instance segmentation, Sensors, 20(4), 1010.
10. L. Leng, M. Li, C. Kim, X.Bee, " Dual-source discrimination power analysis for multi-instance contactless palmprint recognition", Multimedia Tools Application, vol. 76, pp.333–354, 2017.
11. L. Leng and J. Zhang (2013). PalmHash Code vs. PalmPhasor Code, Neurocomputing, vol.108, pp.1-12.
12. J. Miao, X. Zhou and T. Z. Huang (2020). Local segmentation of images using an improved fuzzy C-means clustering algorithm based on self-adaptive dictionary learning, Applied Soft Computing, vol. 91, pp. 106200.
13. M. Koskela and J. Laaksonen (2014). Convolutional network features for scene recognition, In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, USA, pp. 1169-1172.
14. S. Tammina (2019). Transfer learning using vgg-16 with deep convolutional neural network for classifying images, International Journal of Scientific and Research Publications (IJSRP), vol. 9, no.10, pp. 143-150.
15. P. Szuster (2019). Blob extraction algorithm in the detection of convective cells for data fusion, Journal of Telecommunications and Information Technology, vol. 4, pp. 65-73.
16. Y. Xu. T. Wu, F. Gao, J.R. Charlton, and K. M. Bennett(2020). Improved small blob detection in 3D images using jointly constrained deep learning and Hessian analysis, Scientific reports, vol. 10, no.326, pp. 1-12.
17. G. E. Hinton, S. Osindero and Y. W. The (2006). A fast learning algorithm for deep belief nets, Neural Computation, vol. 18, pp. 1527-1554.
18. Y. Liu, S. Zhou and Q. Chen (2011). Discriminative deep belief networks for visual data classification, Pattern Recognition, vol. 44, no.10, pp. 2287-2296.
19. S. Kamada and T. Ichimura (2019). "An object detection by using adaptive structural learning of deep belief network," In 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, pp. 1-8.
20. M. Everingham, S. M. Eslami, L. Van Gool, C. K. Williams, J. Winn et al (2015). The pascal visual object classes challenge: A retrospective, International journal of computer vision, vol. 111, pp. 98-136.
21. J. Shotton, J. Winn, C. Rother and A. Criminisi (2006). "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," In European Conference on Computer Vision, pp. 1-15, Springer, Berlin, Heidelberg.
22. S. Shetty (2016). Application of convolutional neural network for image classification on Pascal VOC challenge 2012 dataset, arXiv preprint arXiv:1607.03785.
23. Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni et al (2015). HCP: A flexible CNN framework for multi-label image classification, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no.9, pp. 1901-1907.
24. P. Tang, X. Wang, B. Shi, X. Bai, W. Liu et al (2018). Deep fishernet for image classification, IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no.7, pp. 2244-2250.

25. Z. Ye, P. Liu, W. Zhao and X. Tang (2015). Hierarchical abstract semantic model for image classification, Journal of Electronic Imaging, vol. 24, no.5, pp. 053022.
26. C. Wu, Y. Li, Z. Zhao and B. Liu (2019). Image classification method rationally utilizing spatial information of the image, Multimedia Tools and Applications, vol. 78, pp. 19181-19199.
27. D. Xie, Q. Li, W. Xia, S. Pang, H. He and Q. Gao (2019). "Multi-view classification via adaptive discriminant analysis," IEEE Access, vol. 7, pp. 36702-36709.
28. Truong, T. D., Prabhu, U., Raj, B., Cothren, J., & Luu, K. (2023). FALCON: Fairness Learning via Contrastive Attention Approach to Continual Semantic Scene Understanding in Open World. arXiv preprint arXiv:2311.15965.
29. Zhang, C. B., Xiao, J. W., Liu, X., Chen, Y. C., & Cheng, M. M. (2022). Representation compensation networks for continual semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7053-7064).