Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.049333

ARTICLE

UNet Based on Multi-Object Segmentation and Convolution Neural Network for Object Recognition

Nouf Abdullah Almujally[1], **Bisma Riaz Chughtai**[2], **Naif Al Mudawi**[3], **Abdulwahab Alazeb**[3], **Asaad Algarni**[4], **Hamdan A. Alzahrani**[5] and **Jeongmin Park**[6,*]

[1]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia

[2]Department of Computer Science, Air University, Islamabad, 44000, Pakistan

[3]Department of Computer Science, College of Computer Science and Information System, Najran University, Najran, 55461, Saudi Arabia

[4]Department of Computer Sciences, Faculty of Computing and Information Technology, Northern Border University, Rafha, 91911, Saudi Arabia

[5]Information Technology Department, College of Computing and Informatics, Saudi Electronic University, Riyadh, 13316, Saudi Arabia

[6]Department of Computer Engineering, Korea Polytechnic University, Siheung-si, Gyeonggi-do, 237, Republic of Korea

*Corresponding Author: Jeongmin Park. Email: jmpark@tukorea.ac.kr

Received: 03 January 2024 Accepted: 25 April 2024 Published: 18 July 2024

ABSTRACT

The recent advancements in vision technology have had a significant impact on our ability to identify multiple objects and understand complex scenes. Various technologies, such as augmented reality-driven scene integration, robotic navigation, autonomous driving, and guided tour systems, heavily rely on this type of scene comprehension. This paper presents a novel segmentation approach based on the UNet network model, aimed at recognizing multiple objects within an image. The methodology begins with the acquisition and preprocessing of the image, followed by segmentation using the fine-tuned UNet architecture. Afterward, we use an annotation tool to accurately label the segmented regions. Upon labeling, significant features are extracted from these segmented objects, encompassing KAZE (Accelerated Segmentation and Extraction) features, energy-based edge detection, frequency-based, and blob characteristics. For the classification stage, a convolution neural network (CNN) is employed. This comprehensive methodology demonstrates a robust framework for achieving accurate and efficient recognition of multiple objects in images. The experimental results, which include complex object datasets like MSRC-v2 and PASCAL-VOC12, have been documented. After analyzing the experimental results, it was found that the PASCAL-VOC12 dataset achieved an accuracy rate of 95%, while the MSRC-v2 dataset achieved an accuracy of 89%. The evaluation performed on these diverse datasets highlights a notably impressive level of performance.

KEYWORDS

UNet segmentation; blob; fourier transform; convolution neural network

1 Introduction

Recent progress in vision technologies, especially in detecting multiple objects and understanding scenes, and feature extraction, relies heavily on how well we can break down images and pick out important features. Semantic segmentation has become a crucial tool in modern computer vision. By continually refining and enhancing different techniques, researchers aim to unlock new possibilities for visual perception and expand the capabilities of computer vision systems to address real-world challenges more effectively.

Over the past two decades, researchers have dedicated significant attention to advancing various aspects of computer vision, with a particular focus on semantic segmentation, multi-object identification, and scene recognition. These endeavors stem from the challenges encountered in object recognition tasks, which include background elimination, feature implementation and enhancement, processing efficiency, and accurate scene categorization. Semantic segmentation stands as a fundamental task in computer vision, aiming to partition an image into semantically meaningful regions. Researchers have explored various approaches to address this task, ranging from traditional methods based on pixel-wise classification [1] to more recent deep learning-based techniques [2]. These efforts seek to improve the accuracy and efficiency of segmentation algorithms, enabling applications such as autonomous driving, medical image analysis, and image-based navigation systems. In parallel, the identification of multiple objects within an image has been the subject of intense investigation. This task involves detecting and classifying multiple objects present in a scene, often requiring the ability to handle diverse object shapes, sizes, and orientations. Researchers have proposed numerous algorithms and frameworks for multi-object identification, leveraging techniques such as object detection [3], instance segmentation [4], and object recognition [5]. These advancements contribute to applications in surveillance, robotics, and augmented reality, among others. Furthermore, achieving accurate object categorization involves capturing high-level contextual information and discriminative features from visual data, often through the integration of deep learning architectures [6] and large-scale datasets [7].

To demonstrate the value of our research, we utilized two datasets: PASCALVOC-12 and MSRC-v2. These extensive databases contain thousands of photos across a wide range of classes, with PASCALVOC-12 comprising 20 classes and MSRC-v2 containing 16 classes. Each dataset presents unique challenges, including variations in illumination, color, occlusion, and object class correspondences. In response, we propose a novel approach to multi-object segmentation and classification. Initially, we employ a bilateral filter for image smoothing and subsequently utilize the UNet model for segmentation. Following segmentation, we annotate the segmented regions using a dedicated tool. After semantic segmentation, we extract distinct features from the identified objects, including Edge features, Blob features, Fast Fourier transform, and KAZE (Accelerated Segmentation and Extraction) features. Edge features provide insight into boundaries and transitions between regions, while the Fast Fourier transform highlights frequency-based information. Blob features identify specific intensity patterns or textures within segmented objects, and KAZE features extract scale-invariant descriptors. These techniques enhance our understanding of segmented objects and facilitate improved analysis, categorization, and decision-making. Finally, we employ convolutional neural networks for object recognition.

The following are the research article’s main contributions:

- We have successfully conducted semantic segmentation employing the UNet deep learning architecture. This approach has proven effective in accurately representing and classifying distinct regions within the input data, leading to meaningful and relevant segmentation outcomes.

- To achieve optimal results by utilizing a smaller feature set, an optimization technique based on frequency-based scale invariant-based features has been implemented.
- Utilizing convolutional neural network (CNN) techniques for object recognition has demonstrated remarkable effectiveness in generating positive and accurate results.
- These advanced techniques significantly enhance our understanding of segmented objects, facilitating feature analysis, improved recognition, and informed decision-making processes.

The following sections represent the remaining part of the paper: Section 2 contains a thorough summary of the appropriate research study. The methodological approach is explained in Section 3, which also includes the presentation of the proposed framework created for object recognition. In Section 4, the empirical evaluation thoroughly examines various datasets and summarizes their findings. In conclusion, Section 5 outlines the key findings and offers recommendations for future research endeavors.

2 Related Work

The area of object identification has been the subject of numerous scholarly investigations, including a range of methodologies involving various methods for semantic segmentation and object detection. Our evaluation of the relevant research includes in-depth solutions to topics including segmentation, item identification, labeling, object categorization, and recognition in dynamic scenarios. Semantic segmentation is thoroughly investigated in the following discussion, with an emphasis on how it can be accomplished through the use of traditional machine learning, current deep learning techniques, and classification techniques as will be detailed in the ensuing subsections.

2.1 Multi-Objects Segmentation

The research introduces an adaptive K-means picture segmentation method that uses a straightforward procedure to get accurate results without the requirement for bilateral K value input [8]. To serve as a technical guide for coaches and athletes in the creation of VR (Virtual Reality) sports simulation systems, the study suggests an enhanced Ada Boost classifier for sports scene detection in films. To improve training efficacy and prevent injuries, the approach concentrates on faithfully imitating a variety of competitive sports. The Ada Boost model used in this paper provides more efficient analysis than conventional scene recognition techniques that use low- or high-level features [9]. The method for precise object segmentation described in this research makes use of active shape and appearance models that change over time in response to traditional appearance clues at form boundaries and the output of a support vector machine (SVM). The process involves segmenting test images by evolving the projected object contour by the probability determined by the SVM classifier, training an SVM classifier using training photos, and creating form and appearance models of the objects [10]. The study presents a unique method for scene comprehension that combines geometric characteristics, Histogram of Oriented Gradient (HOG), and Scale Invariant Feature Transform (SIFT) descriptors to incorporate multiple item detection segmentation and scene labeling [11]. Their work offers a UNet++ approach with deep supervision trained on the dataset for weed detection in sugar beet fields. On the dataset segmentation challenge, they compare UNet, UNet++, and UNet++ with deep supervision architectures [12]. They examine three models for semantic image segmentation: UNet, VGG16_FCN, and ResNet50_FCN [13]. AP-UNet, a novel semantic segmentation model, is suggested. To improve the accuracy of deep and shallow information fusion, a channel attention mechanism based on the UNet network structure is added to the encoder portion. This technique highlights the desired attributes while reducing interference from surrounding noise. A pyramid pooling module is added to the bottom layer of the encoder to keep global context data [14]. The research suggests a contextual hierarchical model (CHM) for semantic segmentation that learns contextual information in a hierarchical framework [15]. This work offers a semantic segmentation-based system for identifying strawberries and guava. Using human-annotated photos, the updated UNet model was trained to correctly separate ripened strawberry and guava fruit [16].

2.2 Multi-Objects Classification

To classify objects, a Convolutional Neural Network (CNN) model is used with exemplary effectiveness. To excel at processing visual data, such as photographs, and correctly categorizing objects within them, this advanced neural network architecture has been purposefully created. The network skillfully captures the rich local patterns and spatial correlations that are inherent to images by utilizing convolutional layers. Affirming its crucial role in applications such as autonomous vehicles, medical diagnostics, and facial recognition, among others, its proficiency has resulted in significant innovations in a variety of fields dependent on precise image-based categorization. In a variety of tasks, including image classification (identifying objects in images), object detection (finding multiple items in an image), image segmentation (segmenting images into separate regions), and others, CNNs have proven to perform exceptionally well. The fundamental objective of this research is to provide a comprehensive overview, considering the current period of rapid progress, of the region-based convolutional neural networks (R-CNN) family, which comprises the R-CNN, Fast R-CNN, and Faster R-CNN. Comparing the faster R-CNN to the R-CNN and Fast R-CNN, simulation results show that the faster R-CNN improves detection speed and accuracy. It has been found that faster R-CNN is very suitable for reliable object detection [17]. The categorizing and recognizing of airplane images in the CIFAR-10 dataset is the specific subject of the paper, which focuses on object recognition in images using a convolutional neural network (CNN). With 25 epochs and 60,000 images, the authors trained the CNN model. On a TensorFlow CPU system, an epoch takes between 722 to 760 s. The model's training accuracy reached 96% after 25 epochs [18]. The paper applies deep learning to multi-class question acknowledgment employing a CNN made with normalized standard initialization and prepared with a dataset of test images from 9 distinctive object categories. The results are compared with vectors extricated from variation approaches of BOW based on a direct L2-SVM classifier. The experiments confirm the adequacy and strength of the CNN demonstrate, accomplishing a precision rate of 90.12% [19]. They used several methods for discovery and classification, such as Convolutional Neural Networks (CNN), Regions with CNN (R-CNN), Fast R-CNN, and Faster R-CNN [20]. A convolutional neural network with two channels (DC-CNN) model is suggested by the researchers to boost the effectiveness of automatic labeling. The model is made up of two separate CNN channels, the first of which was trained on low-frequency samples and the second of which was trained on all training sets. To choose a label, the findings from the two channels are combined. The Pascal VOC 2012 dataset is used to validate the suggested technique, and it achieves good accuracy [21].

3 Material and Methods

This section presents a novel multi-object detection and identification model that can identify and classify several objects in RGB (Red, Green and Blue) photographs. An image is loaded into the model's workflow after which it goes through an analysis and segmentation phase. To efficiently segment the many objects in the image, a UNet architecture is employed in this process. The segmented objects are then evaluated using a variety of feature extraction techniques, including Blob and KAZE features, Energy-based edge detection, and Fast Fourier transform. Fig. 1 provides a visual representation of our suggested model's whole flow.

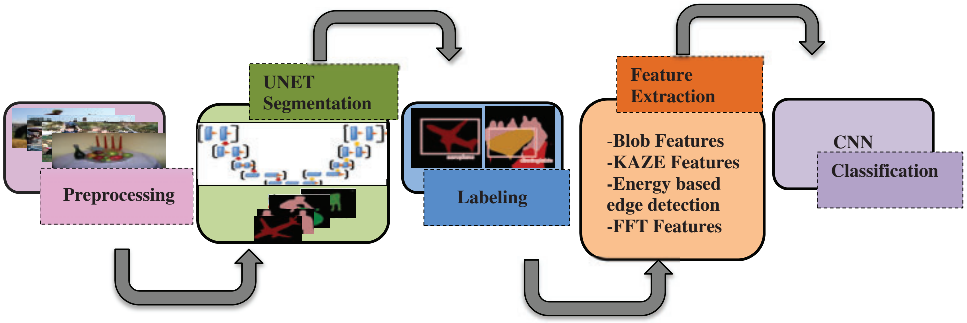

Figure 1: Proposed multi-objects recognition model

3.1 Image Acquisition and Preprocessing

A mutual method of smoothing and reducing noise in photos while keeping crucial edges and details is called bilateral filtering. Traditional linear filters, such as Gaussian filters, apply a weighted average of pixel values within a local neighborhood. While these filters can help reduce noise, they tend to blur edges and details, which might not be desirable in certain applications like image enhancement or computer vision tasks. Contrarily, bilateral filtering considers both the spatial separation between pixels and the range difference (intensity or color difference) between them. The goal is to calculate the weighted average of pixels that fall within a given spatial range, although similarity in intensity or color values also affects the weights. In other words, pixels that are similar to the central pixel in terms of intensity or color will be given a higher weight, whereas pixels that are distinct in terms of those attributes will receive a lesser weight. Spatial domain sigma (s) and intensity range sigma (r) are the two factors that regulate the spatial distance and intensity difference.

During pre-processing, a bilateral filter was applied to the image using a spatial diameter (d) of 15 pixels, and both the color standard deviation (σ_{color}) and spatial standard deviation (σ_{space}) were set to 75. This configuration was chosen to effectively reduce noise while preserving fine structures and edges in the image. The bilateral filter's parameters were carefully selected to strike a balance between noise suppression and the retention of important image features, ensuring high-quality denoising results.

$$BF\,(a,b) = \frac{1}{W_p}\sum\nolimits_{(m,n)\in N} I\,(a+m,b+n)\,.\,w_p(m,n) \tag{1}$$

where $BF\,(a,b)$ is the filtered value at pixel (a,b). $I\,(a+m,b+n)$ is the intensity/color value at pixel (x + i, y + j). $w_p(m,n)$ is the weight associated with the pixel at relative coordinates. N is the neighborhood around the pixel (a, b). W_p is the normalization factor.

The weight w_p (a, b) is calculated using the Gaussian function of both spatial and intensity differences as shown in Fig. 2.

$$w_p\,(a,b) = exp = \left(-\frac{a^2+b^2}{2\sigma_s^2} - \frac{||I\,(a+m,b+n) - I\,(a,b)||}{2\sigma_r^2}\right) \tag{2}$$

3.2 Objects Segmentation

An in-depth discussion of multi-object segmentation is presented in this section. Segmentation plays an important role in image processing as it helps to divide an image into distinct and meaningful regions or objects. These researchers used different UNet models [22,23].

Figure 2: A few examples of PASCAL VOC 2012 dataset (a) original images (b) pre-processed images

We employed a UNet-based architecture for semantic segmentation tasks on the PASCALVOC-12 dataset. The architecture is comprised of a symmetric expanding path that facilitates exact localization and a contracting path that records context. To be more precise, up-sampling and convolutions are used in the expanding path, and a sequence of convolutions and max-pooling layers in the contracting path. We employed Skip connections between the contracting and expanding paths to recover spatial details lost during down-sampling. Next, we used the Adam optimizer to optimize the categorical cross-entropy loss function that we used to train the model. To adapt the model to our dataset, we first convert the XML annotations into binary masks representing the object of interest. These binary masks are used as ground truth labels for training the UNet model. The images and their corresponding masks are resized to 128 × 128 pixels and normalized before being fed into the network. A different collection of images is used to validate the trained model, and the findings indicate that it can effectively differentiate objects from the background. This model shows 72% segmentation accuracy. The IoU (Intersection over Union) score is 78%. Post-processing includes thresholding the predicted probabilities to generate binary masks, which can optionally be color-mapped for visualization. The results are shown in Fig. 3.

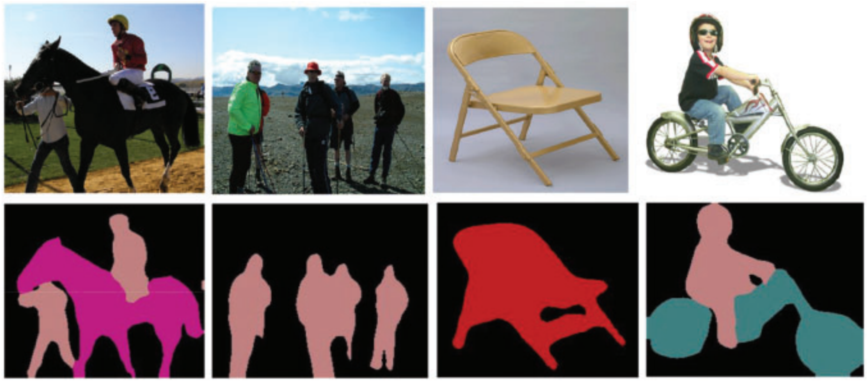

Figure 3: UNet-based segmentation results over some segmented images (row1) and original images (row2) segmented images

The convolution operation is defined as follows for the given input matrix X and filter W:

$$Y\left(i,j\right)=\sum\nolimits_{n}X\left(i+m,j+n\right),W\left(m,n\right) \tag{3}$$

In the final layer, a SoftMax function is applied to each pixel to classify it into C classes:

$$P(c/\boldsymbol{x})=\frac{\exp(\boldsymbol{w}_c.\boldsymbol{x}+b_c)}{\sum_{i=1}^{C}\exp(w_i.\boldsymbol{x}+b_i)} \tag{4}$$

Here $P(c/\boldsymbol{x})$ is the probability of class c given the input $\boldsymbol{x}$ and w_c and b_c are the weights and bias for the class c, respectively. Algorithm 1 shows the working of the segmentation model.

Algorithm 1: UNet segmentation

```
Input: Input image $I$
Output: Segmented image $S$
1. Define the UNet architecture with an encoder-decoder structure.
Encoder: $fe = ge(I)$, where $ge$ denotes the encoder network.
Decoder: $fd = gd(fe)$, where $gd$ denotes the decoder network.
2. Initialize the encoder network to extract features from the input image:
        Encoder network architecture: $ge(I) = fe$
3. Initialize the decoder network to upsample and generate the segmentation mask:
        Decoder network architecture: $gd(fe) = fd$
4. Split the input image into patches to fit the network input size.
5. For each patch:
    5.1. Forward pass the patch through the encoder network to extract features:
        $fe = ge(patch)$, where $patch$ represents the current patch.
    5.2. Forward pass the features through the decoder network to generate the segmentation mask:
        $semask = gd(fe)$, where $semask$ represents the segmentation mask.
    5.3. Combine the segmentation masks from all patches to reconstruct the final segmented image
6. Perform post-processing step thresholding on the segmented image $S$.
7. Output the segmented image $S$.
```

3.3 Labeling

In our efforts to label data, we utilized different annotation tools [24] a precious resource. We specifically employed the Annotely tool for labeling the data. Our labeling operations have been substantially accelerated and improved by its user-friendly design, powerful features, and versatility. Fig. 4 depicts results.

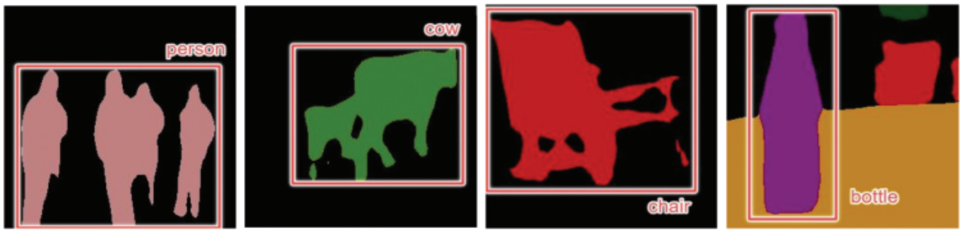

Figure 4: Labeling using the Annotely tool

3.4 Feature Extraction Computation over Segmented Objects

Various features are computed as shown in Algorithm 2 to identify the objects. A basic and crucial stage in pattern identification, computer vision, and image processing is feature extraction. Feature extraction is performed on segmented objects using three distinct algorithms: KAZE feature, energy-based edge detection feature, blob extraction, and Fast Fourier transform. This method carefully extracted important features from specific areas of interest that were segmented. By combining these algorithms, we created a comprehensive set of features, which helped us better analyze, understand, and recognize objects.

3.4.1 KAZE over Segmented Objects

A technique used in computer vision and image processing for finding and describing local features in images is called KAZE. The feature points are designed to be extracted and described using a novel feature descriptor called KAZE-HOG [25]. Researchers have used a variety of algorithms to demonstrate a diversity of feature extraction methodologies [26–29]. When extracting these features, we employed heterogeneous methodology for KAZE feature extraction initially we began with the construction of non-linear scale-spaces and further refined for precise localization. Descriptors are then computed for the key points based on the gradients or intensities of surrounding pixels, encapsulating this information in a form that is both rotation and scale and variant.

The scale space for L (x, y, t) for an image I (x, y) using non-linear diffusion filtering is computed as:

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{t}} = div(c(x, y, y)\nabla L \tag{5}$$

where div is the divergence operator, ∇ is the gradient, and $c(x, y, t)$ is the diffusion coefficient. Detection of key point location the determinant of the Hessian Matrix can be computed as:

$$\text{Det(H)} = \begin{vmatrix} L_{xx} & L_{xy} \\ L_{xy} & L_{yy} \end{vmatrix} \tag{6}$$

Here, L_{xx}, L_{xy}, L_{yy} are the second partial derivatives. It is important to note that for each detected key point, KAZE computes a feature descriptor by capturing the weighted orientation histograms in the local neighborhood of the key point. Fig. 5 represents the extracted keypoints.

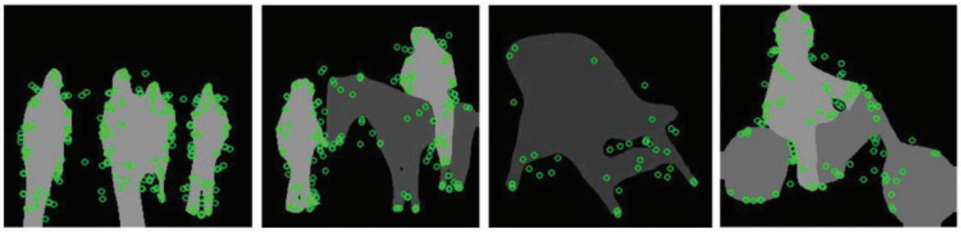

Figure 5: Feature extraction using KAZE features

3.4.2 Energy Function-Based Edge Detection

The motivation behind energy function-based edge detection is to formulate the problem as an optimization work to identify edges in an image. By minimizing or maximizing an energy function also known as a cost function this technique aims to identify the image's edges. The energy function represents the edges in the image such as intensity gradients, color differences, or texture variations to detect edges. Optimization is used to minimize or maximize the energy function. Once the optimization is performed the outcome is a set of points or regions in the image that represent the identified edges, and these points suggest where the edges are most likely to exist. A thresholding step is frequently used to differentiate between edge and non-edge pixels to create a binary edge map. The next step is to visualize the edges that have been found by either presenting them individually or incorporating them into the original image. The grayscale image defined as I (x, y) are the pixel coordinates. The energy function for E (x, y) for the edge detection based on the gradient magnitude can be defined as: Fig. 6 displays the resultant images of energy detection.

$$E\left(x,y\right)=\sqrt{\left[G_x\left(x,y\right)+G_y\left(x,y\right)\right]^2} \tag{7}$$

where G_x (x, y) is the gradient of the image in the x-direction at pixel (x, y). G_y (x, y) is the gradient of the image in the y-direction at pixel (x, y). The gradient of the x and y directions is calculated using derivatives or convolution operations, using the Sobel operator.

$$G_x(x,y)=I\left(x+1,y\right)-I(x-1,y) \tag{8}$$

$$G_y(x,y)=I\left(x,y+1\right)-I(x,y-1) \tag{9}$$

The energy function E (x, y) quantifies the magnitude of the gradient at each pixel in the image. High values of E (x, y) indicate rapid changes in pixel intensity and are indicative of edges.

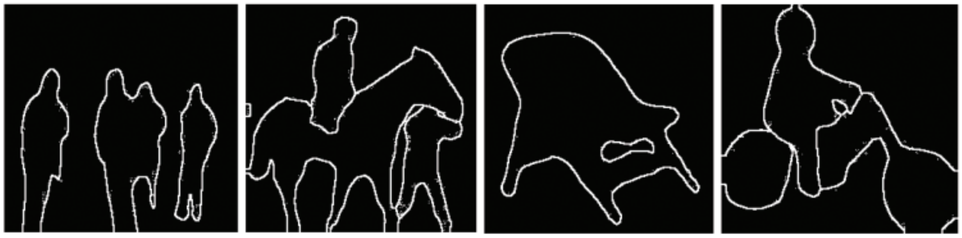

Figure 6: Feature extraction using energy function-based edge detection

3.4.3 Blob Extraction over Segmented Objects

A computer vision approach called blob extraction, commonly referred to as blob detection is used to locate and separate regions of interest in an image. We first begin with converting the image to grayscale, then we segment the image using techniques such as thresholding to separate the object from the background. Furthermore, we employed connected component algorithms like Two-Pass or Union-Fined are applied to tag individual blobs. This labeling is based on pixel neighborhood definition which can be either 4-connectivity or 8-connectivity. Once the blobs are labeled various features like area, centroid, and bounding box are extracted using region properties.

$$I_{gray}=\left(x,y\right)=Grayscale(I_{RGB}\left(x,y\right)) \tag{10}$$

Mathematically, component labeling is described as:

$$L\left(x,y\right)=\begin{cases}n, & if\ I_{gray(x,y)\ is\ connected\ and\ part\ of\ blob\ n}\\0, & otherwise\end{cases} \tag{11}$$

Here “*n*” is the unique integer label for each blob. Therefore, we can represent the feature vector by formulating. In Fig. 7, the results of blob extraction are depicted.

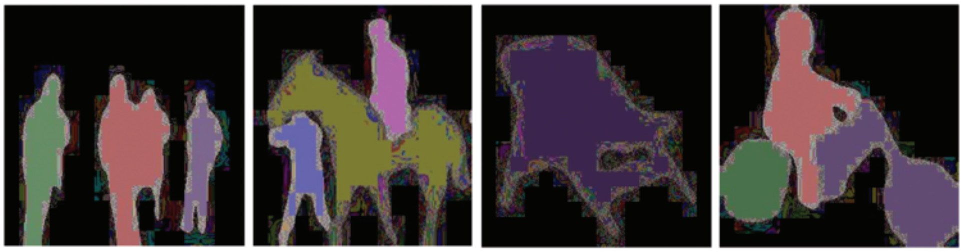

Figure 7: Features extraction using blob technique

3.4.4 Fast Fourier Transform

Fast Fourier transform reveals information about frequencies and patterns in the image by converting spatial data to the frequency domain. Initially image was converted to grayscale to minimize computational complexity and to concentrate on the intensity variation across the image. After that, a 2D Fourier transform was applied to the grayscale image translating it from its original spatial domain to frequency domain. The resulting Fourier transform output is complex and contains both magnitude and phase information to facilitate more visualization and analysis the zero-frequency component was shifted to the center of the spectrum. This repositioning is important for examining the low-frequency component of the image which are often is more informative. We then computed the magnitude spectrum which provides valuable information on the image frequency. Fig. 8 illustrates the extracted results.

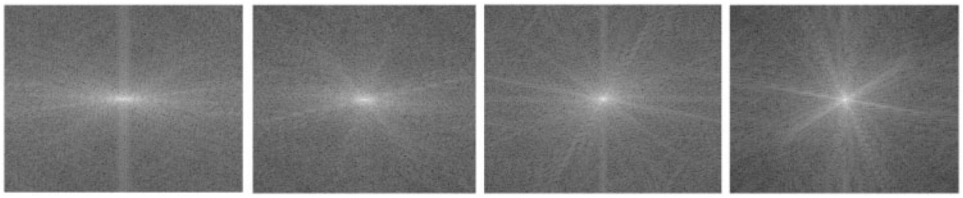

Figure 8: Features extraction using Fast Fourier transform technique

Mathematically, the 2D Fourier transform of an image can be computed as:

$$F(u,v) = F(u,v) = \iint f(x,y)\,.e^{-j(ux,vy)}dxdy \tag{12}$$

where $F = (u,v)$ is the Fourier transform and $f(x,y)$ is the original image.

The magnitude spectrum is calculated using $|F(u,v)| = \sqrt{Re(F(u,v))^2 + Im(F(u,v))^2}$, which involves both the real and imaginary components of

$$F(u,v) \tag{13}$$

The phase spectrum is given as:

$$F(u,v) = arctan\left(\frac{Im(F(u,v))}{Re(F(u,v))}\right) \tag{14}$$

where $Re = (F(u,v))$: The real part of the Fourier transform, which represents cosine wave coefficients. $Im = (F(u,v))$: The imaginary part of the Fourier transform which represents sine wave coefficients.

```
Algorithm 2: Feature extraction
Input: Segmented images
Output: Feature Vector of extracted features
1. F1 ← ExtractKAZEDescriptors(SegmentedImages)
2. F2← ExtractBlobDescriptors(SegmentedImages)
3. F3 ← CalculateFFTFrequencies(SegmentedImages)
4. F4 ← CalculateEnergyBasedFeatures(SegmentedImages)
5. Concatenated_Features ← ConcatenateFeatures(KAZE_Descriptors, Blob_Descriptors, FFT_
   Features, Energy_Features)
6. Feature_Vector ← CreateFeatureVector(Concatenated_Features)
7. Return Feature_Vector
```

3.5 Multi-Object Recognition via Convolution Neural Network (CNN)

Several layers make up this model, including convolution layers, pooling layers, output layers, fully connected layers, dropout layers, flatten layers, and activation functions. This model uses the Keras library with the sequential Application Programming Interface (API). This model takes input images of size (22, 224, 3), padding the same, with three convolution layers, MaxPooling layers, dropout layers, and twso fully connected layers for classification. The first and second convolution layers apply 32 filters to the input images with a 3 × 3 kernel, using the ReLU activation function with a resolution of 224 × 224 pixels. The second convolution layer with the 64 filters and 3 × 3 Kernel size. The feature map's spatial dimensions have been reduced by the maxPooling layer. The output from the previous layer is flattened into a one-dimensional vector using flattened layers. The convolution layers must be connected to the fully connected layers. By randomly changing a portion of the input units to zero during training, the dropout layer helps to prevent overfitting. Following that, two dense, totally linked layers were employed. With 513 units, the first dense layer employs the ReLU activation function. The SoftMax activation function, which is commonly used for multi-class classification tasks, is employed in the second dense layer, which has 20 units. It produces probabilities for every one of the twenty classes. The architecture of the convolution neural network is shown in Fig. 9.

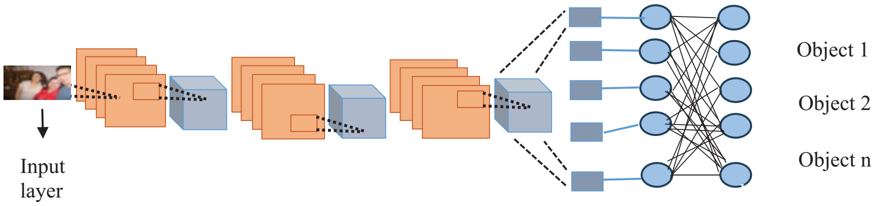

Figure 9: Flow of convolution neural network for proposed object recognition

4 Performance Evaluation

Employing the PASCALVOC-12 and MSRC-v2 benchmark datasets, the suggested model is tested and validated.

4.1 Datasets Description

The PASCALVOC-2012 with 21 classes and no pre-segmented items, [30] focuses on urban street scenes for semantic scene interpretation. The dataset is comprised of 17,000 images of complex scenarios having different indoor and outdoor images, e.g., person, car, potted plant, motorbike, bicycle, bird, airplane, boat, bottle, bus cat, dog, chair, cow dining table, horse, sheep, sofa, TV/monitor.

The MSRC-v2 dataset [31] is comprised of object classes, 591 photos, and precise pixel-by-pixel annotated images. The MSRC-v2 is the extended version of the MSRC dataset. Despite having 23 object classes, only 16 of them are frequently utilized. Given that the current annotation includes both individual object instances and pure class annotation alongside full scene segmentation, the dataset may also be used for object instance segmentation. The object classes are named grass, chair, flower, bird, road, tree, building, horse, signboard, book, aero plane, cycle, sheep, person, cow, and car.

4.2 Experimental Settings and Results

Experiments were conducted using a hardware system equipped with an Intel Core i7 processor operating at 2.11 GHz, along with 16 GB of RAM, with the GPU Intel(R) UHD Graphics 620, running a 64-bit Windows 11 operating system. All experiments were performed utilizing various image processing methods and libraries available in Python. To comprehensively evaluate the proposed framework, multiple sets of experiments were conducted, including assessments of classification accuracy, precision, recall, and F1-score. The experimental results indicate that our proposed segmentation and feature extraction methods effectively differentiate between various object classes within the PASCALV0C-12 and MSRC-v2 datasets. The results of the segmentation accuracy over PASCALVOC-2012 show 74% with the IoU being 78% while the MSRC-v2 shows a segmentation accuracy of 79% with the IoU being 82%. Employing the proposed approach across 16 different objects yielded a notable classification accuracy of 89% on the MSRC-v2 and 95% over the PASCALVOC-12, as demonstrated in Tables 1 and 2. While the computational accuracy examined for the PASCALVOC-12 is 5070s and for MSRC-v2 1250s. Notably, certain object classes, such as bird, person, and airplane showed particularly high accuracy, underscoring their robust recognition performance.

Table 1: Recognition accuracy over the PASCAL VOC-12 dataset

Obj	ae	bd	be	ce	bs	Bt	ch	cr	Ct	cw	Dg	dt	he	Mb	pn	pp	sa	sp	tn	tv
ae	**0.93**	0	0	0	0.07	0	0	0	0	0	0	0	0	0	0.04	0.02	0	0	0	0
bd	0	**0.95**	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03	0	0.02	0	0
be	0	0.05	**0.93**	0	0.05	0	0	0	0.02	0	0	0	0	0	0	0	0	0	0	0
ce	0.01	0	0	**0.96**	0	0	0	0	0	0	0	0	0	0.03	0	0	0	0	0	0
bs	0	0.01	0	0	**0.98**	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0
bt	0	0.09	0	0.08	0	**0.52**	0	0.06	0	0.02	0.03	0	0	0.05	0	0.07	0	0	0.08	0
ch	0	0.05	0.03	0	0	0	**0.96**	0	0.02	0	0	0	0	0	0.02	0	0	0	0	0
cr	0	0.01	0	0	0	0	0	**0.98**	0	0	0	0.01	0	0	0	0	0	0	0	0
ct	0	0.01	0	0	0	0	0	0	**0.99**	0	0	0	0	0	0	0	0	0	0	0
cw	0	0.01	0	0	0.01	0	0	0.02	0	**0.96**	0	0	0.01	0.02	0.01	0	0	0	0	0
dg	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0
dt	0	0	0	0	0	0.01	0	0.01	0.01	0	0	**0.97**	0	0	0	0	0	0	0	0
he	0	0.02	0	0	0	0	0	0	0.01	0	0	0	**0.99**	0	0	0	0	0	0	0

(Continued)

Table 1 **(continued)**

Obj	ae	bd	be	ce	bs	Bt	ch	cr	Ct	cw	Dg	dt	he	Mb	pn	pp	sa	sp	tn	tv
mb	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0
pn	0	0	0	0	0	0	0	0	0	0	0.03	0	0	0	**0.97**	0	0	0	0	0
pp	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0	0	**0.99**	0	0	0	0
sa	0	0.02	0	0	0	0	0.02	0	0	0	0	0	0	0	0	0	**0.96**	0	0	0
sp	0	0	0	0	0	0	0	0	0	0	0.01	0	0	0.01	0	0	0	**0.98**	0	0
tn	0	0	0	0	0.02	0	0	0	0	0	0	0	0	0	0	0	0.04	0.09	**0.84**	0.03
tv	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	**0.99**
Mean accuracy = 95%																				

Note: ae = aeroplane; bd = bird; be = bottle; ce = cycle; bs = bus; bt = boat; ch = chair; cr = car; ct = cat; cw = cow; dg = dog; dt = dinningtable; he = horse; mb = motorbike; pn = person; pp = pottedplant; sa = sofa; sp = sheep; tn = train; tv = television/monitor.

Table 2: Recognition accuracy over the MSRC-v2 dataset

Obj	aer	blg	bok	brd	cye	chr	Car	cow	Flr	hre	psn	Rod	sbd	shp	Tre	grs
aer	**1**	0	0	0	0	0	0	0.	0	0	0	0	0	0	0	0
blg	0	**0.84**	0.06	0.03	0	0	0	0	0.05	0	0	0	0	0	0.02	0.
bok	0	0.02	**0.98**	0.01	0.01	0.	0	0	0	0	0	0	0	0	0	0
brd	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0
cye	0	0.03	0	0	**0.75**	0	0	0.02	0.05	0.01	0	0.04	0	0.05	0	0.06
chr	0	0	0	0	0	**0.89**	0	0.04	0.	0.01	0	0.04	0	0	0.02	0
car	0.07	0	0.05	0	0.05	0	**0.83**	0	0	0	0	0	0	0	0	0
cow	0.01	0	0	0.01	0	0.08	0	**0.70**	0.09	0	0	0	0.05	0.06	0	0
flr	0	0	0	0	0	0	0	0	**0.97**	0	0	0	0.03	0	0	0
hre	0.01	0	0	0.05	0	0.01	0	0	0	**0.93**	0	0	0	0	0	0
psn	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0
rod	0	0	0.04	0	0	0	0	0	0	0	0	**0.86**	0	0	0	0
sbd	0.01	0	0	0.08	0	0.06	0	0.02	0	0.09	0	0	**0.74**	0	0	0.
shp	0	0	0	0	0	0	0	0.04	0	0	0	0	0	**0.96**	0	0
tre	0	0	0	0.01	0.02	0	0	0.04	0.07	0.	0	0	0	0	**0.86**	0
grs	0	0.13	0	0	0	0	0	0	0	0	0	0	0	0	0	**0.87**
Mean accuracy = 89 %																

Note: aer = aeroplane; blg = building; bok = book; brd = bird; cye = cycle; chr = chair; car = car; cow = cow; flr = flower; hre = horse; psn = person; rod = road; sbd = signboard; shp = sheep; tre = tree; grs = grass.

4.2.1 Recognition Accuracy

Using the PASCALVOC-2012 and MSRC-v2 datasets, the four distinct features Blob, Fast Fourier transform, and KAZE were retrieved for this experiment. A confusion matrix for object recognition on the PASCALVOC-2012 dataset is shown in Table 1. Using the PASCALVOC-2012 dataset, an average accuracy of 95% is achieved after 25 repetitions of the experiment. Table 2 shows that 89%

accuracy was obtained for the MSRC-v2 dataset. Table 3 represents the F1, recall accuracies for the PASCALVOC-12 and Table 4 shows for the MSRC-v2.

Table 3: Measurements of Accuracy, Recall, and F1-score over PASCALVOC-2012

Object	Precision	Recall	F1-score	Object	Precision	Recall	F1-score
ae	0.97	0.93	0.95	Cr	1.00	0.98	0.99
bd	1.00	0.95	0.98	Ct	0.99	0.99	0.99
be	0.98	0.93	0.95	Cw	0.99	0.96	0.97
ce	0.94	0.48	0.63	Dg	1.00	1.00	1.00
bs	1.00	0.96	0.98	Dt	0.97	0.97	0.97
bt	0.99	0.98	0.99	He	1.00	0.99	1.00
ch	0.63	0.96	0.76	Mb	1.00	1.00	1.00
pn	0.95	0.97	0.96	Pn	0.95	0.97	0.96
pp	0.80	0.99	0.88	tn	0.97	0.86	0.90
sa	0.73	0.96	0.83	Tv	0.98	0.99	0.98
sp	0.97	0.98	0.98				
Weighted accuracy = 0.95%							

Table 4: Measurements of Accuracy, Recall, and F1-score over MSRC-v2

Object	Precision	Recall	F1-score	Object	Precision	Recall	F1-score
aer	0.92	1.00	0.96	flr	0.75	0.97	0.85
bld	0.90	0.84	0.87	hre	0.92	0.93	0.93
bok	0.72	0.98	0.83	psn	1.00	1.00	1.00
brd	1.00	1.00	1.00	rod	0.83	0.86	0.85
cye	0.75	0.80	0.79	sbd	1.00	0.74	0.85
chr	0.99	0.89	0.92	Shp	0.89	0.96	0.92
car	0.99	0.83	0.90	Grs	0.87	0.90	0.85
cow	0.67	0.70	0.69	Tre	1.00	0.86	0.92
Weighted accuracy = 89%							

4.2.2 Comparison with State-of-the-Art (SOTA) Methods

The differences and improvements between our method and other methods explored in the paper primarily revolve around the effectiveness and efficiency of our segmentation feature extraction and recognition technique. The papers reviewed present innovative approaches in computer vision and image processing, each with its unique contributions and limitations. While they propose novel methods for tasks such as boundary detection, semantic segmentation, feature extraction, and image retrieval, common limitations persist across the board. However, there is still a problem in capturing all the relevant visual information present in the images. These include challenges related to algorithm accuracy, computational efficiency, and dataset bias. This could impact the accuracy and effectiveness of image retrieval and classification tasks, particularly for complex or detailed image content. We found that the accuracy increased by at least 7% when compared to the PASCALVOC-2012 dataset, while a minimum of 4% increase in the accuracy of the MSRC-v2 dataset respectively is achieved during our experiments. Table 5 uses benchmark datasets to compare recognition accuracy results with those from other SOTA methods.

Table 5: Accuracy of recognition (%) in comparison to benchmark datasets

Methods	PASCALVOC-2012	MSRC-v2
Xu et al. [32]	67.9	–
Oquab et al. [33]	70.2	–
Gidaris et al. [34]	73.9	–
Ravi et al. [35]	–	74.0
Ahmed et al. [36]	–	88.75
Proposed	**95**	**89**

5 Conclusion

This research introduces a novel object detection model designed to tackle challenges in indoor and outdoor environments. The approach integrates advanced deep learning architectures, feature extraction methods, and annotation tools to create an efficient pipeline for image analysis and object recognition. The study leverages the PASCALVOC-12 and MSRC-v2 datasets for evaluation and validation. These datasets consist of thousands of photos spanning various classes, with PASCALVOC-12 featuring 20 classes and MSRC-v2 containing 16 classes. These datasets pose challenges in this research images include how bright they are, their colors, things blocking the view, and how different things are shown, including problems with the data, like when there is too much or too little of certain things, which can make the results seem wrong. Another challenge is making the method work when overlapping objects or occlusion. To tackle these challenges, we propose a novel approach to multi-object segmentation and classification. Initially, we employed a bilateral filter to effectively reduce noise and smooth the images. Then, utilize the UNet model for segmentation, followed by annotation using a dedicated tool to label segmented regions. Finally, CNN is utilized for the recognition of each object. These techniques contribute to a comprehensive understanding of segmented objects, enhancing analysis, categorization, and decision-making processes. With recognition accuracies of 95%, and 89% over the PASCALVOC-2012 and MSRC datasets, respectively. In the future, we want to explore new features like geometric-based features, depth, and features of multiple objects to make region extraction and object recognition more accurate. We will use deep learning methods to enhance the effectiveness of our semantic segmentation for scene understanding. This will help us understand how objects relate to each other and the overall scene. We also plan to explore different methods for labeling using deep learning techniques.

Acknowledgement: The authors are thankful to Princess Nourah bint Abdulrahman University Researchers Supporting Project Number (PNURSP2024R410), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Funding Statement: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) Program (IITP-2024-RS-2022-00156326) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation). The authors are thankful to the Deanship of Scientific Research at Najran University for funding this work under the Research Group Funding Program Grant Code (NU/GP/SERC/13/30). The funding for this work was provided by Princess Nourah bint Abdulrahman University Researchers Supporting Project Number (PNURSP2024R410), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the Project Number "NBU-FFR-2024-231-06".

Author Contributions: Study conception and design: Bisma Riaz Chughtai, and Nouf Abdullah Almujally; data collection: Naif Al Mudawi, and Abdulwahab Alazeb; analysis and interpretation of results: Bisma Riaz Chughtai, and Asaad Algarni; draft manuscript preparation: Bisma Riaz Chughtai, Hamdan A. Alzahrani and Jeongmin Park. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All publicly available datasets are used in the study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017. doi: 10.1109/TPAMI.2016.2572683.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Med. Image Comput. Computer-Assisted Interven.*, vol. 9351, pp. 234–241, Nov. 2015. doi: 10.1007/978-3-319-24574-4_28.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks,"*Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 1736–1744, Dec. 2015. doi: 10.48550/arXiv.1506.01497.

[4] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, vol. 1, pp. 2961–2969. doi: 10.1109/ICCV.2017.322.

[5] A. Jalal, A. Ahmed, A. A. Rafique, and K. Kim, "Scene semantic recognition based on modified fuzzy c-mean and maximum entropy using object-to-object relations," *IEEE Access*, vol. 9, pp. 27758–27772, Feb. 2021. doi: 10.1109/ACCESS.2021.3058986.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, Sep. 2014. doi: 10.48550/arXiv.1409.1556.

[7] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, pp. 1452–1464, Jun. 2018. doi: 10.1109/TPAMI.2017.2723009.

[8] K. Venkatachalam, V. P. Reddy, M. Amudhan, A. Raguraman, and E. Mohan, "An implementation of k-means clustering for efficient image segmentation," in *2021 10th IEEE Int. Conf. Commun. Syst. Netw. Technol. (CSNT)*, Bhopal, India, 2021, pp. 224–229.

[9] Y. Yue and Y. Yang, "Improved Ada Boost classifier for sports scene detection in videos: From data extraction to image understanding," in *2020 Int. Conf. Inventive Comput. Technol., (ICICT)*, Coimbatore, India, 2020, pp. 1–4.

[10] S. Luo and J. Li, "Accurate object segmentation using novel active shape and appearance models based on support vector machine learning," in *2014 Int. Conf. Audio, Lang. Image Process., (ICALIP)*, Shanghai, China, 2014, pp. 347–351.

[11] A. Ahmed, A. Jalal, and K. Kim, "Multi-objects detection and segmentation for scene understanding based on texton forest and kernel sliding perceptron," *J. Electr. Eng. Technol.*, vol. 16, no. 2, pp. 1143–1150, 2021. doi: 10.1007/s42835-020-00650-z.

[12] X. Z. Hu, W. S. Jeon, and S. Y. Rhee, "Sugar beets and weed detection using semantic segmentation," in *2022 Int. Conf. Fuzzy Theory and Its Appl., (iFUZZY)*, Kaohsiung, Taiwan, 2022, pp. 1–4.

[13] N. S. Jonnala, N. Gupta, C. P. Vasantrao, and A. K. Mishra, "BCD-Unet: A novel water areas segmentation structure for remote sensing image," in *7th Int. Conf. Intell. Comput. Control Syst., (ICICCS)*, Madurai, India, 2023, pp. 1320–1325.

[14] S. Wang, Y. Hu, and L. Zhang, "Analysis and research of segmentation feature data of FEMORAL CT image based on improved Unet," in *2023 IEEE 3rd Int. Conf. Elect. Commun., Int. Things and Big Data, (ICEIB)*, Taichung, Taiwan, 2023, pp. 114–118.

[15] M. Seyedhosseini and T. Tasdizen, "Semantic image segmentation with contextual hierarchical models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 951–964, May 2016. doi: 10.1109/TPAMI.2015.2473846.

[16] N. Y. Venkatesh and V. Kr, "Deep learning based semantic segmentation to detect ripened strawberry guava fruits," in *2022 IEEE Int. Conf. Elect., Comput. Commun. Technol., (CONECCT)*, Bangalore, India, 2022, pp. 1–6.

[17] O. Hmidani and E. M. Ismaili Alaoui, "A comprehensive survey of the R-CNN family for object detection," in *2022 5th Int. Conf. Adv. Commun. Technol. Netw., (CommNet)*, Marrakech, Morocco, 2022, pp. 1–6.

[18] D. P. Sudharshan and S. Raj, "Object recognition in images using convolutional neural network," in *2018 2nd Int. Conf. Inventive Syst. Control, (ICISC)*, Coimbatore, India, 2018, pp. 718–722.

[19] S. Hayat, S. Kun, Z. Tengtao, Y. Yu, T. Tu and Y. Du, "A deep learning framework using convolutional neural network for multi-class object recognition," in *2018 IEEE 3rd Int. Conf. Image, Vis. Comput., (ICISC)*, Chongqing, China, 2018, pp. 194–198.

[20] G. Priyadharshini and D. R. Judie Dolly, "Comparative investigations on tomato leaf disease detection and classification using CNN, R-CNN, Fast R-CNN and Faster R-CNN," in *9th Int. Conf. Adv. Comput. Commun. Syst., (ICACCS)*, Coimbatore, India, 2023, pp. 1540–1545.

[21] P. Juyal and A. Kundaliya, "Multilabel image classification using the CNN and DC-CNN model on pascal VOC, 2012 dataset," in *Int. Conf. Sust. Comput. Smart Syst., (ICSCSS)*, Coimbatore, India, 2023, pp. 452–459.

[22] J. Karimov *et al.*, "Comparison of UNet, ENet, and BoxENet for segmentation of mast cells in scans of histological slices," in *Proc. 2019 Int. Multi-Conf. Eng., Comput. Inf. Sci*, Novosibirsk, Russia, Oct. 2019, pp. 544–547. doi: 10.1109/SIBIRCON48586.2019.8958121.

[23] F. Yifei, "Image semantic segmentation using deep learning technique," in *Proc. 3rd Int. Conf. Signal Process. Mach. Learn.*, May 2023, vol. 4, pp. 810–817. doi: 10.54254/2755-2721/4/2023439.

[24] J. Aljabri, M. AlAmir, M. AlGhamdi, M. Abdel-Mottaleb, and F. Collado-Mesa, "Towards a better understanding of annotation tools for medical imaging: A survey," *Multimed. Tools Appl.*, vol. 81, no. 18, pp. 25877–25911, Jan. 2022. doi: 10.1007/s11042-022-12100-1.

[25] X. Wang, L. Duan, Y. Fan, and C. Ning, "A multi-sensor image matching method based on KAZE-HOG features," in *2019 IEEE 4th Int. Conf. Image, Vis. Comput.*, Xiamen, China, 2019, pp. 514–517.

[26] Y. Liu, C. Lan, C. Li, F. Mo and H. Wang, "S-AKAZE: An effective point-based method for image matching," *Optik*, vol. 127, no. 14, pp. 5670–5681, Jul. 2016. doi: 10.1016/j.ijleo.2016.03.072.

[27] N. A. A. Khalid, M. I. Ahmad, T. S. Chow, T. H. Mandeel, I. M. Mohammed and M. A. K. Alsaeedi, "Palmprint recognition system using VR-LBP and KAZE features for better recognition accuracy," *Bull. Electr. Eng. Inform.*, vol. 13, no. 2, pp. 1060–1068, Apr. 2024. doi: 10.11591/eei.v13i2.4739.

[28] U. Muhammad, W. Wang, A. Wang, and S. Pervez, "Bag of words KAZE (BoWK) with two-step classification for high-resolution remote sensing images," *IET Comput. Vis.*, vol. 13, no. 4, pp. 395–403, Jun. 2019. doi: 10.1049/iet-cvi.2018.5069.

[29] A. A. Rafique, M. Gochoo, A. Jalal, and K. Kim, "Maximum entropy scaled super pixels segmentation for multi-object detection and scene recognition via deep belief network," *Multimed. Tools Appl.*, vol. 82, no. 9, pp. 13401–13430, Apr. 2023. doi: 10.1007/s11042-022-13717-y.

[30] P. Akiva and K. Dana, "Single stage weakly supervised semantic segmentation of complex scenes," in *2023 IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, 2023, pp. 5943–5954.

[31] Z. Li, Y. Jiang, J. Yue, J. Fang, Z. Fu and D. Li, "A new hybrid PCNN for multi-objects image segmentation," in *Proc. 2012 Asia Pacific Signal and Inform. Process. Assoc. Annual Summit and Conf.*, Hollywood, CA, USA, 2012, pp. 1–6.

[32] X. Xu, F. Meng, H. Li, Q. Wu, Y. Yang and S. Chen, "Bounding box based annotation generation for semantic segmentation by boundary detection," in *Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Taipei, Taiwan, 2019, pp. 1–2.

[33] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *2014 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, 2014, pp. 1717–1724.

[34] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware CNN model," in *IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 1134–1142.

[35] D. Ravì, M. Bober, G. M. Farinella, M. Guarnera, and S. Battiato, "segmentation of images exploiting DCT based features and random forest," *Pattern Recognit.*, vol. 52, no. 4, pp. 260–273, 2016. doi: 10.1016/j.patcog.2015.10.021.

[36] A. Ahmed, A. Jalal, and K. Kim, "A novel statistical method for scene classification based on multi-object categorization and logistic regression," *Sensors*, vol. 20, no. 14, pp. 3871, Jul. 2020. doi: 10.3390/s20143871.