Classification of Road Scenes Based on Heterogeneous Features and Machine Learning

Original Scientific Paper

Sanjay P. Pande
Yeshwantrao Chavan College of Engineering,
Department of Computer Technology,
Hingna, Nagpur, Maharashtra, India
sanjaypande2001@gmail.com

Poonam T. Agarkar*
Ramdeobaba University,
School of Computer Science and Engineering
Katol Raod, Nagpur, Maharashtra, India
agarkarp@rknec.edu

Sarika Khandelwal
G H Raisoni College of Engineering,
Department of Computer Science and Engineering
Digdoh Hills, Nagpur, Maharashtra, India
sarikakhandelwal@gmail.com

Rajani D. Singh
Ballarpur Institute of Technology,
Department of Master of Computer Application
Ballarpur, Chandrapur, Maharashtra, India
rajanidsingh@gmail.com

Pratik R. Hajare
Mansarovar Global University,
Department of Electrical and Electronics Engineering
Raison Road, Bhopal, Madhya Pradesh, India
pratikhajare8@gmail.com

Prashant R. Patil
Smt. Radhikatai Pandav College of Engineering,
Department of Management Studies
Umrer Road, Nagpur, Maharashtra, India
patilnagpur@gmail.com

*Corresponding author

Abstract *– There is a rapid advancement in Artificial intelligence (AI) and Machine Learning (ML) that has extensively improved the object detection capabilities of smart vehicles today. Convolutional Neural Networks (CNNs) based on small, medium, and large networks have made significant contributions to in-vehicle navigation. Simultaneously, achieving higher level accuracies and faster response in autonomous vehicles is still a challenge and needs special care and attention and must be addressed for human safety. Hence, this article proposes a heterogeneous features-based machine learning framework to distinguish road scenes. The model incorporates object-based, image-based, and diverse conventional features from the road scene images generated from four distinct datasets. Object-based features are acquired using the YOLOv5m model and modified VGG19 networks, whereas image-based features are extracted using the modified VGG19 network. Conventional features are added to the object-based and blind features by applying a variety of descriptors that include Matched filters, Wavelets, Gray Level Occurrence Matrix (GLCM), Linear Binary Pattern (LBP), and Histogram of Gaussian (HOG). The descriptors are used to extract fine and course features to enhance the capabilities of the classifier. Experiments show that the proposed road scene classification framework performed better in classifying two scene categories, including crosswalks, parking, roads under bridges/tunnels, and highways achieving an average classification accuracy of 97.62% and the highest of 99.85% between crosswalks and Parking. A marginal improvement of approximately 1% is seen when all four categories were considered for evaluation using a multiclass SVM compared to other competing models.*

Keywords: *Artificial intelligence, Machine Learning, smart vehicles, CNN, object-based, image-based, diverse conventional features, YOLOv5m, and VGG19.*

Received: July 27, 2024; Received in revised form: December 16, 2024; Accepted: December 23, 2024

1. INTRODUCTION

Safer autonomous vehicles work on algorithms based on computer vision that can distinguish certain scenarios and accurately predict labels. Related scenes consist of several details and are infinite. Varying image classification achievements are significant and include a wide range of image classes [1, 2]. Remarkable results have been obtained on the ImageNet dataset using convolutional neural networks and frequent improvements are suggested by many researchers [3]. However, further initiatives are needed in scene categorization to improve visual perception in autonomous driving. Work introduced in [4] considered 2.5 million images for training with 205 categories of worldwide places that included outdoor scenes. This was based

on scene-centric Places and used CNN for recognition/classification. The work was extended in [5] for 365 categories from 2.1 million images. Object-centric feature-based training is easier than scene-level tasks due to their varied scene diversities and possible scene combinations.

Several techniques infer scene categories based on object detection related to certain scenes. Semantic information was also used to guide a mobile robot [6, 7] in the indoor environment for high-level navigation. Thus, semantic mapping within a scene image is widely used for navigating vehicles with better accuracy. It uses semantic information that includes parking lots, objects beside roads, buildings, towers, sidewalks, and constructions to represent rural or urban road scenarios. However, rural conditions are different and are concerned with the abundance of plants, trees, and several vegetation. Open broad roads and heavy vehicles are dominant over highways which are different from the crowded streets in rural and urban areas. Scene categories are prominently defined based on weather and light conditions [8], stationary objects in the scene [9], special traffic scene categories (An example to quote - Square with Street lights), and intersections. Related datasets are not provided with labels, however, they include an overall description and attributes for objects in the scene [10]. Scene videos are typically summarized on a clip basis, since they are not annotated frame by frame, making the labeling process more time-consuming and costly.

In this article, we have collected road scene images from different source datasets and segregated them using a machine-learning tool. The work comprises heterogeneous features that are extracted from the scene images which assist the classifier in distinguishing two classes with better accuracy. The quality features used as input to the support vector machine are based on semantic information about the scene, objects in the scene, and conventional attributes. Pre-trained networks are used to extract the object-based and image-based features whereas diverse descriptors are used to lift different details of the scene images. The scene images are manually selected and labeled [11]. We believe that the proposed scene classification network can be used for offline and real-time applications including driver assistance and mapping semantically autonomous datasets.

Despite several deep-learning and machine-learning models, AI is still not capable of annotating and distinguishing the RS environment autonomously, without human intervention. Numerous experiments were conducted considering the good road conditions and weather, but more recent experiments include real road and weather conditions. Adverse weather conditions such as rain, fog, thick pollution, and snow are still to be evaluated properly for self-driven cars. The resolution of the LiDAR cameras has been enhanced to a great extent and is not the image quality that matters, relating to object recognition and classification. Many findings infer that autonomous robots are no longer a question, but when and how they would be launched in human society. The only question is how safely they will drive on the real roads, irrespective of the geographical structures and conditions.

This emphasizes a critical need for reliable detection of the scene objects using efficient techniques, mathematical modeling, and simulations that can exactly represent reality and converge at the best performance parameters and architectures to adapt to variations in the surroundings. Various contextual factors are required to be considered to improve the generalization capability and confident predictions for the road scene classification, from where the images were acquired. Vision-based perceptual systems are greatly influenced by contextual factors such as geographical locations, weather conditions, and illuminations, geographical or artificial processes.

Findings reported that recent state-of-the-art techniques incorporated deep learning for object detection and scene understanding in the scene images, and there is a broader scope for additional improvements. Still, the performance of CNNs is yet to be investigated under realistic conditions, that when and under what fatal conditions it will cease to operate and can pose a great threat to precious human life in self-driven circumstances.

The weaknesses and deficiencies found after surveying most of the recent and persistent studies are:

1. The inability to detect and classify large objects in the scenes.
2. False detections for small objects.
3. Lack of generalization ability due to changes in weather conditions.
4. Lack of scene content representation, and
5. Complexity of time and computation.

Therefore, there is always room for improvement in distinguishing scenes based on their contents to assist Automated Vehicles. The present research aims to design an intelligent scene classification framework with higher accuracy and lower complexity irrespective of the object dimensions, weather conditions, and uneven illuminations.

The paper claims the following contributions:

1. There is an extraction of object-based features using the VGG19 pre-trained network after detecting the objects with the YOLOV5 network and resizing them to a predetermined dimension.

2. Handcrafted low- and high-level features on gray-scale images are also extracted to improve the disparities among the classes and improve classification. The features include wavelet-based features, local binary pattern-based features, gray-level co-occurrence

matrix features, histograms of Gaussian features, and matched filter coefficients.

3. Blind features (Image-Based features) using VGG19 from the color scene images are extracted from the last fully connected layer of the VGG19 network to obtain the depth level information of the images.

4. The analysis based on the experiments displayed that the proposed Machine learning framework employing a diverse set of features can classify road scenes with higher accuracy.

The remaining paper is framed as follows: Work carried out by different researchers is summarized in the forthcoming section and our proposed scene classification framework is elaborated in the preceding section after the literature review. The last section concludes by discussing the experimental results and avenues of future studies after analyzing the results obtained through our proposed framework.

2. RELATED WORK

The objects associated with the road scene images need accurate detection and classification for precise decisions to assist the driver in taking different actions along the road. Nowadays, for a better 3D perspective, object identification has taken its place as a subdomain in computer vision tasks [12]. The objective is to provide safety, save lives, minimize accidents, and make transportation reliable, and efficient [13, 14]. A variety of techniques are found in the literature for detecting objects in images relative to several applications. Specific objects for specific applications are now a sub-problem of the generalized recognition task. It includes attribute and name assignments for specific objects [15]. The most crucial and challenging part is dealing with 3D objects for autonomous vehicle driving using an optical navigation system. Several sensors are mounted to provide road scene details to the navigation module. In the end, a classifier system is used to collect information and guide the vehicle along a derivable region [16]. Udacity recognized multiple transportation means in the scene by employing HOG features and classified them using various classifier networks. They primarily used the GTI (Grupo de Tratamiento de Imágenes, Madrid, Spain) and the KITTI (Karlsruhe Institute of Technology, Karlsruhe, Germany and Toyota Technological Institute, Nagoya, Japan) benchmark datasets. They obtained superior results using the logistic regression module [17] as compared to SVM and decision trees.

The BDD100K dataset was constructed using several images making it large and comprehensive including a variety of objects acquired in diverse weather situations, places, and times, with occlusions and a wide range of intensity conditions. The YOLO model constructed using the Deep CNN is based on learned features extensively used to detect objects in the real-time environment in videos and images. The work proposed in [12] used YOLOv3 and YOLOv4 models on the BDD100K dataset and obtained significant results improving the detection rate. The authors replaced Leaky RLU with advanced activation functions (MISH and SWISH) and further improved the detection accuracy over the Leaky RLU [12].

Objects of different dimensions (small, medium, and large) were detected in [14] using a single–shot multibox detector (SSD), faster region-based CNN (RCNN), and algorithms present in PyTorch. Experiments were carried out on the BDD100K dataset images. Further research included the KITTI dataset where the performance was measured using average precision for detecting 3D objects in the scene images. The outputs were significantly enhanced by dividing the 3D objects into easy, moderate, and difficult levels. The last level included classifying objects in foggy environments for autonomous vehicles [18].

Object detection in the dark (night) was better in [19] using YOLOv3, Aggregate view object detection, and PointPillars. The techniques resulted in better average precision over the KITTI dataset than others. PointPillars performed the best over objects at night, however, it failed to detect objects in rainy conditions [19]. Sparse LiDAR Stereo Fusion Networks were incorporated in [18] to improve object detection in foggy weather (Multi-fog environment – KITTI) [18]. A combination of YOLOv3 and Darknet-53 was used for detecting and classifying various objects [20]. The work suggested in [21-22] used CNN to convert semantic details from sensory data in the images on the road to recognize cycle riders, pedestrians, vehicles, etc. A novel approach was proposed to process ambulance sounds from a long distance to determine the direction of the emergency vehicle [23].

3D object detection and classification on real and synthetic samples was implemented in [24]. Most studies are compared using the average precision as the evaluation measure. The weaknesses and deficiencies found after surveying most of the recent and persistent studies [25-33] are time complexity, inability to detect and classify large objects in the scenes, lack of generalization ability due to changes in weather conditions, and false detection for small objects. Therefore there is always room for improvement in distinguishing scenes based on their contents to assist autonomous vehicles. The present research aims to design an intelligent scene classification framework with higher accuracy and lower complexity irrespective of the object dimensions, weather conditions, and uneven illuminations.

3. MATERIALS AND METHOD

The authors of this work collected the road scene images from four different sources. The objective was to consider the worst possible scenario for scene classification to assist automated vehicles. The manually separated scene images possess complexity related to multiclass, poor illumination on account of different weather conditions and dimensions. The familiar data-

sets that were used to generate a custom dataset for this work include the LabelMe [34], KITTI [35], BDD100K [36], and Places365 [37] datasets. The crucial parameter to list out the complexity of scene images is their poor imperceptibility. The significant class was still undistinguished even with visual perceptivity. The challenge was to detect the relevant objects in the scene that were not easily detectable. The authors customized the dataset that contained a sum of 2725 scene images from the four benchmark datasets and included road scene images with highway roads (HR), vehicle parking lots (VPL), crosswalks (CRW), and roads under bridges/tunnels (RB/T). Fig. 1 shows road scene images from all four classes.

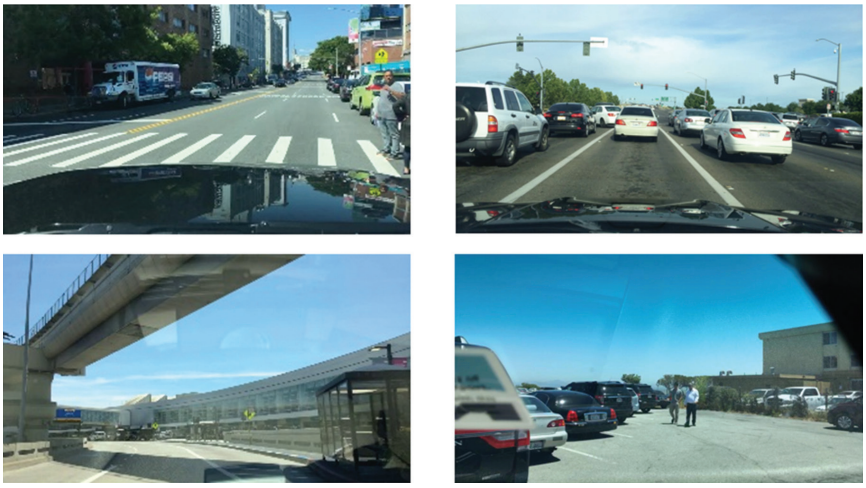


Fig. 1. Road scene images. From top left to bottom right - A crosswalk on the street, vehicles along a highway, a road under an overpass, and vehicles parked at a parking lot

The generated dataset consists of an equal number of images (700 each) for classes HR, VPL, and CRW, while RB/T included 625 images thus resulting in an unbalanced dataset. Fig. 2(a-d) below shows some examples that are difficult to distinguish since the significant objects are either missing or captured during the night due to which significant objects are not clear. Fig. 2(a) shows a highway without vehicles, Fig. 2(b) depicts vehicles parked at night, Fig. 2(c) is an underpass that is not clear, and Fig. 2(d) has a partial crosswalk covered due to vehicles.



(a)



(b)



(c)



(d)

Fig. 2. Sample road scene images from each category. (a) A deserted Highway. (b) A dimmed roadside parking at night. (c) A non-significant under-tunnel road. (d) An Occluded crosswalk

The classification framework is based on the fact that road scenes can be categorized with higher accuracy when the details in the scene images are extracted carefully to represent the road scene correctly. Researchers have suggested that the features extracted from the scene should carry details regarding the objects in the scene, the global or overall characteristics of the scene (image-level), and the fine or local details in the scene images (conventional/handcrafted). Due to the high resemblance among the variety of road scenes, sufficient discriminative information from the scene images is required to properly distinguish scene classes. Redundant information would certainly mislead the classifier, thus increasing the possibility of false detection. Therefore, the proposed scene classification framework is based on an efficient integrated feature-based machine learning approach. Diverse features including overall, fine, and object-based features are integrated using modified pre-trained networks and handcrafted or conventional descriptors. The overall or global features and the object-based features are extracted from the scene images using a modified pre-trained network VGG19 whereas the fine patch-based or window-based features are acquired using eight different descriptors.

3.1. OBJECT-ORIENTED FEATURES

– All the scene images are resized to 256x256 and the scene objects are detected using the YOLOV5m pre-trained network. The capabilities of the YOLOV5m network to identify 80 different objects are utilized to recognize objects in the scene images. Fig. 3 shows an example of object detection on a road scene using the YOLOv5m network. The identified objects include bicycles, cars, persons, and benches. Due to the varying dimensions of objects in the scene, the objects were priory detected from the scene image and then resized to a dimension for further feature extraction. The objective was to consider the contribution of every single object either small or large in dimension from the image. Experimental analysis displayed that every detected object from the scene should be resized to 32x32 so that their contribution is guaranteed. The strength of the feature vector was made dependent on the number of objects detected in the scene. The resized object was subjected to the feature extraction to a modified VGG19 network. The last layer of the VGG19 network was replaced with 1024 and 512 fully connected layers. Thus, for every single object a feature vector of 512 was obtained. The feature vectors obtained from different objects of a single image were then added to determine the strength of each element of the feature vector. The summation also mitigates the presence of zero values in the feature vector.

3.2. SCENE-LEVEL FEATURES

These features are directly obtained from the scene image. The original color image of size 256x256 is subjected to the pre-trained network (modified VGG19) and features are extracted from the image. For each image, a feature vector of 512 lengths was obtained and appended to the object-level feature vector.



Fig. 3. YOLOV5m object detection

3.3. CONVENTIONAL FEATURES

Local features from the scene images (reduced to half-dimension) were extracted using various global and local descriptors. The descriptors include matched filters (kernel-based and orientation-based), wavelets (coarse features using 6 wavelets and fine features using the haar wavelet), linear binary pattern (using 3x3 and 5x5 window), histogram of Gaussian, and Gray level co-occurrence matrix. Kernel-based matched filters [38], haar-based wavelet features [39], and LBP features [40] were used for extreme details whereas orientation-based match filters, wavelet-based (['bior3.1', 'bior3.5', 'bior3.7', 'db3', 'sym3', 'haar']), GLCM [41] and HOG [42] were used to contribute in terms of slight or medium details. The words extreme and medium correspond to details acquired from the image. The former represents details with more elements as compared to the latter one. The total number of feature elements corresponding to all the descriptors was 2310. The following Fig. 4 shows the variety of features and their dimensions.

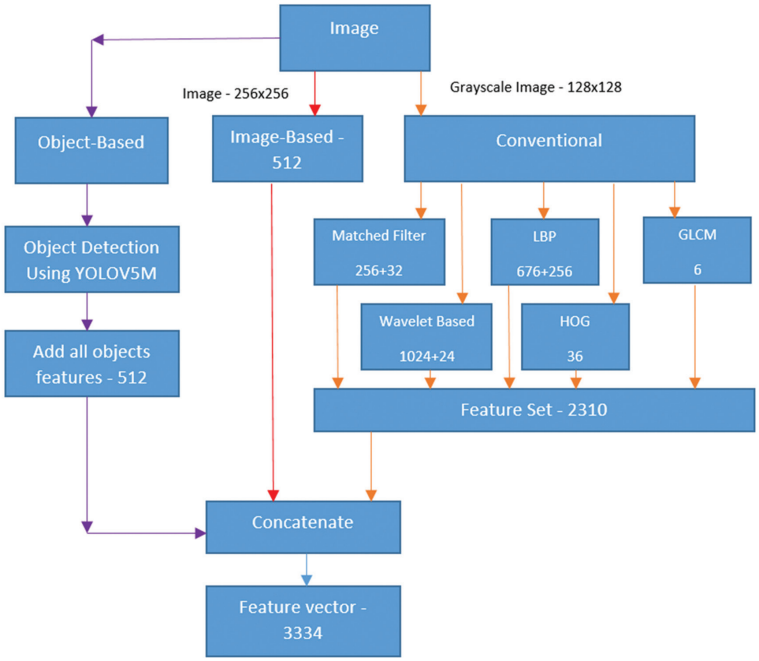


Fig. 4. Object-based, Image-based, and Conventional features with dimensions

Additional details for the conventional/handcrafted features can be found in [43]. These quality features improve the ability of the classifier network or a machine learning classifier. After all the dataset images are completely used for the feature extraction process, the features are normalized using the Max-normalization to fit the values in the range [0 1]. The normalization process was carried over the individual feature column while the missing values in the columns were substituted using the column mean. Fig. 4 shows objects detected by the YOLOv5m network from a street image and Fig. 5 (b-c) depicts the segmented objects from the crosswalk class image shown in Fig. 5(a).

The detailed scene classification framework is shown in Fig. 6. The global FEM uses the VGG19 network without the top layer directly on the input image resized to 256x256. The number of features extracted using the VGG19 network is 512. The blind features thus extracted depend on the scene information and ability of the

network. This is to ensure that regions not belonging to the objects detected using the local FEM contribute to the feature set. The only problems with such features are too many missing values which depend on the quality of the image. Even though the local features are considered using the two-stage deep network framework using the YOLOV5 and the VGG19 networks, the resizing stage for the detected objects may suffer from information loss. A size of 32x32 is considered to uplift the fine features concerning small objects but objects greater than 32x32 would suffer data loss. Therefore, we added fine and coarse features to the local and global features to improve the classification accuracy.

A total of 2310 HF are extracted using various feature descriptors which include wavelet-based, matched filter-based, LBP-based texture, GLCM-based, and the HoG features. The classifier (SVM) is used to learn the representation and predict the sample class for the assessment sample.

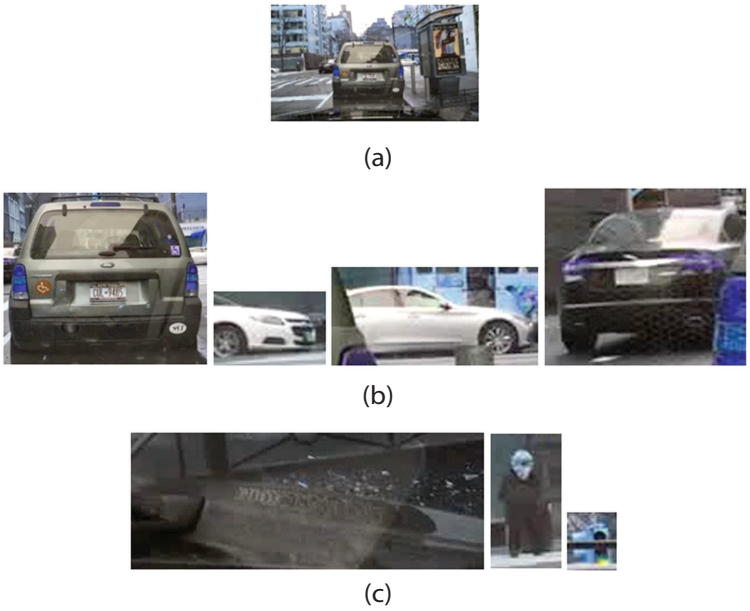


Fig. 5. Object detection using the YOLOV5m network from a single road scene image shown in (a) - Crosswalk Class image). (b)- Objects located in the image by YOLOV5m) Vehicles were detected at the scene. (c) - Other detected objects: keyboard, person and traffic light) Other detected objects include a keyboard, a person, and a traffic light.

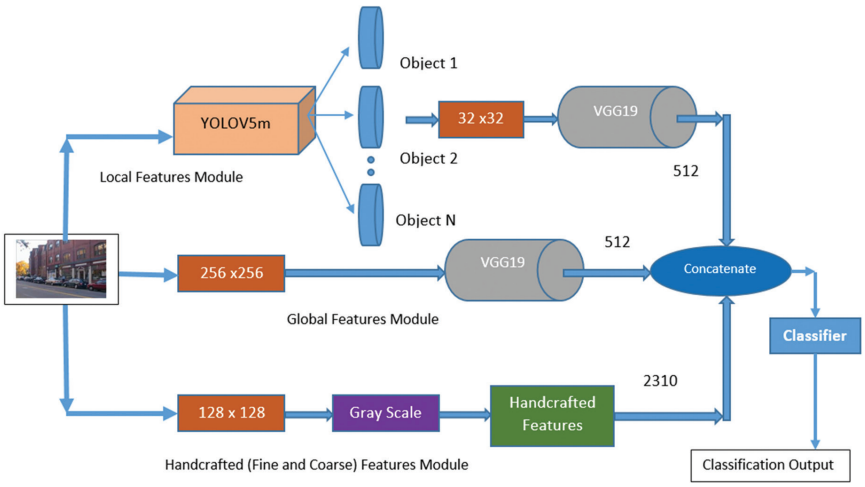


Fig. 6. The framework for the scene classification system

4. RESULTS

The performance metric that is used to evaluate the performance of the proposed road scene framework is classification accuracy. It is simply computed by calculating the ratio of scene samples correctly classified to total samples. It is expressed by the following expression (1):

$$\text{Accuracy} = \frac{\text{Number of scenes correctly classified}}{\textit{Total test samples}} \quad (1)$$

Heterogeneous features of all the 2275 images belonging to four different classes were extracted and stored in a CSV file. The proposed binary scene classification framework was developed in Python 3.9 on

Spyder 5, Windows 11 Environment, i5 Processor, 16 GB RAM, and 512 GB SSD. The feature samples were partitioned in the ratio of 80:20% for training, and testing randomly from the available set.

The training and testing sequence was iterated 10 times to note the classification accuracy between any two classes or categories at any instant. Support vector machine (SVM) with a 'Gaussian' kernel was used to train the feature samples and the maximum accuracy for any two classes was considered. Table 1 below shows the results obtained in terms of classification accuracy between two different categories considered for this research work. The heterogeneous features using different descriptors are uplifted and stored. Finally, the features are split as shown in Fig. 7 for training and testing. The SVM is trained on the training set to learn the scene representation and classify the test samples.

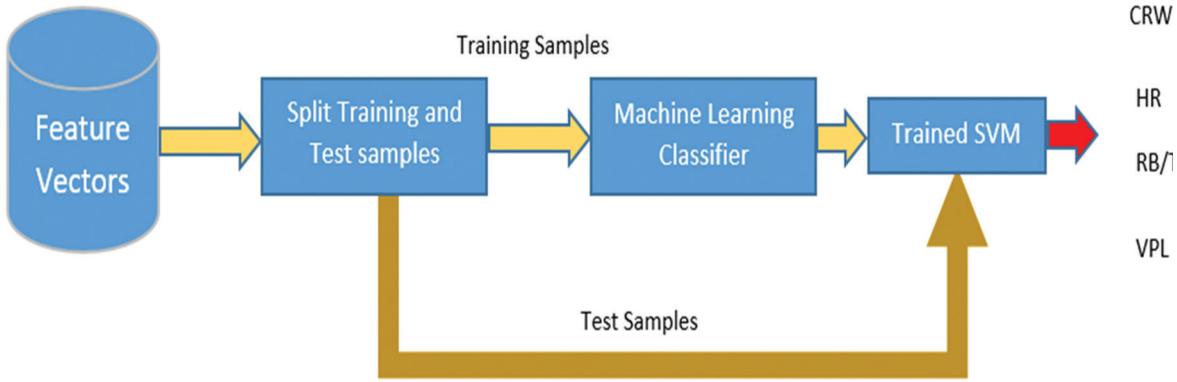


Fig. 7. The SVM-based classification model

Table 1. Classification Accuracy relating to Two Class performance

Class 1	Class 2	Accuracy-Training	Accuracy- Test
CRW	HR	100%	99.22%
CRW	RB/T	100%	93.87%
CRW	VPL	100%	99.85%
HR	RB/T	100%	97.54%
HR	VPL	100%	99.51%
RB/T	VPL	100%	95.74%

As seen from Table 1, the SVM was successful in training the samples with 100% accuracy. However, the test samples were not classified up to the 100% mark. The reason behind the low accuracy particularly in the case of CRW-RB/T, HR-RB/T, and RB/T-VPL is due to the multiple classes' existence in a single road scene image. Fig. 8 and Fig. 9 show some examples from the scene images. Fig. 8 below shows a crosswalk along with an underpass, a crosswalk below a tunnel, and a misleading crosswalk under a tunnel. Similarly, Fig. 9 shows vehicles parked beside a crosswalk, an unseen crosswalk under a tunnel, and far away parking with a crosswalk. Such images when falling under test samples would probably increase the chances of false detection. The research work uses the YOLOv5m network in its standard form and no transfer learning approach has been carried out to train the existing network for scene-based objects. The YOLOv5m network trained on the ImageNet dataset does not include several objects that are associated with road scene images. Therefore, significant objects are not detected by the network from the scenes which also amounts to the reason for low accuracy while detecting scene images. Also, no pre-processing is carried out to eliminate the uneven illumination effects caused by street-side lightning and vehicle lights. Several scene images were acquired during night, rain, and fog which need special attention.



(a) (b) (c)

Fig. 8. Multi-category scene samples.
(a) Crosswalk and an overpass. (b) Crosswalk under a tunnel and (c) Misleading crosswalk under tunnel.



(a) (b) (c)

Fig. 9. Closely correlated objects (Ambiguous) classes.
(a) Partial Crosswalk with parking (b) Unseen crosswalk under a tunnel and (c) Multiclass scene.

Similar work was suggested in [44] that used the contextual semantic relationship between the objects of the scene to classify the indoor and the outdoor scenes. The authors used visual attention regions marked with context-based saliency and deep CNN. They selected scene images from four different datasets including the MIT67, UIUC-Sports, LabelMe, and the Scene 15 dataset. They obtained a maximum classification accuracy of 97.70% over the LabelMe dataset and the lowest over the MIT67 dataset with 72.37%. However, the authors worked to differentiate indoor and outdoor scenes, the work carried out in this work is to classify different street scenarios which is more complex.

We have gone through several research papers to compare our results based on two-class road scenes however we found two research papers [27, 44] that included the categories considered in this work. Work introduced by Jianjun Ni et al. [27] particularly was oriented toward classifying the scenes into five different categories including crosswalks, gas stations, parking lots, highways, and streets using a deep network. While work by Jing Shi et al. [44] classified indoor and outdoor scenes utilizing a deep network. For comparison, we used a multi-class SVM and subjected all scene images for training and testing using the same ratio. Table 2 shows the total samples that were considered for the evaluation purpose.

Table 2. Number of sample images considered in each category for scene classification

Class	Class	Number of images
0	CRW	700
1	HR	700
2	RB/T	625
3	VPL	700

Although the work introduced by Jianjun Ni et al. and Jing Shi et al. is not comparable with our proposed work, we tried to obtain a better picture regarding the framework that considered diverse features extracted from the scene for multi-class configuration using SVM. The work differs in classes that were considered for the research. The competing models used deep learning networks for classifying the scenes, the proposed work utilizes a machine learning classifier. Table 3 shows the comparison between the competing models and results obtained through our proposed framework.

Table 3. Comparative test results based on Average Accuracy

Method	Categories	Classes	Average Accuracy %
Jin Shi et al.	2	Indoor, Outdoor	85.06
Jianjun et al.	5	CRW, Gas Station, HR, VPL and Street	75.99
Proposed Work	4	CRW, HR, VPL, and RB/T	86.01

Although the performance using our framework outperformed the other two competing models by approximately 1%, a lot of research is required in this area to incorporate a scene classification module in an automated vehicle. Better scene representation and advanced classifiers would obtain higher results and assist the unmanned vehicle over densely populated streets under rigorous road and climate conditions.

5. ABLATION STUDY

Maintaining the ML hyper-parameters, the researchers conducted experiments using any two sets of features at a time from object-oriented, scene-level, and conventional features. Table 4 shows the classification accuracies on 20% of test samples which were randomly chosen from the available samples about each of the categories. The average accuracies computed reveal that the object-oriented features and conventional features are crucial in classifying the scene classes but the scene-level features are essential to enhance the performance as seen in Table 1. Also, object-oriented features play an important role in representing the scene as seen from the last column of Table 2. Merely using the scene-level and conventional features would not differentiate complex scene images. Thus dropping any of the features has a greater influence on the performance.

Table 4. Classification Accuracies for Various Feature Combinations

Class 1	Class 2	Accuracy - Test Object-oriented & Scene-level features	Accuracy – Test Object-oriented & Conventional features	Accuracy - Test Scene-level & Conventional features
CRW	HR	91.10	93.94	89.68
CRW	RB/T	89.25	90.45	87.25
CRW	VPL	91.43	91.98	88.39
HR	RB/T	90.22	92.62	89.88
HR	VPL	92.15	93.94	90.10
RB/T	VPL	89.98	90.00	86.97
Average	**90.69**	**92.16**	**88.71**	

6. CONCLUSIONS

This article introduces a road scene classification framework based on heterogeneous features that are extracted at image-level, object-level, and local–level. The features that are extracted, are column normalized, and the missing entries are filled using the column mean. The feature samples are separated for training and testing in an 80:20 ratio and further classified using the support vector machine. The classification results that are obtained, reveal the proposed scene classification framework in classifying two classes that showed higher performance despite partial occlusions, ill-illumination due to diverse weather conditions, low inter-class disparities, multi-class ambiguities, and data imbalance.

There can be an improvement in the classification accuracy by adding quality preprocessing to mitigate the uneven illumination effects from the scene, YOLOv5m transfer learning for common road scene objects, other than the objects found in the ImageNet dataset, and using custom CNN networks. Due to heterogeneous features and three networks (YOLv5m and VGG19), the time required to extract features from the scene images is large. The future work will be based on the concentration of considering more than two classes (multiclass model) that will consider three or four classes as the classifier.

7. REFERENCES:

[1] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, "The Pascal visual object classes (VOC) challenge", International Journal of Computer Vision, Vol. 88, No. 2, 2010, pp. 303-338.

[2] O. Russakovsky, J. Deng, Su H., J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, "Imagenet large-scale visual recognition challenge", International Journal of Computer Vision, Vol. 115, No. 3, 2015, pp. 211-252.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, "ImageNet: A large-scale hierarchical image database", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20-25 June 2009, pp. 248-255.

[4] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, "Learning deep features for scene recognition using places database", Proceeding of the 27th International Conference on Neural Information Processing Systems, Vol. 1, 8 December 2014, pp. 487-495.

[5] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, "Places: A 10 million image database for scene recognition", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 40, No. 6, 2017, pp. 1452-1464.

[6] J. C. Rangel, M. Cazorla, I. García-Varea, J. Martínez-Gómez, E. Fromont, M. Sebban, "Scene classification based on semantic labeling", Advanced Robotics, Vol. 30, No. 11-12, 2016, pp. 758-769.

[7] I. Kostavelis, A. Gasteratos, "Semantic mapping for mobile robotics tasks: A survey," Robot Autonomous Systems", Vol. 66, 2015, pp. 86-103.

[8] S. Wang, Y. Li, W. Liu, "Multi-class weather classification fusing weather dataset and image features", Proceedings of the CCF Conference on Big Data, Springer, Xi'an, China, 11-13 October 2018, pp. 149-159.

[9] I. Sikirić, K. Brkić, P. Bevandić, I. Krešo, J. Krapac, S. Šegvić, "Traffic scene classification on a representation budget", IEEE Transaction on Intelligent Transportation System, Vol. 21, No. 1, 2019, pp. 336-345.

[10] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, "Vision meets robotics: The KITTI dataset", International Journal of Robotics Research, Vol. 32, No. 11, 2013, pp. 1231-1237.

[11] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27-30 June 2016, pp. 770-778.

[12] V. O. O. Castelló, I. S. S. Igual, O. Del Tejo Catalá, J. C. Perez-Cortes, "High-Profile VRU Detection on

Resource-Constrained Hardware Using YOLOv3/v4 on BDD100K", Journal of Imaging, Vol. 6, No. 12, 2020, p. 142.

[13] A. Boukerche, Z. Hou, "Object Detection Using Deep Learning Methods in Traffic Scenarios", ACM Computing Surveys, Vol. 54, No. 2, 2021, pp. 1-35.

[14] M. Mobahi, S. H. Sadati, "An Improved Deep Learning Solution for Object Detection in Self-Driving Cars", In Proceedings of the 28th Iranian Conference on Electrical Engineering, Tabriz, Iran, 4-6 August 2020, pp. 5-9.

[15] H. F. Yoshi, T. Hirakawa, T. Yamashita, "Deep Learning-Based Image Recognition for Autonomous Driving", IATSS Research, Vol. 43, No. 4, 2019, pp. 244-252.

[16] J. Chen, T. Bai, "SAANet: Spatial Adaptive Alignment Network for Object Detection in Automatic Driving", Image and Vision Computing, Vol. 94, 2020, p. 103873.

[17] C. R. Kumar, "A Comparative Study on Machine Learning Algorithms Using Hog Features For Vehicle Tracking Furthermore Detection", Turkish Journal of Computer and Mathematics Education, Vol. 12, No. 7, 2021, pp. 1676-1679.

[18] N. A. M. Mai, P. Duthon, L. Khoudour, A. Crouzil, S. A. Velastin, "3D Object Detection with SLS-Fusion Network in Foggy Weather Conditions", Sensors, Vol. 21, No. 20, 2021, p. 6711.

[19] M. Mirza, C. Buerkle, J. Jarquin, M. Opitz, F. Oboril, K. U. Scholl, H. Bischof, "Robustness of Object Detectors in Degrading Weather Conditions", Proceedings of the IEEE International Intelligent Transportation Systems Conference, Indianapolis, IN, USA, 19-22 September 2021.

[20] G. Al-refai, M. Al-refai, "Road Object Detection Using Yolov3 and KITTI Dataset", International Journal of Advance Computing Science and Applications, Vol. 11, 2020, pp. 1-7.

[21] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks", Advances in Neural Information Processing Systems, Vol. 25, No. 2, 2012, pp. 1097-1105.

[22] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, A. Mouzakitis, "A Survey on 3D Object Detection Methods for Autonomous Driving Applications", IEEE Transaction on Intelligent Transportation Systems, Vol. 20, No. 10, 2019, pp. 3782-3795.

[23] B. Lidestam, B. Thorslund, H. Selander, D. Näsman, J. Dahlman, "In-Car Warnings of Emergency Vehicles Approaching: Effects on Car Drivers' Propensity to GiveWay", Frontiers in Sustainable Cities, Vol. 2, 2020, p. 19.

[24] A. Agafonov, A. Yumaganov, "3D Objects Detection in an Autonomous Car Driving Problem", Proceedings of the International Conference on Information Technology and Nanotechnology, Samara, Russia, 26-29 May 2020, pp. 1-5.

[25] M. Alqarqaz, M. B. Younes, R. Qaddoura, "An object classification approach for autonomous vehicles using machine learning techniques", World Electric Vehicle Journal, Vol. 14, No. 2, 2023, p. 41.

[26] S. A. Khan, H. J. Lee, Huhnkuk, "Enhancing object detection in self-driving cars using a hybrid approach", Electronics, Vol. 12, No. 13, 2023, p. 2768.

[27] J. Ni, K. Shen, Y. Chen, W. Cao, S. X. Yang, "An improved deep network-based scene classification method for self-driving cars", IEEE Transaction on Instrumentation and Measurement, Vol. 71, 2022, p. 5001614.

[28] R. Prykhodchenko, P. Skruch, "Road scene classification based on street-level images and spatial data", Array, Vol. 15, No. 2, 2022, p. 100195.

[29] X. Jia, Y. Tong, H. Qiao, M. Li, J. Tong, B. Liang, "Fast and accurate object detector for autonomous driving based on improved YOLOv5", Scientific Reports, Vol. 13, 2023, p. 9711.

[30] Y. Li, J. Wu, H. Liu, J. Ren, Z. Xu, J. Zhang, Z. Wang, "Classification of Typical Static Objects in Road Scenes Based on LO-Net", Remote Sensing, Vol. 16, No. 4, 2024, p. 663.

[31] G. Dogan, B. Ergen, "A new CNN-based semantic object segmentation for an autonomous vehicle in urban traffic scenes", International Journal of Multimedia Information Retrieval, Vol. 13, No. 11, 2024.

[32] A. Chaudhari, "Smart traffic management of vehicles using faster RCNN based deep learning method", Scientific Reports, Vol. 14, 2024, p. 10357.

[33] J. Guo, J. Wang, H. Wang, B. Xiao, Z. He, L. Li, "Research on Road Scene Understanding of Autonomous Vehicles Based on Multi-Task Learning", Sensors, Vol. 23, No. 13, 2023, p. 6238.

[34] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, T. Darrell, "BDD100k: A diverse driving dataset for heterogeneous multitask learning", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13-19 June 2020, pp. 2636-2645.

[35] S. Goferman, L. Zelnik-Manor, A. Tal, "Context-aware saliency detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34, No. 10, 2012, pp. 1915-1926.

[36] R. McCall et al. "A taxonomy of autonomous vehicle handover situations", Transportation Research Part A, Policy and Practice, Vol. 124, 2019, pp. 507-522.

[37] C. Szegedy et al. "Going deeper with convolutions", Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7-12 June 2015, pp. 1-9.

[38] S. K. Saroj, V. Ratna, R. Kumar, N. P. Singh, "Efficient Kernel-based Matched Filter Approach for Segmentation of Retinal Blood Vessels", Solid State Technology, Vol. 63, No. 5, 2020, pp. 7318-7334.

[39] M. Wulandari, R. Chai, B. Basari, D. Gunawan, "Hybrid Feature Extractor Using Discrete Wavelet Transform and Histogram of Oriented Gradient on Convolutional-Neural-Network-Based Palm Vein Recognition", Sensors, Vol. 24, No. 2, 2024, p. 341.

[40] P. Lakshmi, M. Sivagami, "LT-LBP-Based Spatial Texture Feature Extraction with Deep Learning for X-Ray Images", Journal of Computer Science, Vol. 20, No. 1, 2024, pp. 106-120.

[41] C. I. Ossai, N. Wickramasingha, "GLCM and statistical features extraction technique with Extra-Tree Classifier in Macular Oedema risk diagnosis", Biomedical Signal Processing and Control, Vol. 73, 2022, p. 103471.

[42] D. R. Sulistyaningrum, T. Ummah, B. Setiyono, D. B. Utomo, Soetrisno, B. A. Sanjoyo "Vehicle detection using histogram of oriented gradients and real Adaboost", Journal of Physics: Conference Series, Vol. 1490, 2020.

[43] M. D. Narlawar, D. J. Pete, "Occluded Face Recognition: Contrast correlation & edge preserving enhancement based optimum features on CelebA dataset", Journal of Harbin Engineering University, Vol. 44, No. 8, 2023, pp. 1192-1204.

[44] J. Shi, H. Zhu, Y. Li, Y. Li, S. Du, "Scene classification using deep networks combined with visual attention", Journal of Sensors, Vol. 2023, No. 1, 2023. (retracted)