International Journal of Advance and Applied Research

www.ijaar.co.in

ISSN – 2347-7075 Impact Factor – 7.328

Peer Reviewed Bi-Monthly

Vol.10 No.6 July – August 2023

A Comprehensive Review Of Distinctive Approaches To Object Recognition Using Deep Learning

Mr. Prashant Bhat[1] & Dr. Anil Kumar[2]

[1]*Ph.D. Research Scholar, Department of Mechanical Engineering,*
Shri J.J.T. University, Rajasthan, India

[2]*Professor & Research Guide, Department of Mechanical Engineering,*
Shri J.J.T. University, Rajasthan, India

Corresponding Author - Mr. Prashant Bhat

DOI - 10.5281/zenodo.10029166

Abstract:

The human brain can detect the position of an item inside a picture and recognise it as soon as it sees it, but a machine requires time and a vast quantity of data to do the same operation. The human brain also does this in a much shorter length of time. When it comes to object recognition and classification, a deep neural network that is built on a convolutional neural network yields excellent results and a high level of accuracy. It takes a significant amount of time and a vast quantity of data (pictures and videos, for example) in order to train deep neural networks. Because the computational cost of computer vision is so high, the approach of transfer learning, in which a model that has been trained on one job is reused on another task that is linked to it, produces superior results. Authors have developed a variety of deep learning-based methods for object recognition and classification, such as You Only Look Once, Mask Region based Convolutional neural network, Fast Region based Convolutional neural network, and Faster Region based Convolutional neural network. An analysis of the similarities and differences between a number of distinct algorithms is presented here.

Keywords: *Deep Learning, Object Detection, Classification, Convolutional Neural Network, YOLO, R-CNN.*

Introduction:

When people look at a picture, they instantly understand what items are shown in the image, where those objects are located, and how the objects are connected to one another. On the other hand, if the calculations for the preparation of pictures could be accurate and rapid enough, personal computers would presumably be able to do autonomous driving without the need of specialised sensors, and assisting devices would probably provide customers with continuous scene data.

In a similar vein, if these computations were able to carry out Deep Learning tasks with the same level of expertise and quality of performance as humans do, then and only then would it be considered as true artificial intelligence. Therefore, the categorization of objects, the localisation of objects, and the detection of objects are the primary goals [1] of the process of image preparation; the primary challenges are accuracy, execution time, processing speed, and the financial sustainability of the Endeavour.

The problem of categorization is expanded in several ways, including via the process of object detection. When presented with a picture and a number of different classes of things, the purpose of object detection is to assess whether or not the image includes any objects that fall into the specified classes, as well as to pinpoint precisely where in the picture these objects are located. When it comes to the categorization of objects, it just concerns the nearness or non-nearness of these items. When it comes to semantic division, on the other hand, it seeks to describe lone pixels as being or not being a portion of an object of the specified classes. Object classification [2] is now a subtask of object detection, which is a subtask of semantic fragmentation as a result. The process of localising items in an image entails identifying the location of an object inside the picture, while the process of categorising things determines the possibility of an object being present in a picture. The algorithms that were used in the process of object localization provided the coordinates of the region that an object occupied in relation to the image.

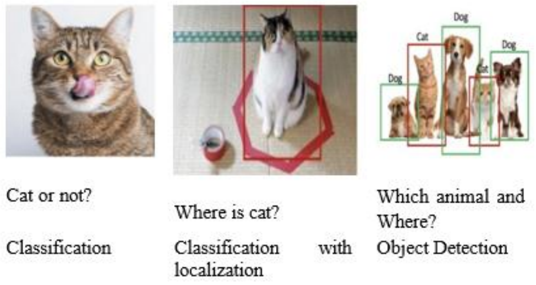

Figure 1: Common computer vision tasks. [3]

The researchers had investigated a variety of application areas of object detection, including the identification of faces and texts, detection of pedestrians and vehicles, detection of questionable identities via the use of surveillance systems, and analysis of medical images, amongst others. Object detection may be accomplished by either deep learning or machine learning techniques, which are the two primary methods that are used. In machine learning-based approaches, it detects the features using Haar, Scale Invariant Feature Transform (SIFT), and Histogram of Oriented Gradient (HOG), and then for classification it can use support vector machine (SVM) [4]. On the other hand, in deep learning approaches, it generally uses the Convolutional neural network for the object detection, which does not require any knowledge about the features. Background subtraction, the Gaussian mixture model, Region proposed Convolutional neural network (R-CNN) [5, quick R-CNN [6, Faster R-CNN [7], Mask R-CNN [8], and You only look once (YOLO) [9] are some of the different deep learning algorithms for object identification.

The confusion matrix is used in order to do the assessment of the procedure. The judgments that were made by the classifier may also be represented by another method known as the confusion matrix. An study of the results that are to be anticipated from a classification challenge is what is known as a confusion matrix. Count numbers are used to indicate the amount of right and wrong predicted outcomes, and these values are then sorted according to each class. At the moment when predictions are being made, the confusion matrix's job is to set out all of the ambiguities that the classification model must contend with. It enlightens us not just about the faults that a classifier is making, but also, and perhaps more importantly, regarding the kinds of errors that are being produced. In the confusion matrix, there are four different categories: [10] data entries that have been successfully categorised as positives are referred to as "True-Positives" (TP). False-Positives (FP) [10] are records that should have been marked as negative but were instead given the positive designation. True- Negatives (TN) [10] refers to those negatives that have been categorically identified as negatives. Last but not least, the term "false-Negatives" (FN) [10] refers to good cases that have been mistakenly classified as negative. In table 1, the following is a representation of the structure of a confusion matrix for binary classification:

Table 1: Binary classification confusion matrix

	Actual-Positive	Actual-Negative
Predicted-Positive	True-Positive	False-Positive
Predicted-Negative	False-Negative	True-Negative

Recall and accuracy are the two evaluation metrics that were produced using the confusion matrix as their source. Precision [11] may be defined as the total quantity of positive instances that are properly categorised, also known as true-positive examples, divided by the total quantity of positive examples that are labelled by the system, also known as (true positive + false positive examples). It is possible to judge it by:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall [27] is a metric that may be defined as the entire quantity of positive instances that are properly categorised, i.e. true-positive examples, divided by the total quantity of positive examples that are included in the data, i.e. (true positive + false negative examples). It is possible to judge it by:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Object Detection:

Deep Learning-based algorithmic frameworks There are certain machine learning algorithms that don't make use of deep learning, yet deep learning algorithms are all considered to be within the category of machine learning algorithms. Estimation models that rely on Deep Learning are constructed consisting of a number of hidden layers. These layers aid in the process of learning data representation along with consideration at each level. The fundamental focus of deep learning is on the algorithms used in deep neural networks, where "deep" refers to the number of hidden layers in the network. The primary goal of deep learning is to solve learning problems by modelling their solutions after the operation of the human brain. Deep learning is being employed by the system, and despite changes in the structure of the system, it is continuously becoming better. The best approach to improve the system's performance is to either do some alterations that are reliant on the original system or use certain tricks. The following is a list of the numerous algorithms that may be used for object identification and classification.

R-CNN:

Region-based First, an area search is carried out by the convolutional network, as the name suggests, and then classification is carried out. A procedure known as region search may be used to find an item inside a picture. The exhaustive search approach was created by J.R.R. Uijlings et al. [5] in 2012 as an alternative to the selective search method, which is one of the ways that may be used to search for regions. The selective search method is one of the methods that can be used. It starts by calculating tiny sections or areas in an input picture, and then it groups those computations into a hierarchical structure. As a result, the complete picture is contained inside the final group, which functions as a hierarchical structure. When it comes to the process of grouping the identified areas, there are two aspects that are taken into consideration: the colour space and the similarity metrics. Therefore, the final picture is created by grouping together a number of smaller areas, which ultimately results in a proposal for a given number of regions.

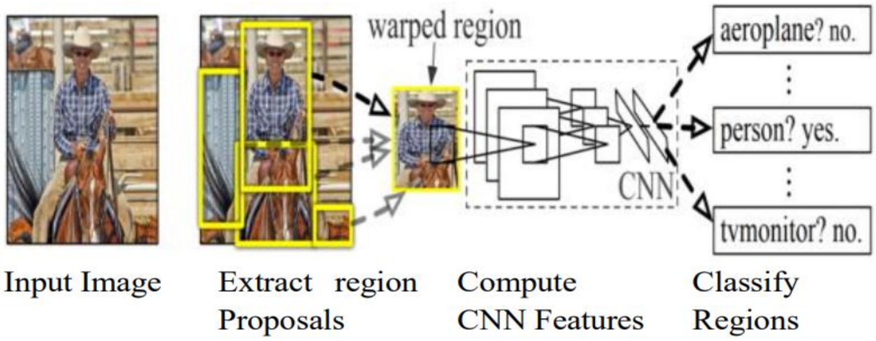

Figure 2: Region-based Convolutional neural network (RCNN) Framework. [5]

In R-CNN, the detection of area proposals is accomplished via the use of a selective search methodology. Subsequently, deep learning is used for the purpose of determining the locations of objects inside the discovered region proposals. After that, CNN is used, and in order to fit the input size of the CNN, the individual area suggestions are scaled. This assisted in the extraction of the feature vectors with a dimension of 4096. These feature vectors are employed as an input by the numerous classifiers that are used in the process of predicting the likelihood for each class. Then, the likelihood of recognising the objects by making use of these feature vectors is projected for specific classes that already have a pre-trained support vector machine (SVM). The usage of linear regression in area proposal, which alters the shapes and sizes of the bounding boxes, may be used to reduce the amount of inaccuracy that occurs while attempting to localise objects.

FAST R-CNN:

Fast Region-based Convolutional Network (Fast R-CNN) was developed in 2015 by R.Girshick et al. [6]. Although it is comparable to R-CNN in some respects, its primary goal is to reduce the amount of time needed for evaluating each region proposal that is associated with a significant number of models.

In rapid R-CNN, the whole picture is taken into consideration as an input for the CNN, which makes use of numerous convolutional layers, in contrast to traditional R-CNN, which needs CNN for each proposed part of the image. On the feature maps that are produced by CNN, the selective search strategy has been employed in order to locate the area of interest (RoI). By using the RoI pooling layer to lower the size of the feature maps, one may more accurately determine the area of interest while also satisfying the length and breadth requirements as the primary criterion. The result of each individual RoI layer's analysis is used as an input for the fully-connected layers, which then generates a feature vector as their final output. After that, a softmax classifier is used in conjunction with these feature vectors in order to identify the objects, and a linear regressor is employed in order to adjust the localization of the objects. Afterwards, the object detection process is complete.

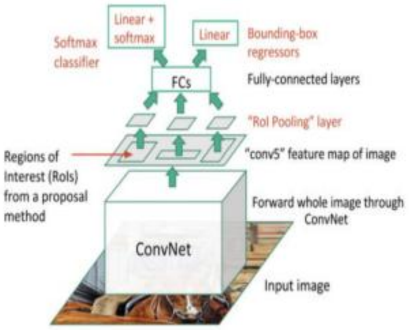

Figure 3: Fast region-based Convolutional network (Fast R- CNN) Framework. [6]

FASTER R-CNN:

In 2016, S.Ren et al. [7] developed a new method for the detection of objects, prediction of bounding boxes, and region proposal generation. This method is known as region proposal network (RPN), and it was designed to overcome the cost issue that is present in the traditional method, which makes use of selective search methods for the generation of region proposals. Therefore, the faster region proposal Convolutional neural network may be computed by the combinations of models that use the Region Proposal Network and the fast region proposal Convolutional neural network. An full picture is taken into consideration here as an input, and that information is then used to build feature maps. The generation of a feature vector that is linked to two layers that are entirely connected occurs when a window of size 3x3 is dragged over the whole of a feature map. There are two layers, both of which are completely linked. The first layer is used for box regression, while the second layer is used for box classification. When using layers that are completely linked, a large number of region ideas are found. When the upper limit of k regions is set to be permanent, the size of the output for the box regression layer is 4k, while the size of the output for the box classification layer is 2k. When a k area proposal is found by using a sliding window, this kind of proposal is referred to as an anchor.

Following the detection of the anchor boxes, the score was decided upon based only on the usefulness of the boxes that had been produced by the application of a threshold to the objectness. The feature maps and anchor boxes that are produced as a result of the main CNN model are the outputs that are then used as an input for the rapid R-CNN model.

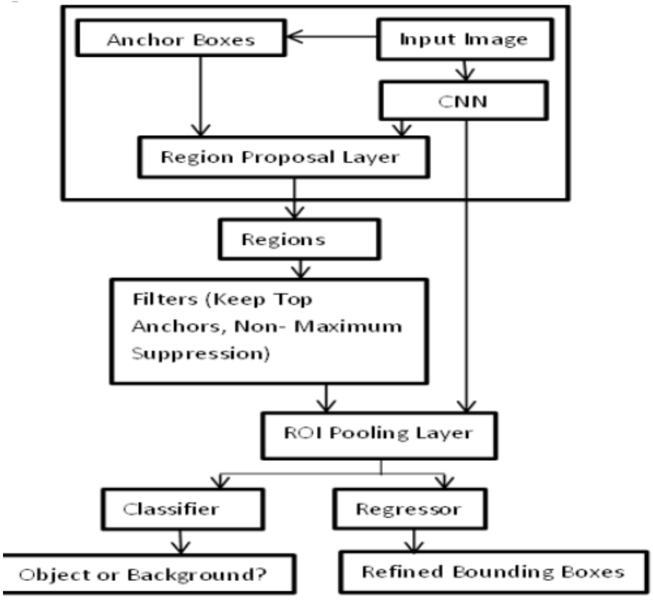

Figure 4: The basic architecture of Faster Region-based Convolutional neural network.

RPN, as opposed to the selective search approach, is employed by Faster R-CNN so that training and testing speed may be increased, and performance can be improved. RPN is applied over the ImageNet dataset, which was used for pre-training, and then it is applied over the PASCAL VOC dataset, where it is fine-tuned. This allows for classification. The last step in training a rapid R-CNN involves the generation of region proposals coupled with anchor boxes. As a result, it is a process that repeats itself.

Mask R-CNN:

K. He et al. [8] created the Mask Region-based Convolutional Network (Mask R-CNN) in 2017. This network was designed to function concurrently for the detection of bounding boxes, which aids in the prediction of object masks. The mask is understood to represent the splintering of the objects in the picture into individual pixels. When compared to other strategies, such as those for the identification of a key point, detection of bounding box and object, and instance fragmentation, the performance of these techniques in the COCO challenges has been superior for the past four times in a row. The Faster R-CNN software is employed by the Mask Region-based Convolutional Network, also known as the Mask R-CNN. This aids in the production of three outputs, which are an object mask, an offset for bounding boxes, and a label for classes. In addition to this, it makes use of the Area proposal network (RPN) in order to generate bounding boxes, and it concurrently generates all three outputs for each and every region of interest (RoI).

When running a faster version of R-CNN, the main RoIPool layer is switched out for a RoIAlign layer. After that, the elimination of the division of the real RoI coordinates and the evaluation of the localization comes next. The items that make up the region. After that, two convolutional layers are added to the second branch, and ultimately, the third branch is the one that is employed for the detection of the object's mask. The actions that are carried out by these three different branches are connected to various loss functions, all of which are eventually brought together. Finally, in order to create improved performance, this combined value is decreased. This is done because successfully completing the fragmentation job improves object localisation, which in turn improves the classification rate.

YOLO:

In the YOLO approach, the detection of bounding boxes and the prediction of the probabilities associated with them are both accomplished via the use of a single assessment and a single neural network model. Because it is capable of making predictions in real time, using it requires very little effort.

It is possible to describe how the YOLO approach works by noting that it takes a full picture as an input and then fragments the image into aSxS network. The B bounding boxes are then predicted along with their confidence ratings by utilising individual cells of this network. It is possible to define the confidence score as the product of the likelihood of identifying objects and the value of the IoU included inside the ground truth boxes and the anticipated values.

The GoogleNet model of CNN is used in YOLO, and an inception module has been suggested [9]. It has two fully connected layers that are followed by 24 convolutional layers. The inception module may be switched out for a 3x3 convolutional layer that is followed by a dimensionality reduction layer and any one of the following filters: 1x1, 2x2, or 3x3 filters. There are three different versions of YOLO (v1, v2, and v3), with the most recent version, YOLO v3, being the quickest of the three. It has nine convolutional layers, a specific number of filter proposals, and the RoIAlign layer, which supports translation-equivariance and scale-equivariance.

In this method, the picture itself is used as an input for ResNet, which then processes it using 101 layers. In order to process the RoIs that have been discovered, the RoIAlign layers are employed. This network has a fully-connected layer that is linked to it via one of its three branches. These branches assess the bounding box coordinates and forecast the likelihood of the In YOLO, the output generated by the last layer for predicting every cell of the grid is S*S*(C+B*5), where S*S represents the total size of the grid, C represents the number of estimated probability for each class, and B represents a limited number of anchor boxes for each cell (these boxes are connected to the cell's confidence score and its four coordinates).

When classifying images, we make use of the ImageNet dataset, which has already been pre-trained with more than fifty percent of the Convolutional layers.

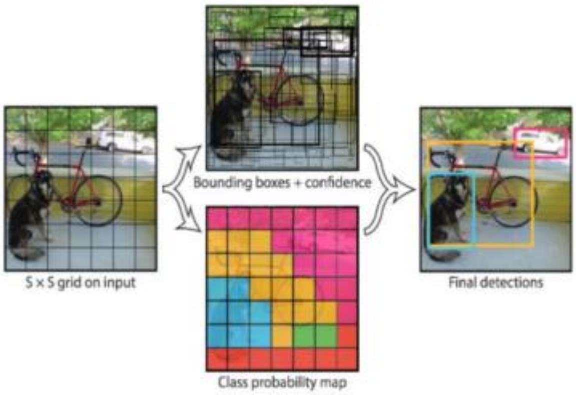

Figure 5: You Only Look Once (YOLO) framework [9]

When more traditional approaches are taken into consideration, the majority of the time, object detection is accomplished by the use of the prediction of bounding boxes. When it comes to the YOLO approach, there are a significant percentage of bounding box predictions that do not include the item. The Non-Maximum Suppression (NMS) strategy is used at the very last node of the network in order to remedy this problem. It does this by combining the bounding boxes of extraordinarily coincident identical items into a single box, however there is still a possibility of the identification of some false positives due to the nature of the algorithm.

Comparison of Various Algorithms

Algorithm	**Advantages**	**Disadvantages**	**Application Areas**
R-CNN	It is the first neural network based on region proposal for achieving higher detection quality.	It requires a large amount of data, energy, processing, and duration.	It can be used where the data set is large.
FastR-CNN+ VGG16	The quality and speed of detection are improved.	It is an expensive method as the region proposals are obtained using another model.	It can be used where quality and speed is a major concern as it out performs the R-CNN method.
FasterR-CNN +VGG	Region proposals are obtained using region proposal network.	Speed is the only drawback.	It is faster than fastR-CNN as it utilizes region proposal network instead of selective search method.
FasterR-CNN +Inception ResNet V2	Improve the speed 3 times when using 50 proposals instead of 300.	Region proposal is limited to 50.	It can be used for faster processing of object detectors.
SSD + MobileNet	It is fast and gives the best accuracy. It out performs FasterR-CNN.	Poor results for small objects.	It can be used where accuracy is a Major concern.
MaskR-CNN +ResNet101	Best object instance segmentation and bounding box object detection results are achieved.	It is simple but the small amount of over head is added to faster R-CNN.	It can be used for segmentation purpose.
YOLOV1	It is fast and out performs DPM and R-CNN.	It increases localization errors and Not good for small objects.	It can be used for detection of large objects.

Conclusion:

In this research, we have provided a synopsis of the many different deep learning strategies that may be used for the identification and categorization of objects. We have just scratched the surface of the inner workings of R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN, and YOLO at this point. As a means of drawing a conclusion from this body of research, a summary of a number of algorithms and a comparison of their underlying models, including pros and cons as well as potential uses, will be presented. The comparison result demonstrates that the R-CNN method may be used in situations when the dataset is huge. On the other hand, the YOLO v1 technique is used in situations where the objects being detected are larger in size since it is both quicker and more effective than the R-CNN algorithm. Even though R-CNN provides a greater quality of detection, rapid R-CNN combined with VGG16 provides both a higher quality and a faster rate of object detection. Because it uses the region proposal approach rather than the selective search method, the faster R- CNN + VGG algorithm is much quicker than its predecessor, fast R-CNN + VGG 16. R-CNN Enhanced Speed Plus Inception It is possible to implement the ResNet V2 algorithm in order to speed up the object detector's processing. When compared to the speedier R-CNN method, the SSD + MobileNet technique provides the highest level of accuracy; nevertheless, it is unable to provide superior results when dealing with smaller objects. The best results for bounding box object identification and segmentation are achieved by combining Mask R-CNN with ResNet 101.

References:

[1]. Du, J. (2018). Understanding of Object Detection Based on CNN Family and YOLO. Journal of Physics Conference Series 1004(1):012029 , 1-8.

[2]. Zhao, Z.-Q., Zheng, P., Xu, S.-T., & Wu, X. (2019). Object Detection With Deep Learning: A Review. IEEE Transactions on Neural Networks and Learning Systems , 1- 21.

[3]. Grover, P. (2018, February 15). Towards Data Science.

[4]. Retrieved April 16, 2019, from https://towardsdatascience.com/evolution-of-object- detection-and-localization-algorithms-e241021d8bad

[5]. Sun, Z., Bebis, G., & Miller, R. (2006). Monocular precrash vehicle detection: features and classifiers. IEEE Transactions on Image Processing , 2019 - 2034.

[6]. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2015). Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence , 142 - 158.

[7]. Girshick, R. (2015). Fast R-CNN. IEEE International Conference on Computer Vision (ICCV) (pp. 1-9). Santiago, Chile: IEEE.

[8]. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R- CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems 28 (NIPS 2015) (pp. 1-14). Neural Information Processing Systems Foundation, Inc.

[9]. He, K., Gkioxari, G., Dollar, P., &Girshick, R. (2017). Mask R-CNN. 2017 IEEE International Conerence on Computer Vision (ICCV) (pp. 2980-2988). Venice, Italy: IEEE.

[10]. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1-10). IEEE.

[11]. Davis, J., &Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. 23 rd International Conference on Machine Learning, (pp. 1-8). Pittsburgh, PA.

[12]. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing and Management , 427-437.

[13]. Harsha, S. S., & Anne, K. R. (2016). Gaussian Mixture Model and Deep Neural Network based Vehicle Detection and Classification. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 9 , 17-25.

[14]. Zhou, Y., Nejati, H., Do, T. T., Cheung, N. M., & Cheah, L. (2016). Image-based Vehicle Analysis using Deep Neural Network: A Systematic Study. IEEE International Conference on Digital Signal Processing (DSP). Beijing, China: IEEE.

[15]. Gao, Y., & Lee, H. J. (2015). Moving Car Detection and Model Recognition based on Deep Learning. Advanced Science and Technology Letters , 57-61.

[16]. Kaur, R., & Talwar, M. (2016). Automated Vehicle Detection and Classification with Probabilistic Neural Network. IJARIIT , 1-4.

[17]. Chan, Y. M., Huang, S. S., Fu, L. C., Hsiao, P. Y., & Lo, M.F. (2012). Vehicle detection and tracking under various lighting. IET Intelligent Transport Systems , 1-8.

[18]. Berg, A., Ahlberg, J., &Felsberg, M. (2015). A Thermal Object Tracking Benchmark. 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1-7). Karlsruhe, Germany: IEEE.

[19]. Bhartee, A. K., Srivastava, K. M., & Sharma, T. (2017). Object Identification using Thermal Image Processing. International Journal of Engineering Science and Computing, 11400-11401.

[20]. Rodin, C. D., Lima, L. N., Andrade, F. A., Haddad, D. B., Johansen, T. A., &Storvold, R.

(2018). Object Classification in Thermal Images using Convolutional Neural Networks for Search and Rescue Missions with Unmanned Aerial Systems. 2018 International Joint Conference on Neural Networks (IJCNN) , 1-8.

[21]. Nam, Y., & Nam, Y. C. (2018). Vehicle classification based on images from visible light and thermal cameras. EURASIP Journal on Image and Video Processing , 2-10.

[22]. Moranduzzo, T., &Melgani, F. (2014). Detecting Cars in UAV Images With a Catalog-Based Approach. IEEE Transactions on Geoscience and Remote Sensing (Volume: 52 , Issue: 10 , Oct. 2014) , 6356 - 6367.

[23]. Sivaraman, S., & Trivedi, M. M. (2013). Integrated Lane and Vehicle Detection, Localization, and Tracking: A Synergistic Approach. IEEE Transactions on Intelligent Transportation Systems, Vol. 14, No. 2,June 2013 , 906-917.

[24]. Chen, Y.-L., Chen, T.-S., Huang, T.-W., Yin, L.-C., Wang,S.-Y., &Chiueh, T.-C. (2013). Intelligent Urban Video Surveillance System for Automatic Vehicle Detection and Tracking in Clouds. IEEE 27th International Conference on Advanced Information Networking and Applications (AINA) (pp. 814-821). Barcelona, Spain: IEEE.

[25]. Tuermer, S., Kurz, F., Reinartz, P., & Stilla, U. (2013). Airborne Vehicle Detection in Dense Urban Areas Using HoG Features and Disparity Maps. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing , 2327 - 2337.