Object Recognition in Image Sequence using Heuristic Convolution Neural Network

Mahesh Kini M

Abstract: ***The human visual system can make a distinction of tiger from cat very easily without taking any efforts. But in case of a computer system, it is a very complicated job. Identifying and differentiating task has to deal with many challenges but the human brain makes it effortless. Self learning or heuristic techniques are most relevant in this area. The recognition task is to search for the particular object of same shape, color and texture and so on, of the trained objects and match with input. The geometrical distinction such as zoom in, zoom out, rotation etc result in poor performance. This paper uses convolution neural network models Alexnet and VGGNet on object recognition problems which are added with novel heuristic method. We have used CIFAR-10 dataset. The performance and computation speeds are found efficient.***

Keywords: component; convolution neural networks; object recognition; feature extraction; classification; AlexNet; VGGNet;

I. INTRODUCTION

A object recognition system that are trained during learning phase recognizes the target object from the inputs. However, recognizing a geometrically variant image is a complex computer vision problem. The object recognition problem we have separate two phase: *feature extraction* and *classification*. The feature extraction phase deals with finding the feature of the objects like shape, color and so on. The output of the feature extraction phase , i.e., learning model, is given as input to classification phase. The classification phase discriminate the feature of each object.

An image of an video sequence is a two dimensional array pixels. These pixels portray the features of the image like the intensity, color etc. A grey scale image contains a two dimensional array of pixels and a color image contains three set of two dimensional pixel array for red, blue and green channels.

The human brain can easily distinguish between the tiger and cat image but for computer it is a challenging task. The image is just some array of pixel values for computers. The human brain (is a complex structure made of neurons) made the detection task easy to humans. The neural network is a technique that works exactly as brain, contains interconnected neurons exchanges messages between them. .

Each interconnection is assigned with some weights and these weights are optimized during training phase. In neural network a set of various object views is given as the input to the network. The classification and feature extraction are done by the network. The network is organized in multiple layers. So that the first layer detects very basic features like lines, edge etc with considering the zoom in, zoom out, rotation of the object and the higher layers detects complex features like shape. The back propagation method is used to reduce the error rate. In the training phase the weights of nodes are adjusted to improve the performance

A traditional image recognition algorithm has been applied to the hand-crafted features like SIFT features. The features are viewed based on Bag-of-visual-words model and took after by clustering algorithm or Learning algorithm like Support Vector Machines (SVMs). In clustering method, similar images are grouped using various clustering algorithms and extracted the center of the cluster as the key feature to represent that cluster. The weakness of this method is the visually not similar scene can have similar SIFT values. So there may have a case were visually dissimilar image classified under same cluster. Accordingly, these algorithms performance crucially relies upon the features utilized and that is the primary weakness of the traditional image recognition methods. The researcher or the programmer needs to choose which feature should use for the particular query. As year passed the image became more complex and these bringing up about a trouble with choosing feature extraction. In the meantime, a few researchers developed models based on machine learning models. These are models comprised of feature extraction from multi layers. This layered approach lead to beginning of deep learning models. Some early deep learning models like RBM (Restricted Boltzmann Machines) and DBN (Deep Belief Networks) got great outcomes on small datasets. These are unsupervised learning models. Unsupervised learning allows an approach to the problem with no or little idea of what our outcome should resemble. We can obtain structure from data where we don't essentially know the impact of the variable.

In 1998 LeCun et al. [1] proposed LeNet architecture was the first convolution neural networks. The LeNet architecture is small and efficient network. It can even run on CPU or GPU and was mainly utilized for handwritten digit or character recognition with minimal preprocessing. A convolutional neural network with gradient descent based learning algorithm and back propagation is to increase overall performances. The features are extracted in convolution layers. The features are not explicitly defined as in traditional recognition methods. The features are learned from the experience. So many training data is required as the model is built without any prior knowledge. The pooling layer is applied to decrease the dimension of the features.

In 2012, Alex Krizhevsky et al. [2] proposed ALexNet architecture was the first model, reduced the error rate to 15.3%, beating all other available approaches. It was an unsupervised machine learning model and feature are extracted automatically by the model. The AlexNet gave significant break to deep

Revised Manuscript Received on October 05, 2019.

Mahesh Kini M, Department of Computer Science and Engineering, N.M.A.M. Institute of Technology, (Visvesvaraya Technologcal University, Belagavi) Nitte 574110, Udupi District, India.

learning model and several architectures were proposed to reduce the error rate further.

In traditional neural network the input image is flatten to one dimensional array and given to network. The issue with the network is it requires huge memory space and complex computations. If the image size is 250X250 then 62,500 nodes are required. The convolutional neural network model is the one the best model of neural networks for object recognition. This paper uses the convolutional neural network models AlexNet and VGGNet on image recognition problems.

II. RELATED WORK

Srinivas et al. [3] describes convolutional neural network in detail. They featured feature extraction from the traditional image classification, object classification methods and the beginning of the convolutional network to the recent development in convolutional neural network. This main drawback of traditional image classification and object detection algorithm is the features are explicitly defined. The accuracy of the result depends on the feature defined. If we selected wrong features then the performance of the model decreases. The multilayer approach leads to the beginning of deep learning. In the convolutional neural network the feature are not explicitly defined and the weights of neural network are adjusted during training. This paper is the detailed study of the convolutional neural networks like AlexNet, RNN, multi layer model and hybrid CNN model. They also mentioned some of the current demerits of the CNN: More training data is required; CNN is give false result for artificial images, robust for small geometric changes etc.

Browne et al. [4] describes the architecture of convolutional neural network and some real world examples of convolutional neural networks in Robotics. They described the land mark detection problem and sewing pipes crack detection problem using CNN.

Cai et al. [5] proposed a 3D CNN model to detect the strange behavior in the examination of surveillance. This model can also be applied to the abnormal behavior they are not yet designed. The “optical flow” is calculated for the training and testing dataset using fameback’s algorithm. The optical flow cannot process directly so it is converted into flow image. The flow image is given to the 3D CNN to build a model. The CNN for 3D has the ability to learn the spatial and behavior in the video clip. The main difference between 2D and 3D CNN is the spatial dimensions. The model has the ability to work with number of abnormal behaviors and performs better then motion blob, template matching and skin SVM algorithms.

Sun et al. [6] proposed a semantic attribute based approach on deep convolutional networks. These semantic attributes are automatically discovered from a joint image and text corpora. Standard *corenlp* toolkit is used to automatically discover the caption from the image. The input frames are passed to the trained CNN which extract the visual features of the image. The visual and semantic features are fused by vector concatenation. Bundling center clustering algorithm is used to cluster the frames. The some frames from the center of the cluster are chosen to represent the video shot. The number of the cluster is obtained by dynamic programming approach. The approach will not work well in title-based approach.

Li et al. [7] proposed SingleNet; this method aggregates various layers of features and guides them to uniform space to reduce the error rate and to increase the performance. Various dense boxes are used to approximate the localization of the objects in the image. The feature map is generating by using a VGG16 full convolutional network and feature maps from each layer are aggregated by sum operation and un-pooling. Lastly the bounding box and classification is applied to the various feature maps.

Xiao et al. [8] proposed a salient object detection model with the combination of eye tracking data; to mainly focus on the human interested sessions and super pixel segmentation. Initially the input images are preprocessed. Simple linear clustering algorithm is used convert the segmented image into super pixels and the adjacent super pixel pairs are grouped with eye tracking data.

Guo et al. [9] proposed a real-time object detection method based on multiple feature fully convolutional network. This method predicts the class and percentage with very less detection computation speed. The anchor box technique is used to localize the object in the image.

Ahmad [10] proposed an object detection method in compressed MPEG-I videos. This method produce smooth boundary detection and solid motion vector information. Initially the motion vector are extracted from the P-frames of the input MPEG-I video. I-frames are used to extract the DCT information. The backgrounds are segmented using texture characteristics and finally the objects are localized directly in the DCT domain.

Bazi, and Melgani [11] proposed an object detection method for unnamed aerial vehicles (UAVs); the convolution support vector machine network. The model contains numerous convolutional layers and reduction layers similar to pooling layers and linear SVM at end. Unlike the CNN the weights are tuned using forward supervised learning strategy .A set of linear SVMs are used as filter in each convolutional layers. The network will perform well small training data.

Sachdeva et al. [12] proposed a bag of visual model for image –to- image matching problems. Initially the local feature of the images are extracted to distinguish image using SIFT algorithm and then clustered as bag of visual word. Using Euclidean distance between the each bag of visual are calculated . The some top neighbors are extracted. This method work well even with geometrically invariants images and need less data for training.

Oliveira et al. [13] proposed a Fast and Light Weight Object Detection (FLODNet) model based on Convolutionalneural network which is faster than CNN model and can execute in CPU within few seconds. The FLODNet is a shallow convolutional neural network with 10 layers and fixed kernel size. This model initially resize the image and given to the network. The return the class name and confidence score and finally the output image are stretched to look similar to input image and bounding box are assigned.

Tenguria et al. [14] proposed object detection model in Robotics. The method basically contain three parts; hardware, communication channel and software. The Raspberry Pi camera is used to capture the image and SMTP protocol is used to communicate to the software. The software is YOLO network trained on COCO dataset. The camera is controlled by robot. According to object detection the robot manages the actions.

III. PROPOSED METHOD

The proposed method is a two dimensional convoultional neural with several convolution layers, pooling layers and activation layers and one or two full connected layer at the end. The CIFAR-10 dataset is used for training. The Fig.1 is the architecture of proposed model. The method consists of three phase. Preprocessing phase, training phase and evaluation phase respectively. The raw input image cannot give as the input because the image may be in different dimensions. So initially the processing method rescales the input image dimension into 32X32 and some data augmentation method like rotation and zoom in applied. The preprocessed images are given to CNN models to train the model and evaluated with testing set.

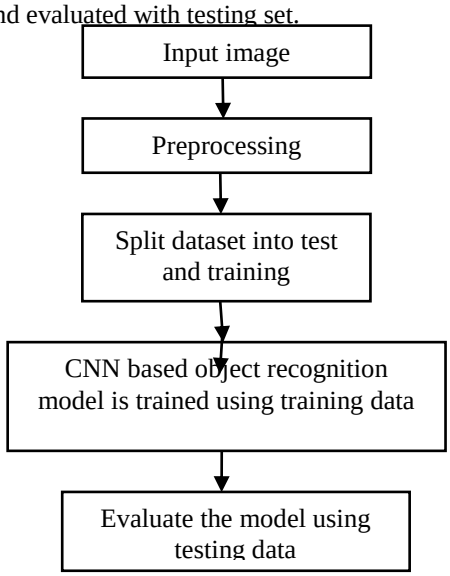

Fig.1 The proposed method

A. Dataset

CIFAR_10 dataset is used for training the model. The dataset contains ten object classes like cat; dog etc. The Fig 2 is the sample CIFAR-10 dataset images with corresponding classes. This dataset consists of about 50,000 training set images and 10,000 testing images. The each class contains different variety images i.e. the cat class contains image of different variety cat from small cat to big cat in different angles and rotations.

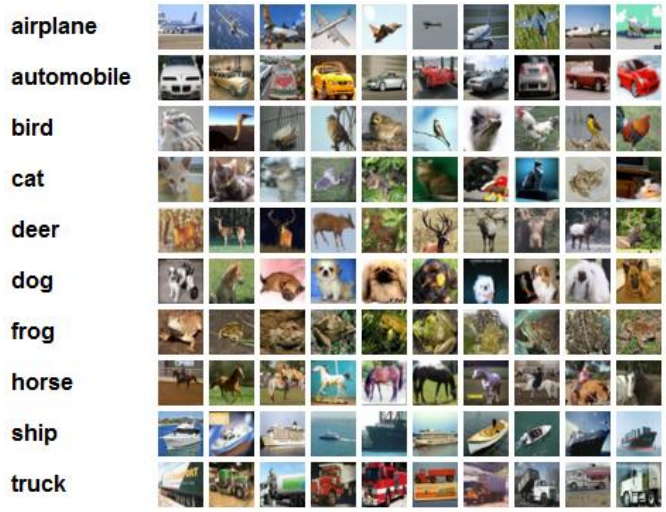

Fig.2 Sample CIFAR-10 data with class label and images

B. Model Architecture

The model will be the combination of convolutional, max-pooling and activation layers. The convolution layers extract various features. The first layer extract basic feature and next layer extract feature of first layer and so on. The input images are of dimension 32X32X3. The convolution operation is the element wise multiplication of image and kernel over the image. The weights are shared among the neurons is the main advantage over neural network.

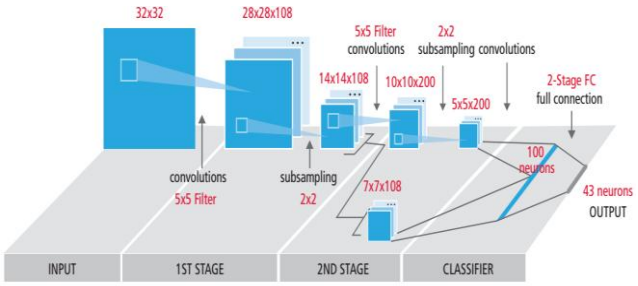

Fig.3 Convolutional neural network

The max-pooling reduce the dimension of the image. The 2X2 max-pooling layer segments the image into 2x2 blocks and select maximum value from four values is selected from each segment. ReLU activation layer reduce the non-linearity, it convert all the negative value to zero and fully connected layers are added at the end for classification. The last fully connected layer contains 10 neurons. 10 is the number of classes. In fully connected layer all the neurons are connected. So before first fully connected layer, the two dimensional array is flatten to one dimensional vector. The softmax activation function is to map the probability of class match in between 0 to 1. The class with highest percentage of class label is predicted as result. The dropout was introduced in some layers to reduce over fitting. The Fig.3 is the sample convolutional networks with two hidden layers.

C. AlexNet

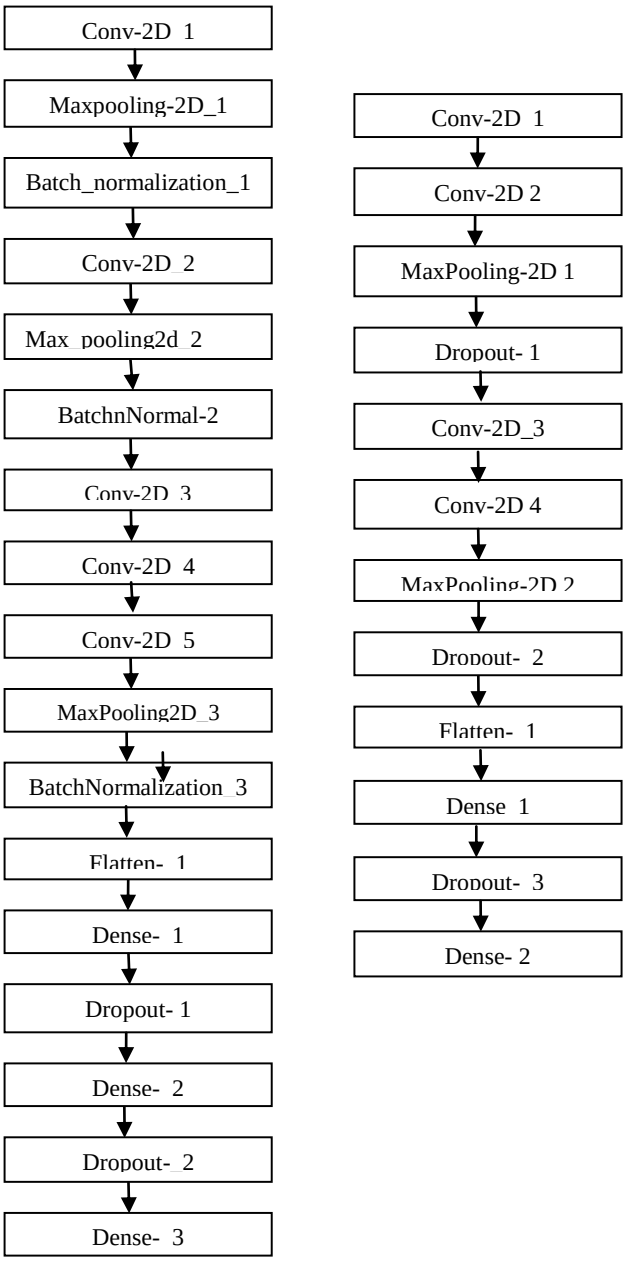

Fig.4 1 Alexnet Layer Architecture And 2) Vggnet Layer Architecture

Alex Krizhevsky et al. [2] proposed ALexNet reduced the error rate to 15.3%.The AlexNet is made up of eight layers. The 3X3 is taken as the size of kernel. The maxpooling is applied to first, second and fifth layer. The batch normalization is applied to same layers as max pooling layer. Finally the dropout layer is applied to last two fully connected layers. The Fig.4 first block diagram is the architecture of the AlexNet.

D. VGGNet

Karen Simonyan and Andrew Zisserman proposed 19 layers CNN model layers with 3x3 kernels, 2x2 maxpooling layers, stride and padding 1. The CIFAR-10 dataset size is 32X32, so the number of layer is reduced to 6 layers. The max pooling layers is applied after every two convolutional layers. The dropout layer is applied after max pooling and there after in fully connected layers. The Fig.4 second block diagram is the architecture of the VGGNet.

IV. RESULT AND DISCUSSIONS

All the training and testing are done on Google colaboratory GPU based system with NIVIDA K80. The k80 have 4992 cores. The python programming language upon Keras toolkit with Tensorflow as backend is used as for programming. All the training is done for 100 epochs with 32 batch size. The Figs 5-8 are the graphs obtained after training and testing.

Fig.5 AlexNet train loss vs. validation loss

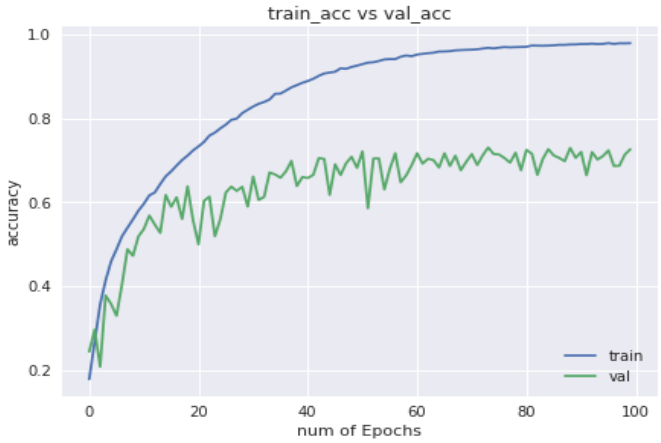

Fig.6 AlexNet train accuracy vs. validation accuracy

The fig.5 describes the loss details of AlexNet for training and testing data over hundred epochs. For the training data initially the error rate was high as the epochs increased the error rate is decreased near to zero. For the testing data initially the error rate was low as the epochs increased the error rate is increased near to 4. The fig.6 describes the accuracy details of AlexNet for training and testing data over hundred epochs. For training data initially the accuracy was low as the epochs increased the accuracy reached 1(100%). For the testing data initially the accuracy was low as the epochs increased the accuracy reached around 70.

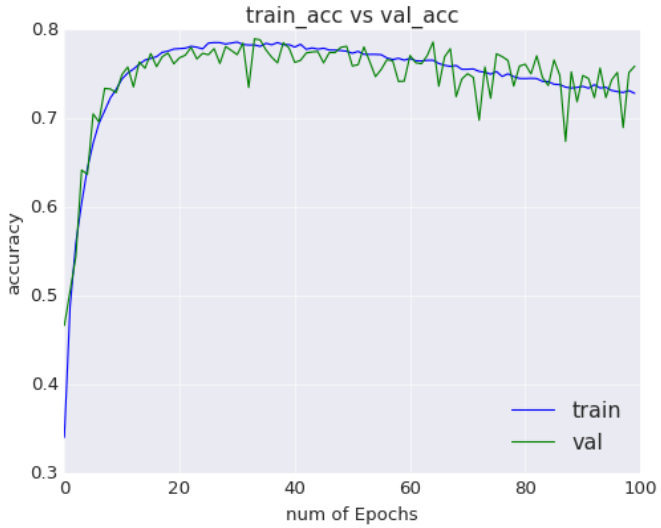

Fig.7 VGGNet train accuracy vs. validation accuracy

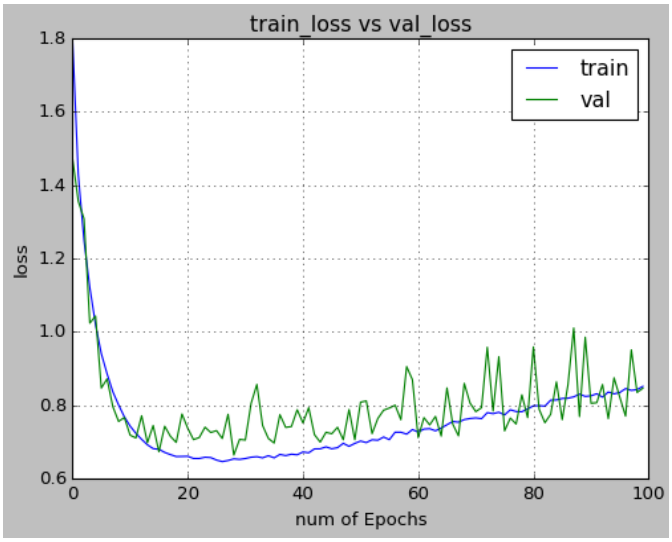

Fig.8 VGGNet train loss vs. validation loss

The fig.7 describes the loss details of VGGNet for training and testing data over hundred epochs. For training and testing data initially the error rate was high as the epochs increased the error rate is decreased near to 0.8. The fig.8 describes the accuracy details of VGGNet over hundred epochs. For training and testing data initially the accuracy was low as the epochs increased the error rate is increased near to 0.8 (80%). The table 1 compares the performance and speed of AlexNet and VGGNet. The AlexNet took 50 minutes to execute and gave the average test accuracy as 72.6% but VGGNet took only 18 minutes and average test accuracy is 75.88%.

TABLE I
Results of two models

	Test Score	Accuracy (%)	Time(sec)
AlexNet	3.5530472145080565	72.6	3017.552
VGGNet	0.8461340543746948	75.88	1134.176

V. CONCLUSION AND FUTURE WORK

In the recent development in object recognition, many techniques and algorithms have been proposed. In this paper we used the Convolution neural network model AlexNet and VGG16 heuristically. From the result it is clear that VGG16 is better over AlexNET. The CNN model VGG16 found to be better in performance and speed than AlexNet.

The convolution neural network normally have one or more number of hidden layers, as the hidden layer number increases the performance increases. The dropout layer is introduced in CNN to circumvent over-fitting. In training process, high computation is done so high performing system is required to reduce the computation time. In future, the image recognition problem can be implemented in new deep convolution networks.

REFERENCES

1. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, Gradient-based learning applied to document recognition. IEEE, 86(11), 2278-2324.
2. Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems 25, 1097–1105. Curran Associates, Inc.
3. Srinivas, S., Sarvadevabhatla, R. K., Mopuri, K. R., Prabhu, N., Kruthiventi, S. S., & Babu, R. V. (2016). A taxonomy of deep convolutional neural nets for computer vision. Frontiers in Robotics and AI, 2, 36.
4. Browne, M., Ghidary, S. S., & Mayer, N. M. (2008). Convolutional neural networks for image processing with applications in mobile robotics. In Speech, Audio, Image and Biomedical Signal Processing using Neural Networks (pp. 327-349). Springer, Berlin, Heidelberg.
5. Cai, X., Hu, F., & Ding, L. (2016, December). Detecting Abnormal Behavior in Examination Surveillance Video with 3D Convolutional Neural Networks. In Digital Home (ICDH), 2016 (pp.20-24). IEEE.
6. Sun, K., Zhu, J., Lei, Z., Hou, X., Zhang, Q., Duan, J., & Qiu, G. (2017,July). Learning deep semantic attributes for user video summarization. In Multimedia and Expo (ICME), 2017 IEEE (pp.643-648).
7. Li, J., Qian, J., & Yang, J. (2017, September). Object detection via feature fusion based single network. In Image Processing (ICIP), 2017 IEEE (pp. 3390-3394).
8. Xiao, F., Peng, L., Fu, L., & Gao, X. (2018). Salient object detection based on eye tracking data. Signal Processing, 144, 392-397.
9. Guo, Y., Guo, X., Jiang, Z., Men, A., & Zhou, Y. (2017, September). Real-time object detection by a multi-feature fully convolutional network. In Image Processing (ICIP), 2017 IEEE International Conference on (pp. 670-674). IEEE.
10. Ahmad, A. M. (2006). A Novel Object Detection Technique in Compressed Domain. In Computer Vision and Graphics (pp. 689-694). Springer, Dordrecht.
11. Bazi, Y., & Melgani, F. (2018). Convolutional SVM Networks for Object Detection in UAV Imagery. IEEE Transactions on Geoscience and Remote Sensing.
12. Sachdeva, V. D., Fida, E., Baber, J., Bakhtyar, M., Dad, I., & Atif, M. (2017, December). Better object recognition using bag of visual word model with compact vocabulary. In Emerging Technologies (ICET), 2017 13th International Conference on (pp. 1-4). IEEE.
13. de Oliveira, B. A. G., Ferreira, F. M. F., & da Silva Martins, C. A. P. (2018). Fast and Lightweight Object Detection Network: Detection and Recognition on Resource Constrained Devices. IEEE Access, 6, 8714-8724.
14. Tenguria, R., Parkhedkar, S., Modak, N., Madan, R., & Tondwalkar, A. (2017, April). Design framework for general purpose object recognition on a robotic platform. In Communication and Signal Processing (ICCSP), IEEE 2017 International Conference on (pp. 2157-2160).

AUTHORS PROFILE

Mahesh Kini M is pursuing Doctorate in Computer Science and Engineering from Visvesvaraya Technological University, Belagavi, Karnataka. Research is on Video Analysis based on Machine Learning Predictions and done M.Tech in Computer Science and Engineering at N.M.A.M.I.T, Nitte under VTU, Belagavi. He has total academic teaching experience over 15 years and has published papers in National and International Journals. Artificial Intelligence, Image Processing and Computer Vision are the area of interest. He is also member of various National and International professional societies. Currently he is serving as Assistant Professor at N.M.A.M. Institute of Technology, Nitte, Karnataka.