Object Detection and Recognition in Images

[1]Sandeep Kumar, [2]Aman Balyan, [3]Manvi Chawla
Computer Science &Engineering Department,
Maharaja Surajmal Institute of Technology, New Delhi, India.

Abstract **- Object Recognition is a technology in the field of computer vision. It is considered to be one of the difficult and challenging tasks in computer vision. Many approaches have been proposed in the past, and a model with a new approach which is not only fast but also reliable. Easynet model has been compared with various other models as well. Easynet model looks at the whole image at test time so its predictions are informed by global context .At the prediction time, our model generates scores for the presence of the object in a particular category. It makes predictions with a Single network evaluation. Here object detection is a regression problem to spatially separated bounding boxes and associated class probabilities.**

Index Terms **- Computer vision, image detection, Feature Extraction.**

I. INTRODUCTION

All Object Recognition has two parts- Category Recognition and its detection [4] .Category Detection deals with distinguishing the object from the background. And Category Recognition deals with classifying the object into one of the predefined categories. It is a identifying process of specific object in a digital image or video. Generally, Object recognition algorithms rely on matching, learning, or pattern recognition algorithms using appearance-based or feature-based techniques [5]. For example, it is used to find instances of real life objects like bicycles, fruits, animals and buildings in images or videos.

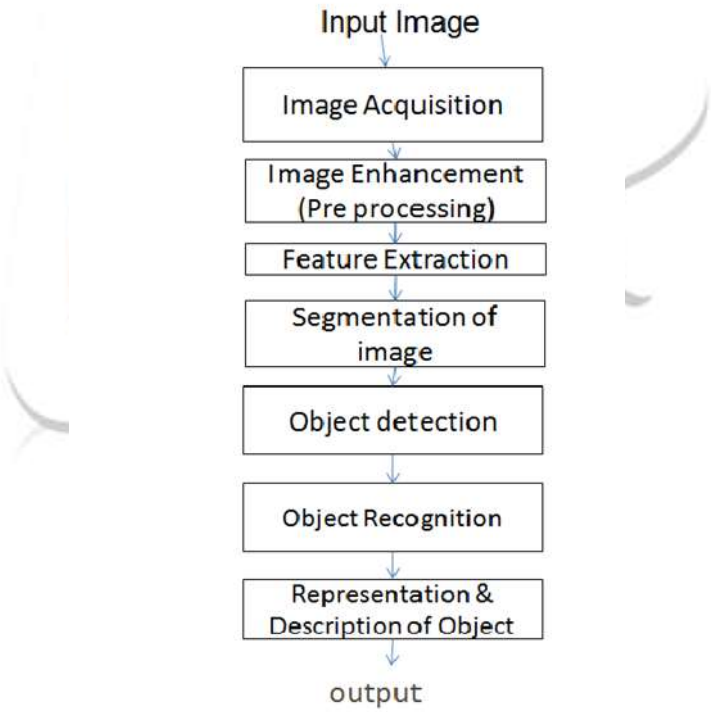

Fig 1: Model Diagram

As shown in fig1. the Object detection algorithms use features which can be extracted to recognize a particular object. This model is very simple and easy to implement. Here, object detection is a single regression problem which detects directly from bounding box coordinates and class probability. Every object has its own class such as all circles are round, which are used while recognizing the objects.
follow.

II. PREPROCESSING

It is the lowest level of abstraction. The process of preprocessing improves the image intensity by suppressing the unwanted features or enhancing them for further processing.[2].It resizes the image size to 448*448 and also normalizes the contrast and

brightness effects. The image is also cropped and resized so that feature extraction can be performed easily. The input images are pre-processed and very easily normalize the contrasts and brightness. Preprocessing step can be done by subtracting the mean of image intensities and divide by the standard deviation. New brightness value can be found by using the neighborhood of a pixel in the input image. The fig 2 below shows the preprocessing of image.

Fig 2. Preprocessing (The image has been resized) [9].

III. FEATURE EXTRACTION

Its main motive is to simplify the image by considering only the important information and leaving out the extra information which is not necessary for recognition. It uses the method of edge detection which can only retain the essential information [3].It represents the reduced part of an image as a feature vector. This approach is used when the size of image is very large. Hence, by this process, image recognition becomes easier. It starts from the already measured data and features which provides some kind of information facilitating the further steps.

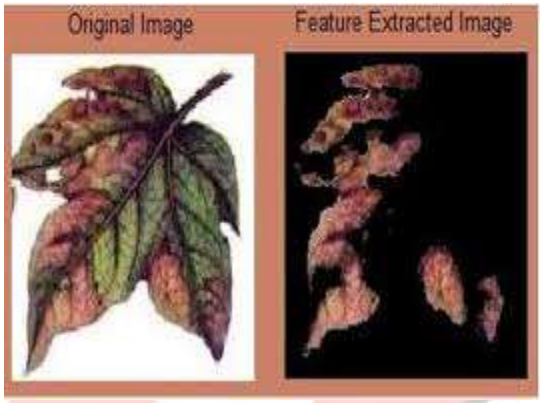

Fig 3: Feature Extraction (The image has been simplified considering only the important information) [10].

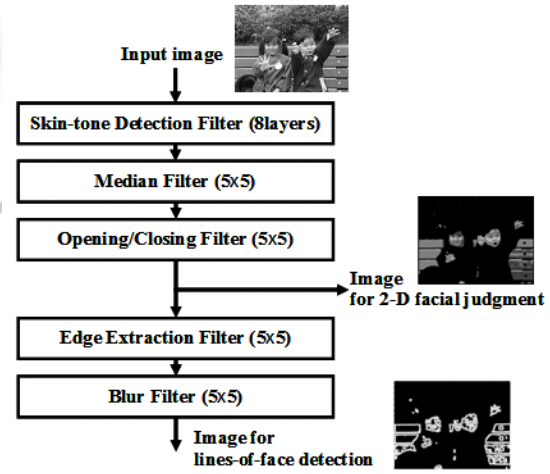

Fig 4: Combined process of Preprocessing and feature extraction [9].

IV. DETECTION IN IMAGES

The different components of object detection are integrated into a single neural network, which uses features from the whole image to predict a bounding box [4].

The bounding boxes for other classes are also predicted at the same time.

Hence the neural network analyses the full image and also the different objects in the image.

The image is input to the system which is divided into a grid of SxS cells. If the center of our image falls in a grid cell, it is responsible for analyzing that object. A grid cell predicts B bounding boxes. A bounding box is a rectangle enclosing an object. Each box has a confidence score corresponding to it, which shows a percentage indicating the extent to which it is certain that the box actually encloses some object. This score doesn't tell us anything about the nature of the object in the box. If no object

exists in a cell , the confidence score is zero[1]. For every bounding box, the cell also predicts a class from all the possible classes of our dataset. The confidence score for a box and class prediction are combined into a single score that tell us the probability that this particular bounding box contains a specific type of object [1]. Every Bounding box has 5 parameters: x, y, w, h, and confidence. The x and y coordinates represent the center of the bounding box. The width (w) and height (h) are predicted for the image and the confidence score is also predicted. For the PASCAL VOC dataset, a 7x7 grid is used i.e. S=7 and 2 bounding boxes for each cell i.e. B=2. As the PASCAL VOC has 20 classes so C= 20. Hence, our final prediction is 7x7x30 tensor. As there are 7x7= 49 grid cells and 2 bounding boxes for each cell, the total number of bounding boxes turns out to be 98. Most of these have very low confidence scores and are thus discarded.

V. DESIGN

A convolutional neural network is used for our model. A convolutional neural network is similar to an ordinary neural network and contains neurons and weights for each neuron [7]. While a regular neural net doesn't scale well to take full images as input, a convolutional neural net can take large images as input and their architecture is designed accordingly. Three main layers are used for a convolutional neural net: Convolutional Layer, Pooling Layer, and Fully-Connected Layer [8]. The initial layers of the neural net are used for feature extraction while the fully connected layers predict the coordinates and output probabilities. This network has 24 initial convolutional layers and 2 fully connected layers.

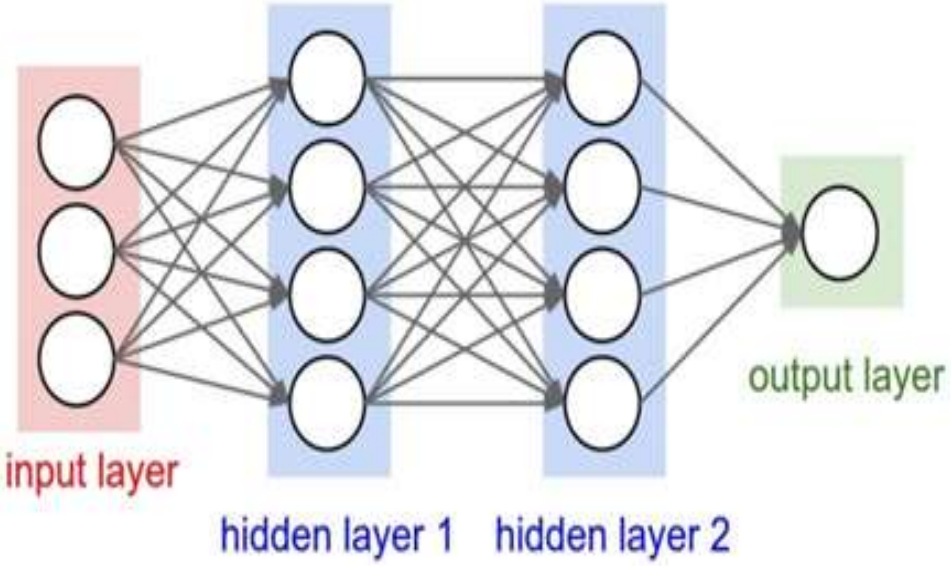

Fig 5 : Convolution neural network[10].

Finally, an input image (resized to 416x416 pixels) is passed to the convolutional neural net in a single pass, which comes out as a 7x7x30 tensor, describing the bounding boxes for grid cells. The final scores for the bounding boxes are calculated and the ones having low scores are discarded [5], [9].

VI. DATASET

The PASCAL VOC (visual object classes) 2007 is a dataset which contains 9,963 images belonging to 20 different classes. The classes are mentioned below -

- Person: person.
- Vehicles: bicycle, motorbike, bus, car, train, boat, aero plane.
- Animals: bird, cat, dog, cow, sheep, horse.
- Indoor objects: bottle, tv/monitor, chair, dining table, sofa, potted plant.

Here is a sample image which shows one object from each of the 20 different classes -

Fig 6: Dataset PASCAL VOC 2007[11].

VII. CHALLENGES FACED IN OBJECT RECOGNITION

Change in size, cropping out the background are some of the factors influencing the accuracy of the system. The accuracy of the model might change by scaling the image.

Adjusting Brightness and Contrast of the image may also make it difficult for the system to recognize the objects in the image.

There may be cases when the object might not be visible enough for the system to recognize it. The Object Recognition System must handle these cases of low visibility.

The system may fail in cases where similar objects occur in groups and are too small in size.

Various lightning conditions and shadows in the image may also pose difficulty for the system to recognize the object[6].

VIII. APPLICATIONS OF OBJECT DETECTION AND RECOGNITION

1. Self-Driving Cars-Self Driving Cars may use Object detection and recognition system to identify pedestrians and cars on the roads and then make the suitable decision in accordance.
2. Face Detection-Another application of Object detection and recognition is Face Detection .e.g.- Facebook recognizes people before they are tagged in images.
3. Medical Science-Object Detection and recognition system may help Medical science to detect diseases. For e.g.- Detecting Tumors and various cancers.
4. Text Recognition-Text recognition deals with recognizing letters/symbols, individual words and series of words. Ex- Recognizing handwriting of a person.
5. Hand Gesture Recognition- Hand Gesture Recognition deals with recognition of hand poses, and sign languages.

IX. COMPARISON WITH OTHER DETECTION SYSTEMS

This model has been compared with other Detection Systems such as RCNN (Region based Convolution Neural Network), FASTER RCNN, SDD (Single Shot Detector) using PASCAL VOC 2007 dataset.

1. RCNN-RCNN uses Selective Search to create the Bounding Boxes. RCNN looks at the image through windows of different sizes [9]. It extracts the region proposals and then pass them though the CNN to generate CNN features. At last it adds SVM(support vector machine) which
   helps in classifying whether there is an object in the region proposed ,and if yes then what object it is .RCNN produces around 1800-1900 bounding boxes, while our system produces only around 100 which is far less than that produced in RCNN.
2. FASTER RCNN-FASTER RCNN is similar to RCNN except that it uses ROIPOOL (Region of interest Pooling).It runs the CNN just once per image and shares its computation to other sub regions. Faster RCNN thus uses only one pass of the original image.
   It can also be used for region proposals.
   It has mAP of about 70.
   Its drawback is real time performance, which this model overcomes.

X. EVALUATION METRICS

Method	mAP	FPS	Batch size

RCNN	66.1	6	1
FASTER - RCNN	73.2	7	1
EASYNET MODEL	69.4	45	1

Fig 7 : Evaluation metrics (Comparison of Easynet with RCNN and Faster RCNN)

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP(q)}}{Q}$$

Mean average Precision (MAP) for a set of queries is the Mean of the average precision scores for each query, where Q is the number of Queries.

XI. OUTPUTS

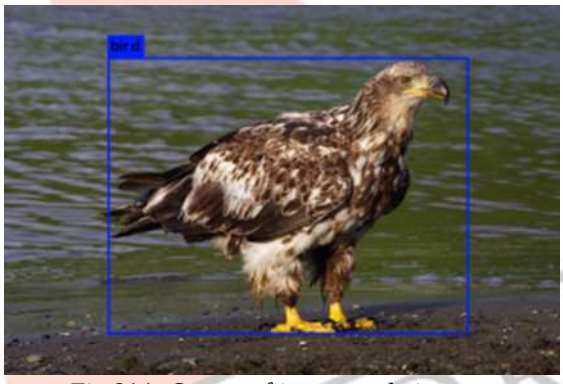

Fig 8(a): Output of image eagle.jpg.

OUTPUT : Bird

ACCURACY : 91%

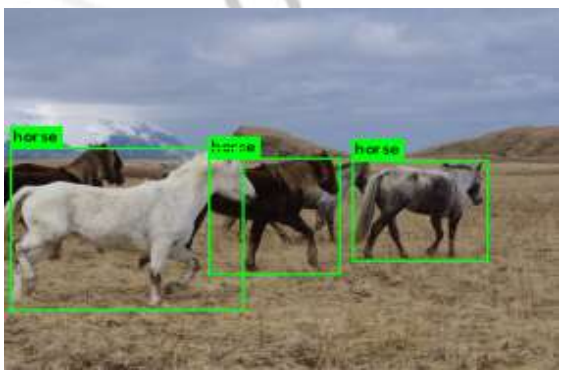

Fig 8(b): Output of image horse.jpg

OUTPUT : Horses
ACCURACY : 86%

XII. CONCLUSION

Easynet model is very simple to implement and build. It is unified for object detection. It generalizes the domains and can be trained easily on full images. It can also consist of object tracking along with detection. Also, acquiring the pertained Dataset PASCAL VOC 2007 made the work easier and hence the model could be implemented on hardware with no interruptions. Different types of identification can be done and multiple objects can be detected by Easynet model. In object detection, background subtracting approach has been used when an image is taken from a single camera with a static background. In future, the work can be extended by detecting the moving objects with non-static background.

REFERENCES

[1] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.2016.
[2]Bazeille, Stephane, et al. "Automatic underwater image pre-processing." *CMM'06*. 2006.
[3] Lazebnik, Svetlana, Cordelia Schmid, and Jean Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories." *Computer vision and pattern recognition, 2006 IEEE computer society conference on*. Vol. 2. IEEE, 2006.
[4] Yeo, Boon-Lock, and Bede Liu. "Rapid scene analysis on compressed video." *IEEE Transactions on circuits and systems for video technology* 5.6 (1995): 533-544.
[5] Belongie, Serge, Jitendra Malik, and Jan Puzicha. "Shape matching and object recognition using shape contexts." *IEEE transactions on pattern analysis and machine intelligence* 24.4 (2002): 509-522.
[6]LeCun, Yann, Fu Jie Huang, and Leon Bottou. "Learning methods for generic object recognition with invariance to pose and lighting." *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. Vol. 2. IEEE, 2004.
[7] Hinton, Geoffrey E., et al. "Improving neural networks by preventing co-adaptation of feature detectors." *arXiv preprint arXiv:1207.0580* (2012).
[8] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.
[9] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014
[10] Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems*. 2015.
[11] Khan, Fahad Shahbaz, et al. "Color attributes for object detection." *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.