Automatic Object Tracking using Deep Learning Technique

Prof. Jyoti Wadmare[1], Mr. Fenil Desai[2], Ms. Aditi Ushir[3], Ms. Ketaki Warke[4]

Department of Computer Engineering,
K. J. Somaiya Institute of Engineering and Information Technology,

Sion, Mumbai, University of Mumbai

Department of Computer Engineering, K. J. Somaiya Institute of Engineering and

Information Technology, Sion, Mumbai

Abstract

Due to the rapid increase of necessity in security and military applications, surveillance systems have become a necessary area of study. Asking human operators to keep watch for long hours is not only a cumbersome task but it also increases the chance of error. Thus, to assist human operators identify events which are important, Automatic Object Tracking is proposed. An object is tracked by, firstly, detecting the object using any of the various object detection methods in frames present in the input video. These methods make use of the spatial domain, temporal changes, presence etc. of the objects present. Every object is then tracked using any of the various methods. This can be used for monitoring traffic, animation, robot vision and video surveillance. In the proposed system, YOLO v2 is being used for Object Detection and Kalman Filter along with Non-Maximum Suppression will be used for Automatic Object Tracking.

Keywords: *Object Detection, Object Tracking, Keyframe Extraction*

I. INTRODUCTION

Automatic object tracking, when used for surveillance applications, improves the system's capacity to study information. The data obtained during surveillance results in gathering vast amounts of knowledge for a huge amount of time which makes data optimization a must. Keyframe extraction is the first step taken for data optimization. The required information, data or knowledge can be easily extracted from the keyframes. After keyframe extraction, the objects are detected and are later tracked automatically within the video [1].

Object detection techniques or methods generally fall into two categories. The first category is 'Machine Learning based approaches' and the second approach is 'Deep Learning based techniques'. Machine learning methods are preferred when the system's objective is to classify objects after detection whereas the Deep learning techniques are used for end-to-end object detection which means that the objects can be detected even without giving the system classifying parameters.The main aim of object detection is to locate where objects are present in a given image. The input provided to the system is a surveillance video of a particular system like Unmanned Aerial Vehicle, Drone-based Surveillance, surveillance for some premises etc. The keyframes from the particular video are then extracted by using the method of finding histogram difference of the frames. These keyframes make it easier for the user to extract information quickly and easily. The YOLOv2 algorithm is used for object detection in images. Non-maximum suppression is used for more precise detection. The last step includes tracking objects that have been detected by YOLOv2 using Kalman filter.

II. LITERATURE SURVEY

Li et al. had proposed a system which selected the video is divided into segments and the first frame of each of these segments is extracted and labelled as a keyframe. However, this method doesn't provide effective results as none of the other frames of the segments are looked at to determine if they contain valuable information. Zhao et al. explored the study of curve segmentation for extraction of key-frames. The video is divided into segments and every frame is represented by a coloured histogram. The difference between all the consecutive frames is calculated and plotted in a 2D plane. The end result of the curve which has been plotted is studied to find the sharp corners. The frames that correspond to the sharp corners in the graph are labelled as the keyframes for their respective segments [2].

Regunathan Radhakrishnan describes a key frame extraction technique which takes into account the user's intuition. It is assumed that more the motion in the video, more the number of keyframes extracted. The system divides the video into segments in such a way that the motion or activity involved in each of these segments is equal. The keyframes are those which are located halfway through each of the segments [3].

Mukherjee et al. proposed a system to extract keyframes from a video on the basis of the randomness of the frames. Unique features that are present in each and every frame are obtained in order to calculate the randomness between consecutive frames. The keyframes are those that correspond to high randomness [4].

As per the literature survey, the previous works of key-frame extraction shows us that a predefined number of keyframes are selected to represent each video. The method to be chosen should make sure that predefined number of keyframes are not extracted as the number of keyframes differ for different videos. Thus, the method of comparing absolute difference of the histogram of consecutive frames to a threshold will give an accurate output for keyframe extraction.

The next task is object detection. The easiest way to carry out object detection is to retrieve various regions of the user's interest from the image and then use these regions for classification based on CNN. The drawback of this method is the need of different spatial locations and aspect ratios within the image. In order to overcome these problems, algorithms like Faster R-CNN and YOLO are used.

Faster R-CNN:

Ross Girshick et al designed a fast and efficient algorithm called Faster R-CNN. In this method, the entire image is given as an input to the algorithm and a convolution feature map is generated. After this process determination of region proposals is done and all these proposals are combined into squares, with the help of Region of Interest(RoI) pooling layer all the proposals are converted to a predefined size. Followed by this the softmax layer is used for prediction of the region. The benefit associated with Faster R-CNN in comparison to R-CNN is that in Faster R-CNN the feature map is generated only one time and 2000 region proposals need not be fed to the neural network for each iteration [5].

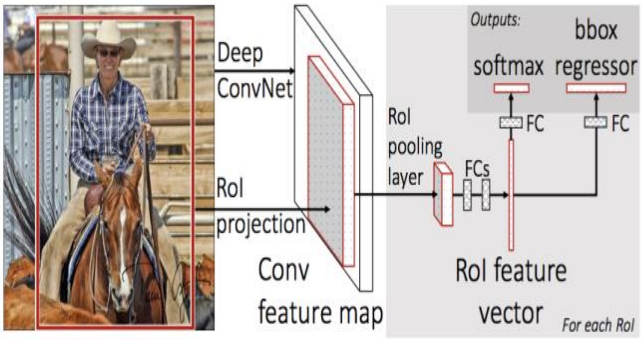

Figure 1. Faster R-CNN

III. PROPOSED SYSTEM

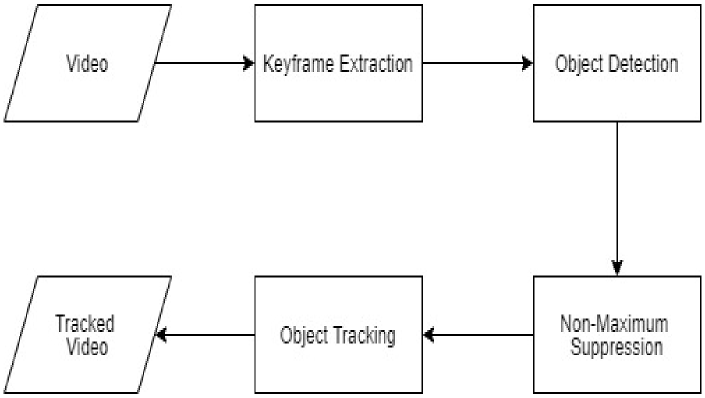

Figure 2. Proposed System

i) Video: An input video is provided.

ii) Keyframe Extraction: The proposed method for keyframe extraction is the "Absolute Difference of Histogram-based technique". This method extracts distinct key-frames while maintaining the order of these frames as they appear in the input video based on threshold obtained from mean and standard deviation of absolute differenceof histogram of consecutive frames. The first step is to find the number of frames that the input video consists of. This helps the system to estimate the number of comparisons it will have to make in order to find the keyframes. The second step is to convert each frame into its respective gray scale image. For each iteration, the histogram difference between the gray scale images of each of the consecutive frames is calculated. After this, the mean and standard deviation of the histogram differences is calculated. The threshold, which will be used to distinguish keyframes and regular frames, is calculated based on the values of the mean and standard deviation which is obtained. For the next step, for each iteration, the threshold value that has been obtained is compared to the histogram difference of the respective frames.. If this difference exceeds the value of the threshold then the second image is considered as a keyframe. After the algorithm is completely executed, a set of keyframes is obtained. These keyframes are those frames that summarize the input video precisely and effectively without the loss of any valuable information [6].

iii) Object Detection: The process of finding real-life objects like human beings, animals, cars, etc in visual media is called object detection. There are various object detection algorithms available in order to carry out object detection. All these algorithms work on different aspects of the characteristics extracted from an image. Object Detection is used in many systems like OCR, Object Tracking, etc. In our system we are going to use the YOLO algorithm. YOLO stands for You Look Only Once and as the name suggests the algorithm studies the properties and features of the input image only once and then uses this knowledge to carry out object detection.

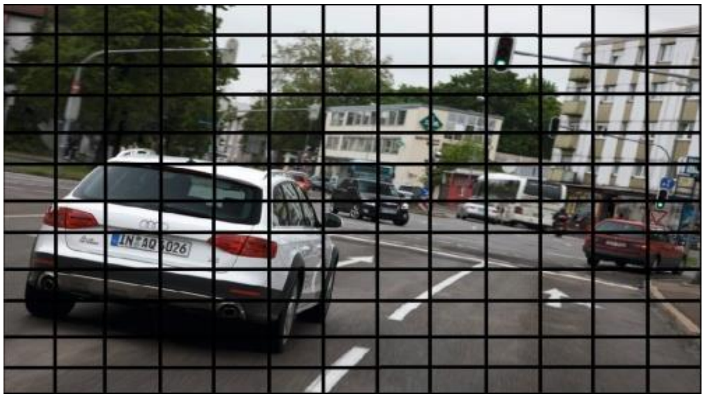

Figure 3. YOLO 13 X 13 grid

YOLO breaks every image into a 13 by 13 grid. Each grid is responsible for prediction of five rectangular boxes often known as bounding boxes. After these rectangular regions are identified and predicted, YOLO outputs the confidence and certainty associated with the prediction. These outputs do not give any information about the kind of object that is enclosed by the rectangular region [7].

iv)Non-maximum Suppression:

Non-Maximum Suppression is an important part of computer vision and it is the core of most proposed approaches when it comes to detecting an edge, corner or an object. It is needed because the ability of algorithms to locate the object it is interested in is not perfect. This results in detection of the same object several times around the real location of the object. [8].

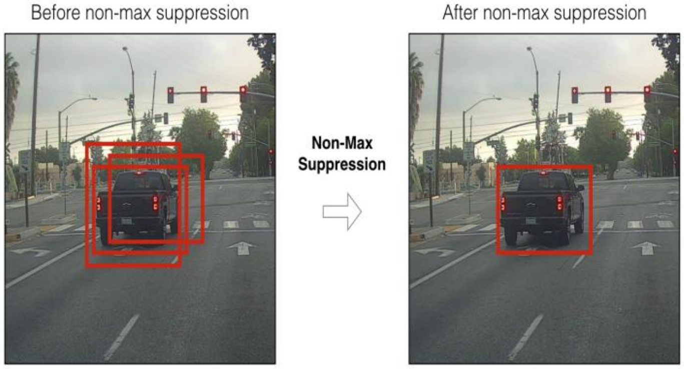

Figure 4. Non-Maximum Suppression (NMS)

v) Object Tracking: Object tracking is performed by using Kalman Filter after the use of NMS. Tracking is performed by feeding the centroid of the bounding box to the Kalman Filter which predicts the particular object's future trajectory based on it's current direct and speed. Kalman Filter performs 2 steps: Correction and Prediction. The previous state's motion model is vital to predict the object's current state. When the object's centroid is provided, an update is performed [9].

IV. IMPLEMENTATION

i) Keyframe Extraction:

Keyframe Extraction can be summarized as a dual stepped algorithm in which the first part is to calculate the threshold value which will be used to select the keyframes. Threshold is obtained with the help of the formula:

$$T= aM+ bS \quad (1.1)$$

where M is the mean, S is the standard deviation of the absolute difference respectively and a,b are constants. The latter part involves extraction of keyframes. Once the threshold is calculated, the keyframes are extracted by comparing the absolute difference of the histogram of every consecutive frame to the value of the threshold.

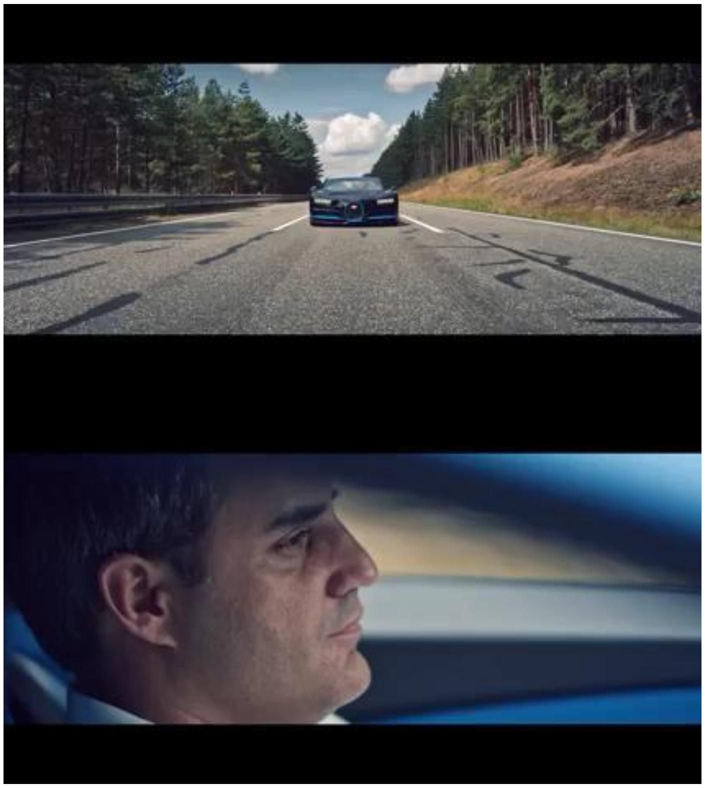

Figure 5,6. Keyframes Extracted Using Histogram Based Technique

ii) Object Detection:

It is not possible to detect objects in motion without tracking them. Hence, object detection will first be done on a dataset of images containing various vehicles. The YOLOv2 algorithm is used for object detection in these images. The first step to implement this deep learning technique is to create and train a model. Various images of vehicles are provided to the model. This dataset of images is divided as 60% of the images for training and 40% of the images for testing. The set of training images is given to the model in order to train the detector. More the images provided to the model for training, the better the accuracy while detecting images after training. Instead of providing a large number of different images for training, a simple augmentation function is used. This augmentation function randomly changes the images. These transformed images act as new training data for the detector. In this manner, the training accuracy is increased by simply augmenting the original set of training data rather than using new sets for training. After the model is trained, the images kept aside for testing are given to the model to check how accurately the model detects objects in unseen images. The set of testing images is not augmented in order to unbiasedly evaluate the detector's accuracy. The object is detected in an image with a confidence rate which suggests how confident the algorithm is that it has detected an object.

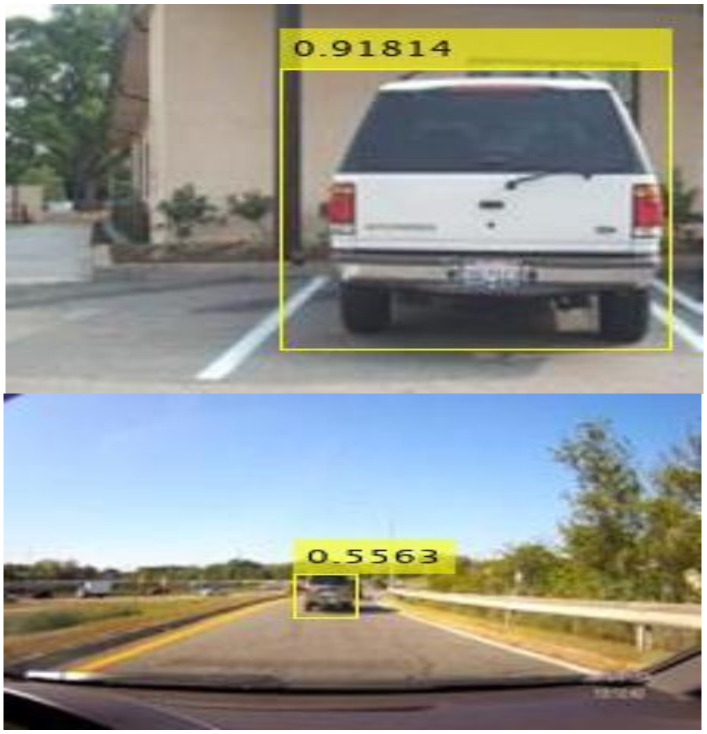

Figure 7,8. Object Detection In Images Using YOLOv2

V. PERFORMANCE METRICS

It is important to evaluate how well the detector performs while detecting objects. The average precision performance metric gives a number that combines the ability of the detector to accurately locate an object and the detector's ability to look for all the related data which is commonly known as precision and recall respectively. The graph below shows the precision at different levels of recall. The average precision value of the trained model is 0.83. Ideally, the average precision for a model should be 1. Using more images for training the model can improve the average precision however doing so would also increase the time taken for training the model.

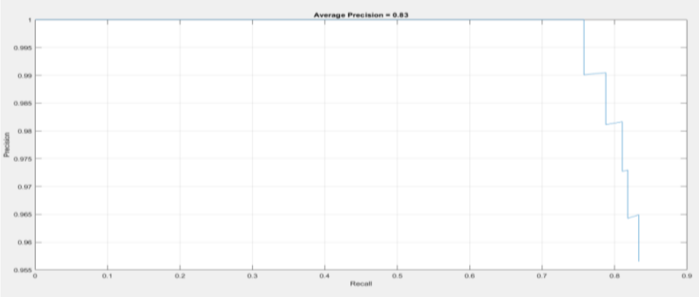

Figure 9. Precision vs Recall Graph Plotted To Calculate Average Precision

VI. CONCLUSION

The task of tracking an object employs several different techniques in a particular sequence in order to detect it precisely. On reviewing various methods for keyframe extraction, the histogram-based method was chosen as the one with the best outcome. Object detection was performed on a dataset of images in order to detect the objects. The YOLOv2 algorithm was used as it provides rapid and accurate detection of an object. The training speed of the YOLOv2 algorithm is unmatched and thus, was chosen for implementation. Object tracking is important as it provides several applications like surveillance systems, sports summary, better analysis etc. We will be employing Kalman Filter for the purpose of tracking an object detected by YOLOv2 algorithm.

REFERENCES

[1] Ancy A. Micheal, K. Vani, “Automatic object tracking in optimized UAV video”, The Journal of Supercomputing, 2019.

[2] Sheena C V and N. K. Narayanan, “Key-frame extraction by analysis of histograms of video frames using statistical methods”, 4th International Conference on Eco-friendly Computing and Communication Systems, 2015, p. 36-40.

[3] Ajay Divakaran and Regunathan Radhakrishnan and Kadir A Perker, “Motion Activity Based Extraction Of Key Frame From Video Shots”, IEEE 2002.

[4] Mukherjee et al., “Key-frame Estimation in Video using Randomness Measure of Feature Point Pattern”, IEEE transactions on circuits on systems for video technology, Vol.7, No.5, May 2007, p. 612-620.

[5] Rohith Gandhi, “R-CNN, Fast R-CNN, Faster R-CNN, YOLO Object Detection Algorithms”, https://towardsdatascience.com /r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e”, July 2018.

[6] Sanjoy Ghatak, “Key-frame Extraction Using Threshold Technique”, International Journal of Engineering Applied Sciences and Technology, 2016, Vol. 1, Issue 8, p. 51-56.

[7] Matthijs Hollemans,“Real-time object detection with YOLO”, “https://machinethink.net/blog/object-detection-with-yolo/”, May 2017.

[8] Non-max Suppression”, https://www.coursera.org/lecture/convolutional-neural-networks/non-max-suppression -dvrjH.

[9] “Kalman Filter”, “https://en.wikipedia.org/wiki/Kalman_filter”.