International Journal of Research in Engineering, Science and Management
Volume 4, Issue 12, December 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

Self-Directed Tracking of Objects using Deep Learning Technique

Shreyash S. Kalal[1*], N. Niveditha[2]

[1]Data Analyst, Department of Pre-Sales, Aurigo Software Technologies, Bangalore, India

[2]Software Engineer, Department of Quality Engineering, Aurigo Software Technologies, Bangalore, India

***Abstract*: Security has become a necessity. Asking humans to keep watch for long hours is a cumbersome task and increases the chance of error. Thus, to assist human operators in identifying important events, Automatic Object Tracking is proposed. Firstly, an object is tracked by detecting the object using any of the various object detection methods in frames present in the input video. These methods use the spatial domain, temporal changes, presence, etc., of the objects present. Everything is then tracked using any of the various techniques. 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 and Non-Maximum Suppression will be used for Automatic Object Tracking.**

***Keywords*: Keyframe extraction, Object detection, Object tracking.**

1. Introduction

Automatic object tracking, when used for surveillance applications, improves the system's capacity to study information. The data obtained during surveillance gathers vast amounts of knowledge for a considerable amount of time, making 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. In contrast, Deep Learning techniques are used for end-to-end object detection, meaning that the objects can be detected without giving the system classifying parameters. The main aim of object detection is to locate where things 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 specific video are then extracted by finding the 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 the Kalman filter.

2. Literature Survey

Li et al. proposed a system that selected the video divided into segments, and the first frame of each of these segments is extracted and labeled 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 a colored histogram represents every frame. The difference between all the consecutive frames is calculated and plotted in a 2D plane. The result of the curve, which has been planned, is studied to find the sharp corners. The frames that correspond to the sharp corners in the graph are labeled as the keyframes for their respective segments [2].

Regunathan Radhakrishnan describes an essential frame extraction technique that takes into account the user's intuition. It is assumed that the more the motion in the video, the more the number of keyframes extracted. The system divides the video into segments so 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 based on the randomness of the frames. Unique features present in every frame are obtained 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 show us that a predefined number of keyframes are selected to represent each video. The chosen method should ensure that the predefined number of keyframes is not extracted as the number of keyframes differs for different videos. Thus, comparing the 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

*Corresponding author: shreyashkalaled@gmail.com interest from the image and then use these regions for classification based on CNN. The drawback of this method is the need for different spatial locations and aspect ratios within the image. To overcome these problems, algorithms like Faster R-CNN and YOLO are used.

A. 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 input to the algorithm, and a convolution feature map is generated. After this process, determination of region and proposals is done. All these proposals are combined into squares; with the help of the Region of Interest(RoI) pooling layer, all the proposals are converted to a predefined size. Followed by this, the softmax layer is used for the prediction of the region. The benefit of Faster R-CNN compared to R-CNN is that in Faster R-CNN, the feature map is generated only once, and 2000 region proposals need not be fed to the neural network for each iteration [5].

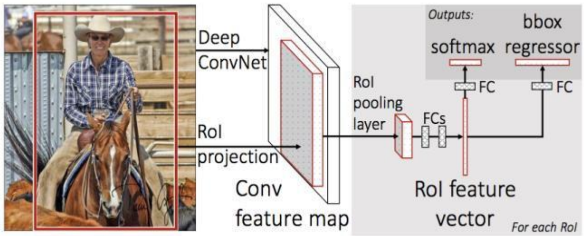

Fig. 1. Faster R-CNN

3. Proposed System

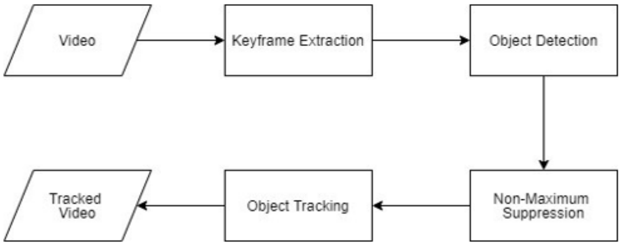

Fig. 2. Proposed system

A. Video

An input video is provided.

B. 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 the threshold obtained from the mean and standard deviation of absolute difference of histogram of consecutive frames. The first step is to find the number of frames that the input video consists of. This helps the system estimate the number of comparisons it will have to make to find the keyframes. The second step is to convert each frame into its respective grayscale image. For each iteration, the histogram difference between the grayscale images of each consecutive frame 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 mean and standard deviation 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 threshold value, then the second image is considered a keyframe. After the algorithm is entirely 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].

C. 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 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 approach, 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.

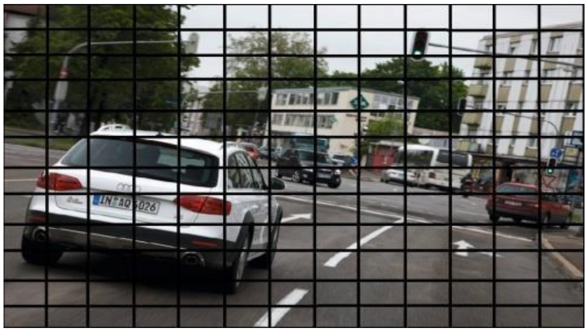

Fig. 3. YOLO 13 x 13 grid

YOLO breaks every image into a 13 by 13 grid. Each grid is responsible for the 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 enclosed by the rectangular region [7].

D. Non-maximum Suppression

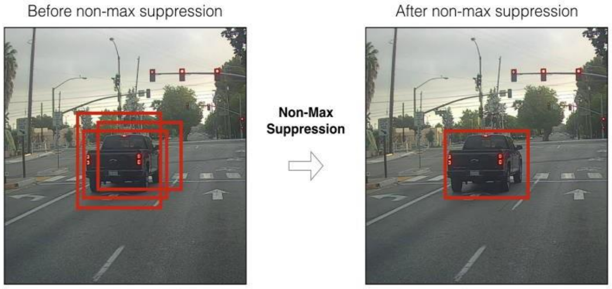

Fig. 4. Non-Maximum Suppression (NMS)

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

E. 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 its current direction and speed. Kalman Filter performs two steps: Correction and Prediction. The previous state's motion model is vital to predicting the object's current state. When the object's centroid is provided, an update is performed [9].

4. Implementation

A. Keyframe Extraction

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

$$T= aM+ bS \quad (1)$$

M is the mean, S is the standard deviation of the absolute difference, respectively, and a, b are constants. The latter part involves the 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.

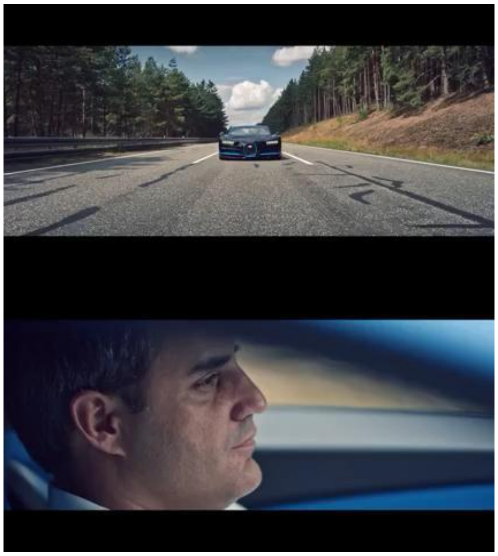

Fig. 5. Keyframes extracted using Histogram based technique

B. 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 implementing 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 into 60% of the images for training and 40% for testing. The set of training images is given to the model to train the detector. More the images provided to the model for training, the better the accuracy while detecting images after training. Instead of giving a large number of different ideas 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 training data set 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 to evaluate the detector's accuracy unbiasedly. The object is detected in an image with a confidence rate which suggests how confident the algorithm has detected an object.

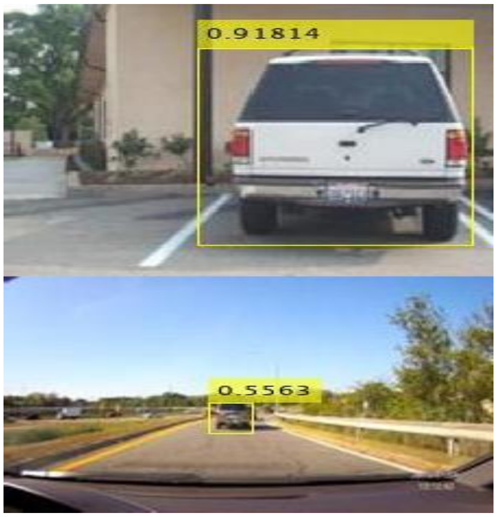

Fig. 6. Object detection in images

5. Performance Metrics

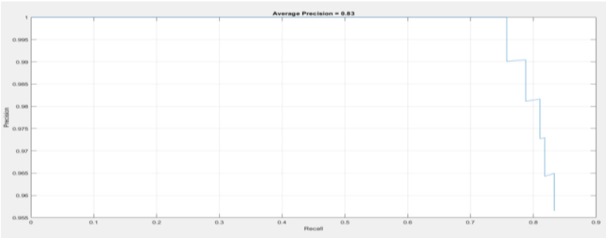

Fig. 7. Precision v/s Recall graph plotted to calculate average precision

It is essential 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 accuracy at different levels of memory. 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 accuracy however doing so would also increase the time for training the model.

6. Conclusion

The task of tracking an object employs several different techniques in a particular sequence 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 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 essential 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] D. P. Mukherjee, S. K. Das and S. Saha, "Key Frame Estimation in Video Using Randomness Measure of Feature Point Pattern," in *IEEE Transactions on Circuits and Systems for Video Technolog*y, vol. 17, no. 5, pp. 612-620, May 2007.

[2] 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.

[3] S. Ghatak, "Keyframe Extraction Using Threshold Technique," in *International Journal of Engineering Applied Sciences and Technology*, vol. 1, no. 8, pp. 51-56, 2016.

[4] A. A. Micheal and K. Vani, "Automatic object tracking in optimized UAV video," in *J Supercomput*., vol. 75, pp. 4986–4999, 2019.

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

[6] A. Divakaran, R. Radhakrishnan and K. A. Peker, "Motion activity-based extraction of key-frames from video shots," *Proceedings. International Conference on Image Processing*, 2002, pp. I-I.