https://www.ijsrtm.com
Vol.4 Issue 1 Mar 2024: 06-09
Published online 11 March 2024
E-ISSN: 2583-7141

International Journal of Scientific Research in Technology & Management

A Deep Analysis of Moving Object Detection & Tracking for Various Machine Learning Approaches

Amit Saxena
Dept. of Computer Science & Engineering
Rabindranath Tagore University
Bhopal, Madhya Pradesh, India
amit.saxena78@gmail.com

Sitesh Kumar Sinha
Dept. of Computer Science & Engineering
Rabindranath Tagore University
Bhopal, Madhya Pradesh, India
siteshkumarsinha@gmail.com

Sanjeev Kumar Gupta
Dept. of Computer Science & Engineering
Rabindranath Tagore University
Bhopal, Madhya Pradesh, India
sanjeevgupta73@yahoo.com

***Abstract*— Object detection is a significant process for performing computer vision related task. It plays an important role in the field of visual object tracking. If it is talking about the real world then it is a challenging task to detect object with high precision rate because of the mazy in appearance. There are various deep learning approaches through which object can be detected precisely but detecting the moving object and tracking it in every frame is bit more challenging. Moving object detection and tracking have wide variety of application such as border surviellence, road activity detection, forest monitoring and many more. There are various deep learning networks such as CNN, R-CNN, SSD, YOLO and many more, these networks are very much capable to detect the object with large traning model. Object tracking is next process after object detection because it is required to target the intial point with target object that has been detected and need to track as per the motion. The intention of this paper is to deep analyze the machine learning approaches which are used to detect the moving object with tracking.**

Keywords*— Moving Object Detection, Object Tracking, CNN, R-CNN, SSD, YOLO, Pattern Recognition.

I. INTRODUCTION

In the field of object detection; it is required to detect the region of interest (ROI) and the background area. Several methods for foreground identification which is the target part of the image were put forth by several researchers over the past ten years. However, the issues have drastic shifts and target drift during tracking, which is not solved by several methods. Accurately estimating the object position is the fundamentally difficult especially in moving object identification and tracking. Identifying moving objects is a crucial stage in the field of computer vision technologies. It is utilized in a variety of applications, including military usage, medical imaging, and video analysis. Typically, a frame comprises information about the foregrounds and backgrounds but it is not the worth technique through accurate information can be obtained [1]. The ROI's feature points in this foreground item represent it, while the remaining features are regarded as background. The two main components of a surveillance system are motion estimation and the detection of moving objects. The backdrop pixel information affects the object detection, which is the initial phase. Video data must be compressed because it is redundant and out of place in both space and time. The movie may be compressed by minimizing the spatial and temporal information. Many academics have numerous approaches that concentrate on picking up things in a video clip. Many of them combine and intersect many procedures, and many of them employ numerous strategies. The technique known as background subtraction is used to remove the interesting moving item from the frames of a video. The majority of non-stationary backdrop and light fluctuations have an impact on the background subtraction. The optical flow method can really eliminate this flaw; however in congested environments it causes false alarms for tracking systems. The object trackers are impacted by background data in the majority of background subtraction scenarios, although this results in misclassification. Additionally, choosing a reliable classifier is a difficult for improving the accuracy [2].

II. RELATED WORKS

There are different approaches proposed by various researchers that having certain flaws. The motive is to deep analyze the researches related to the moving object detection along with object tracking. In the discipline of computer vision, object detection is a fundamental study area. For more difficult computer vision tasks like target tracking, pattern recognition, and semantic comprehension, it serves as a crucial precursor. It seeks to properly identify the category, locate the target of interest inside the image, and provide the bounding box of each target. Small data size, low portability, lack of pertinence, high temporal complexity, and lack of robustness only in certain basic situations are the primary drawbacks of this technology.

A. *Convolutional Neural Network (CNN)*

Shraddha Mane et al. [3] presented the research related to robust object detection using Convolutional Neural Network (CNN). The advantages of CNN for detecting object is that it has high accuracy and less false alarm rate. But the biggest disadvange is that it requires lots of training data to produce the precise model. It has various burdens such as gradient problem, overfitting, balancing and handling large datasets. The result of this model is based on sensitivity, specificity and accuracy. These parameters are based on the confusion matrix for predicting the object on the basis of features and correct and incorrect outcomes.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100\ \% \quad (1)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} * 100\ \% \quad (2)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} * 100\ \% \quad (3)$$

Here the primary model for detecting the object is Tensorflow and for tracking the detected or recognized object is CNN.

B. *Region-Convolutional Neural Network (R-CNN)*

Girshick, R et al. [4] presented an object detection model using R-CNN and in this model; it first uses the selective search feature to extract the region that enhances the object detection approach. The retrieved object characteristics are then sent into the SVM classifier for classification after being uniformly scaled to a fixed-length feature vector. Finally, a linear regression model is trained to carry out the bounding box regression process. The R-CNN does significantly enhance accuracy when compared to the conventional detection approach, but it requires a lot of calculations and does it inefficiently. Second, converting the region suggestion to a fixed-length feature vector directly could distort the objects.

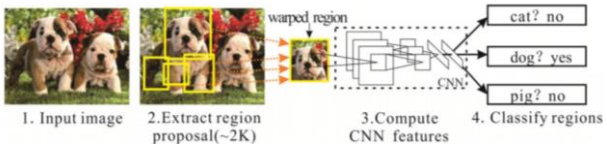

Fig. 1. R-CNN Architecture [4]

C. *Faster-Region Convolutional Neural Network (F-RCNN)*

Girshick, R et al. [4] presented an object detection model using R-CNN and in this model; it first uses the selective search feature to extract the region that enhances the object detection approach. The retrieved object characteristics are then sent to the model for detection and tracking. Proposed approach replaces the prior Selective Search technique for region proposal generation with region proposal networks. The model is composed of two modules: the Fast R-CNN detection method and a fully convolutional neural network used to produce all region proposals. These two modules share a set of convolutional layers. The input picture is sent via the CNN network to the shared convolutional layer at the very end. In order to create a higher-dimensional feature map, the picture is transmitted forward to the specified convolutional layer on the one hand, and the feature map for the input of the RPN network on the other.

D. *TrackNet (Track Network)*

Chenge Li et al. [5] presented an object detection and tracking approach that is related to the TrackNet. It has potential to detect and track the multiple objects at same time. It uses the spatial temporal features as VGG network does. TrackNet is newly introduced network that acts as VGG-19 but the problem is that both the networks have bulky in nature because of heavy traning and higher weight models that is why it takes more time to detect the object and also consumes higher memory as compare to the various networks. In tracking, systems need to detect the object at each and every frame in real time and if it is a time consuming process then the frame rate will get degraded and system will not effectively work.

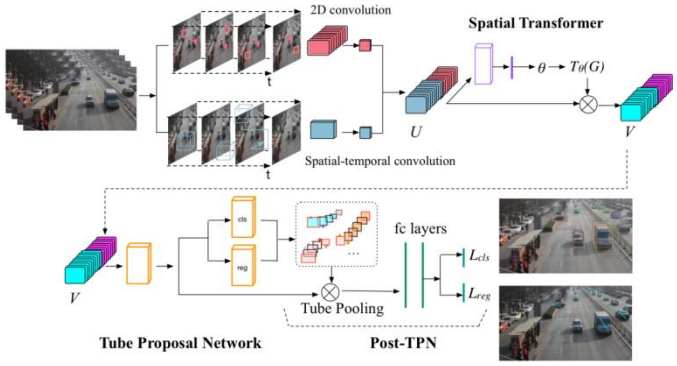

Fig. 2. TrackNet Structure [5]

E. *Mask R-CNN (Mask Region-Based Convolutional Neural Network)*

Faster R-CNN was expanded by the developers of Mask R-CNN at Facebook AI Research (FAIR) to include instance segmentation in addition to the class and bounding box operations. By doing instance segmentation, which combines object detection with semantic segmentation, all objects in an image are first detected and then each instance is segmented while being distinguished from the other instances. Mask R-CNN generates a binary mask for each RoI in parallel with the class and bounding box in the second stage, whereas the first stage (region proposal) is identical to that of its predecessor. Without taking into account the categories, this binary mask indicates whether the pixel is a part of any item. The model is much simpler to train because the class for the pixels would be determined just by the bounding box that they are located in. The second stage also differs in that RoIAlign is used instead of the RoI pooling layer (RoIPool) that was first introduced in Fast R-CNN. When using RoIPool to perform instance segmentation, the feature map is mismatched with respect to the source picture and has several pixel-wise errors. The reason for this is because RoIPool quantizes the areas of interest, which involves rounding the floating-point values

in the resultant feature map to decimal values. On the other hand, the enhanced RoIAlign accurately aligns the extracted features with the input by completely eliminating any quantization and instead computing the precise values of the input features via bilinear interpolation [6].

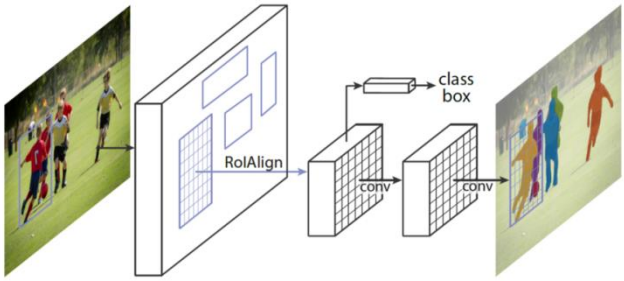

Fig. 3. Mask R-CNN framework [6]

F. YOLO (You Only Look Once)

Joseph Redmon et al. [7] developed YOLO, which is now the most well-liked object detection technique, and for good reason. While maintaining decent accuracy, it can analyse real-time movies with little latency. Furthermore, as the name implies, all objects in a picture may be detected with just one forward propagation. The same inventor of YOLO, Joseph Redmon, also developed Darknet, an open source neural network framework implemented in C and CUDA. In comparison to the previous generation, YOLOv3 is bigger, more accurate on tiny things, but somewhat less accurate on larger items. In contrast to the previous Darknet-19 (19-layer CNN) for YOLOv2, Darknet-53 (53-layer CNN with residual connections) is utilised in YOLOv3. YOLOv3 predicts bounding boxes at 3 distinct sizes and on various layers of the network, in contrast to earlier YOLO versions that just output the bounding box, confidence, and class for the box. Non-max suppression (NMS), a straightforward technique that eliminates bounding boxes that overlap with one another more than a predetermined intersection-over-union (IoU) threshold, is used to determine the final item detections on the picture. When two bounding boxes overlap, the one with the highest YOLO-assigned confidence wins, and the other two are thrown out.

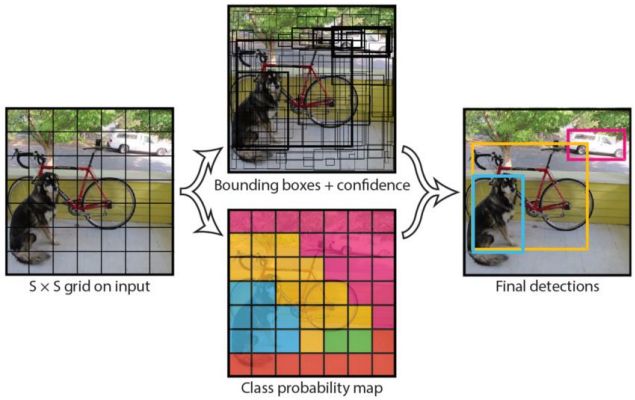

Fig. 4. YOLO Regression Problem [7]

The box values are related to reference anchors, much like in Faster R-CNN. Nevertheless, it employs k-means clustering on the training dataset to choose the best anchors for the task, as opposed to using the same hand-selected anchors for every task. With YOLOv3, 9 anchors are the standard. It could also come as a surprise that several independent logistic classifiers trained with binary cross-entropy loss, rather than softmax, are employed for class prediction [7].

G. SSD (Single Shot MultiBox Detector)

Wei Liu et al. [8] presented SSD which is a multibox detector for a single shot detention. A few months after YOLO, Single Shot MultiBox Detector (SSD) was released as a respectable substitute. Similar to YOLO, object detection occurs during a single network forward propagation. The input picture is passed through many convolutional layers in this end-to-end CNN model, creating candidate bounding boxes at various sizes along the way. SSD uses the identified items as the baseline for training, and any other bounding boxes that do not coincide with the positives are considered negative instances. It turns out that by building the dataset in this manner, it is extremely unbalanced. Because of this, SSD does hard negative mining immediately following NMS. In order to limit the ratio of positive to negative instances to no more than one to three, hard negative mining selects only the negative examples with the largest confidence loss. This results in a more stable training phase and speedier optimization.

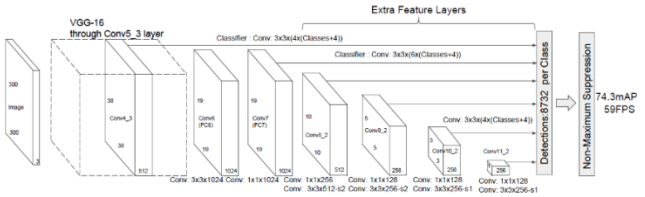

Fig. 5. SSD Architecture [8]

H. RetinaNet (Retina Network)

Tsung-Yi Lin et al. [9] presented RetinaNet where researchers from FAIR suggested RetinaNet back in 2017. Moreover, it uses a one-stage architecture similar to YOLO and SSD, which compromises accuracy for speed rather than using a two-stage framework like the R-CNN variants. The RetinaNet creates a rich, multi-scale convolutional feature pyramid using a ResNet + FPN backbone. As usual, two subnetworks are connected to the top, one for categorising anchor boxes and the other for creating offsets from the anchor boxes to the object boxes that represent the ground truth.

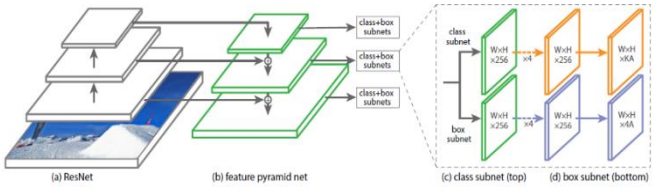

Fig. 6. RetinaNet Architecture [9]

As previously indicated, the cross entropy loss is outweighed by the imbalance of classes during training of dense detectors. By reducing the amount of simple negatives and focusing training on a small number of challenging cases, the novel focused loss enhances accuracy. This is accomplished by altering the loss function such that it values difficult samples less highly than easy ones.

I. ROLO (Recurrent YOLO)

Guanghan Ning et al. [10] presented ROLO a single object tracking technique that combines recurrent neural networks and object identification. YOLO and LSTM are combined to form ROLO. In addition to location inference priors, the object detection module also employs YOLO to gather visual information. The LSTM accepts an input feature vector of length 4096 at each time step (frame) and outputs the tracked object's position.

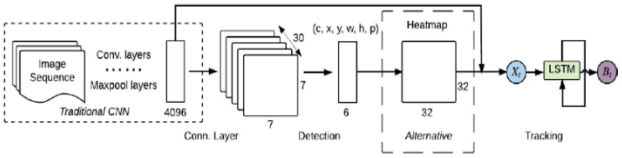

Fig. 7. ROLO Architecture [10]

J. SiamMask (Siamese Mask)

Qiang Wang et al. [11] presented SiamMask as an object tracking model. SiamMask is a fantastic option for monitoring a single item. It is built on Google's Facenet, which popularised the endearing siamese neural network. It provides class-independent object segmentation masks in addition to rotating bounding boxes at a rate of 55 frames per second. SiamMask must be started with a single bounding box in order for it to monitor the intended item in order to do this. A substantially slower object detector would result from changing the model to accommodate multiple object tracking (MOT), which is not possible with SiamMask.

III. CONCLUSION

This paper is intented to analyse various approaches that can detect the moving objects and track as per the positions for the same. It is bit challengable for most of the models to track the object in every frame due to blur in motion; if an object is moving very fast. There are many approaches that can detect the object very precisely such as CNN, RNN, YOLO and many more but these network models get distracted due to lack of feature prediction and then it get more complicated to evaluate the target attribute. SiamMask is bit popular in this challenge because it uses Facenet network and Facenet is very much proficient to detect and track object when it is in motion also. YOLO is good for object detection and it has various versions. In comparison to the previous generation, YOLOv3 is bigger, more accurate on tiny things, but somewhat less accurate on larger items.

REFERENCES

[1]. Enrique J. Fernandez-Sanchez *, Javier Diaz and Eduardo Ros, “Background Subtraction Based on Color and Depth Using Active Sensors”, Sensors 2013, 13, 8895-8915; doi:10.3390/s130708895.

[2]. Junda Zhu, Yuanwei Lao, and Yuan F. Zheng, “Object tracking in structured environment for video surveillance applications”, IEEE transactions on circuits and systems for video technology, vol.20, February 2010.

[3]. S. Mane and S. Mangale, "Moving Object Detection and Tracking Using Convolutional Neural Networks," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 1809-1813, doi: 10.1109/ICCONS.2018.8662921.

[4]. Girshick, R., Donahue, J., Darrel, T.,Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: Computer Vision and Pattern Recognition. Columbus.2014, pp. 580-587.

[5]. Li, Chenge & Dobler, Gregory & Feng, Xin & Wang, Yao. (2019). TrackNet: Simultaneous Object Detection and Tracking and Its Application in Traffic Video Analysis.

[6]. He, Kaiming and Gkioxari, Georgia and Dollár, Piotr and Girshick, Ross, Mask R-CNN, arXiv 2017.

[7]. Redmon, Joseph and Divvala, Santosh and Girshick, Ross and Farhadi, Ali, You Only Look Once: Unified, Real-Time Object Detection, arXiv 2015.

[8]. Liu, W. et al. (2016). SSD: Single Shot MultiBox Detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science(), vol 9905. Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_2

[9]. T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2999-3007, doi: 10.1109/ICCV.2017.324.

[10]. G. Ning et al., "Spatially supervised recurrent convolutional neural networks for visual object tracking," 2017 IEEE International Symposium on Circuits and Systems (ISCAS), Baltimore, MD, USA, 2017, pp. 1-4, doi: 10.1109/ISCAS.2017.8050867.

[11]. Wang, Qiang & Zhang, Li & Bertinetto, Luca & Hu, Weiming & Torr, Philip. (2019). Fast Online Object Tracking and Segmentation: A Unifying Approach. 1328-1338. 10.1109/CVPR.2019.00142.