Pioneer: Journal of Advanced Research and Scientific Progress (JARSP)
Volume: 02 Issue: 06 | 2023 ISSN: 2751-7551
http://innosci.org

A Brief Overview of Object Tracking Techniques in Images and Videos

Nitish
P.G. Student, Department of ECE, Sat Kabir Institute of Technology and Management, Haryana, India

Sumit Dalal, Kirandeep
Assistant Professor, Department of ECE, Sat Kabir Institute of Technology and Management, Haryana, India

Abstract: An essential and difficult task in the field of computer vision is the attempt to identify, track, and locate objects over a series of images known as video. The process of identifying specific items in pictures or movies is known as object recognition. In lieu of having human operators watch over computers, it is helpful to comprehend and describe object behaviour. It seeks to find moving things in video files or security camera footage. Finding an object or objects using a single camera, many cameras, or a particular video file is known as object tracking. With identification of objects, object depiction, and object tracking, we provide an overview of the fundamentals of object detection and tracking.

Keywords: Object Tracking, Object Detection.

I. INTRODUCTION

The process of detecting an object's position in photos and videos taken by security cameras is known as object identification and localization. Several computer vision (CV) methods can be used to categorise and identify the objects in actual time [1]. Additionally, the objects are defined by using bounding boxes that are rectangle in shape. On the basis of ML (Machine Learning) and DL (Deep Learning), there are numerous applications of object detection in business and academic research, including recognising faces [2], recognising text [3], recognising pedestrians [4], logo recognition [5], recognising objects in videos [6], vehicle identification [7], identifying illnesses [8], imaging for medicine [9], and a lot more.

Traditional techniques based on ML (Machine Learning) have been used mostly for object detection; the object area is computed first using a sliding window, then various features are mined, and finally, traditional classifiers such as SVM (Support Vector Machine) are used to classify the objects. Although the results are satisfactory, these methods are incapable of accurately detecting and classifying objects ignoring the underlying deep features. A researcher used the Haar feature descriptor to extract the linear, center, diagonal, and edge features before classifying the objects using a Support Vector Machine. Moreover, employing a hand-crafted feature descriptor requires human effort. As a result, researchers are concentrating their efforts on deep learning algorithms like CNN, R-CNN, and YOLO, which have greatly improved object detection performance.

The goal of current deep learning algorithms for object detection is to streamline the network and accelerate the detection procedure. These outcomes largely depend on the precision of the derived hypotheses, like, for instance, when the researcher uses a quicker R-CNN approach to identify and count automobiles. Despite the fact that this method can speed up the detecting process, it is less accurate than other more established techniques. The biggest drawback of these techniques is their inability to find far-off cars.

II. RELATED WORK

As an outcome of advancements in the field of CV and their numerous monitoring applications, a large array of methodologies has been established in the recent past, attracting the necessary interest of researchers. As it may be used in so numerous uses, such as human detection, face detection, car identification, hammer identification, weapon detection, blade detection, and many more, recognising objects in images becomes more and more crucial as a result of advancements in computer vision. The traffic system is more and more reliant on automatic vehicle identification devices as a result of technological advancements and a growth in the number of vehicles on the road globally. As a result, the vehicle detection and recognition system must be effective in a time-based environment.

Formerly many handcrafted elements had to be deleted in order to detect vehicles, which calls for manual action. The three most frequently used feature descriptors were HAAR [10], hog [11], and LBP [12]. The classification framework's effectiveness at detecting automobiles was assessed, and a significant number of automobiles were found. Furthermore, the SVM classifier is frequently utilised with remarkable success in vehicle detection when combined with the HOG feature. Additionally, statistical techniques utilising vertical and horizontal edge attributes were started for the recognition of cars and vehicle tracking at night by positioning the tail lights, together with the aforementioned characteristics and learners with extensive applications in vehicle detection tasks.

Object Tracking vs. Object detection

Once the starting point of the target object is known, object tracking pertains to the capability to anticipate or forecast the current location of the target object in each subsequent frame of a video.

Contrarily, object detection is the act of identifying a target object inside a frame of a video or an image. Only when the target image is discernible on the input will object detection function. It cannot detect the target object if it is obscured by any disturbance. Regardless of the occlusions, object tracking is trained to follow the object's route.

Single Object Tracking (SOT) vs. Multiple Object Tracking (MOT)

Rather than tracking several objects, (SOT) seeks to track one object from a single class. It is also known as Vision Object Tracking on occasion. The target object's bounding box in SOT is established in the opening frame. The technique's objective is to find the identical object in the remaining frames. SOT falls under the class of detection-free tracking since the tracker requires the user to individually supply the first bounding box. As a result, Single Object Trackers ought to be capable to track any object they are given, even if there isn't a trained categorization model for it [13]. Multiple Object Tracking (MOT) is a technique where the tracking algorithm tracks each and every object of interest in the video, first determining the number of objects within every frame, then tracking each object's identity from a single image to the next until the object leaves the frame.

III. ISSUES IN OBJECT TRACKING

Any tracking method must use an object detection methodology to identify the target item in either every frame of the video or only in the frame where it first appears. However, certain object detection systems use the frame sequence's global knowledge resister to reduce the number of erroneous detections. Focusing on object tracking algorithms can provide a number of difficulties. On a straight road or in an uncomplicated setting, it is straightforward to track an object. The target object will experience distortion, occlusion, background noise, etc. in a real-world setting.

Occlusion

The tracking algorithm declines track of the item due to an interference phenomena where the background or foreground has an impact on the object. In other words, as more things draw close, the algorithm becomes confused. This results in the problem of the initially detected object being

tracked again (erroneously) as a new object. Occlusion sensitivity can be used to stop it. The user can determine which particular characteristic of the item is perplexing the network thanks to occlusion sensitivity. Identical photos can be utilised to correct biases and aid the network in extracting features that distinguish the objects after being recognised.

Background clutter

The background of the photos input into the algorithm causes a lot of problems in any ML or DL activity. Object tracking models work in a similar way.

Theoretically, it becomes more challenging to identify features, detect, or even track the object of interest when the backdrop is densely packed. In addition to slowing the network's ability to learn and optimise, a background that is densely filled contributes duplicate data or noise that makes the network less responsive to essential features. A carefully selected dataset with a sparse background can be used to avoid background clutter.

Fluctuating spatial scales

One of the problems with object tracking is that the target objects can have a wide range of sizes and shapes; this kind of information may perplex the learning algorithm and result in generalization mistake.

IV. Deep Learning-based approaches to Object Tracking

A number of the techniques used old or conventional machine learning techniques like Support Vector Machine or k-Nearest Neighbour. These methods are effective at identifying the intended object, but they call for the extraction of significant and discriminating features by experts. Deep learning techniques, on the other hand, independently retrieve these crucial features and interpretations.

MDNet

Multi-Domain Net is a sort of object tracking technique that uses extensive training data. Its goal is to teach you about numerous possibilities and spatial connections. It takes several annotated movies from various domains since MDNet is trained to acquire the shared depiction of targets from these films. Pretraining and visual tracking are components of MDNet online: The network must learn multi-domain description during pretraining. The algorithm is trained on numerous annotated films to learn representational and spatial properties in order to accomplish this. After pre-training, the network only contains shared layers made up of learnt representations after the domain-specific layers have been eliminated. A binary categorization layer is inserted throughout the inference process and is trained or improved live.

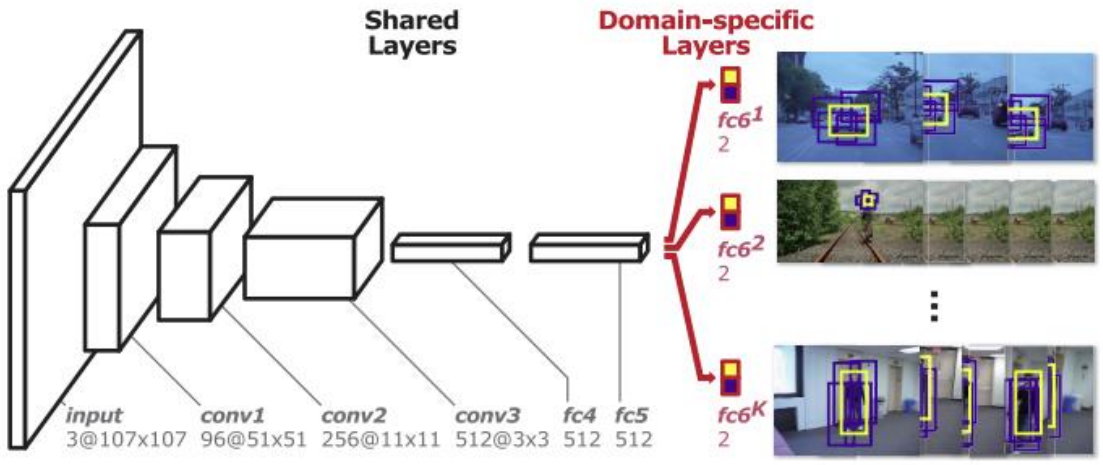

Fig.1.: An instance of MD Network[14].

GOTURN

Deep Regression Networks are systems with offline training. This technique can be employed to monitor objects that do not present in the training set since it learns a general link between object motion and appearance. Since they cannot benefit from a huge number of movies to enhance their efficiency, online tracker algorithms are slow and do not work effectively in real-time. Alternatively, offline tracker systems can be trained to handle rotations, shifts in perspective, changes in lighting, and other challenging situations. Generic Object Tracking Using Regression Networks, or GOTURN, tracks objects using a regression-based method. They basically perform a single feed-forward pass over the network before regressing immediately to find the target items. A search area from the present frame and a target from the previous frame are the network's two inputs. The target object in the present image is then located by the network by comparing these images.

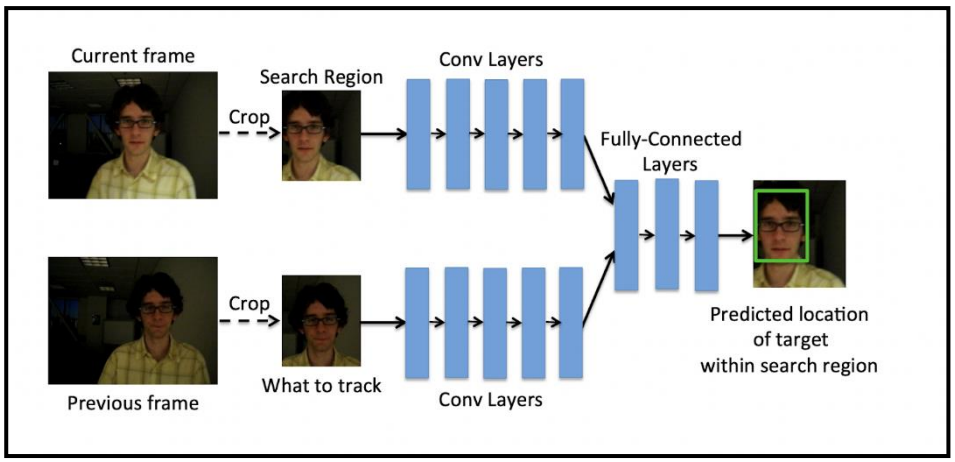

Fig.2.: An instance of GOTURN [15]

ROLO—Recurrent YOLO

ROLO combines YOLO and recurrent neural networks. In general, LSTM and CNN are chosen together. In ROLO, two different neural network architectures are combined: an LSTM network is used to determine the trajectory of the target object, and a CNN is utilised to extract spatial information. Spatial data is taken out and transmitted to the LSTM at each time step, which then provides the location of the object.

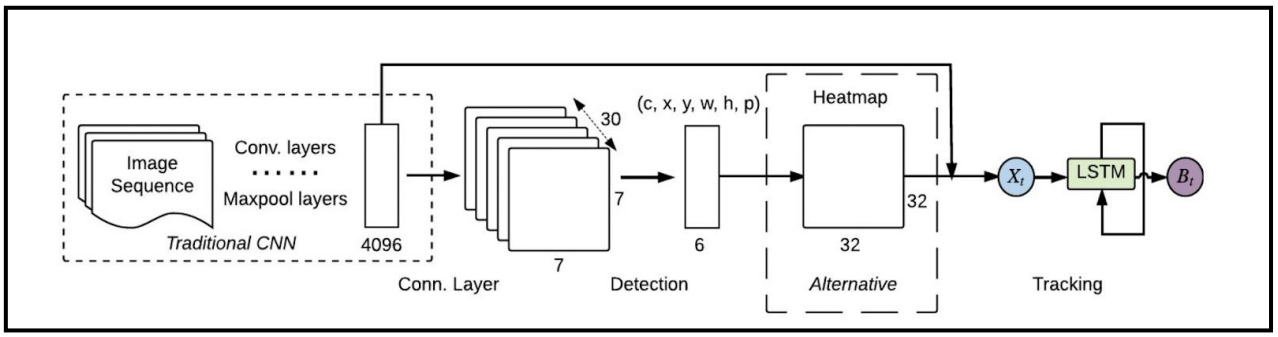

Fig.3: An instance of ROLO Model [16]

ROLO combines YOLO and recurrent neural networks. Overall, LSTM and CNN are chosen together. In ROLO, two different neural network architectures are combined: an LSTM network is used to determine the trajectory of the target object, and a CNN is utilised for gathering spatial information. Spatial data is taken out and transmitted to the LSTM at each time step, which then provides the object's location.

DeepSORT

One of the most well-liked object tracking algorithms is DeepSORT. It is an addition to the online-based monitoring algorithm known as Simple Online Real-time Tracker, or SORT. In order to estimate the location of an object given its past location, a method called SORT employs the Kalman filter. The occlusions are extremely well handled by the Kalman filter. Three parts make up SORT: detection, estimation and association. After covering the fundamentals of SORT, we can use deep learning strategies to improve the SORT algorithm. Because deep neural networks can now characterise the properties of the target image, SORT is able to predict the object's location with a significantly better degree of precision. In essence, a task-specific dataset is used to train the CNN classifier until it obtains high precision. Once it is accomplished, the classifier is removed, leaving us with simply the dataset's retrieved features. The SORT technique is then used to track objects using this extracted characteristic.

SiamMask

The goal of SiamMask is to enhance the fully-convolutional Siamese network's offline training process. Siamese networks use two inputs to create a dense spatial feature depiction: an image that has been cropped and a bigger search image. The Siamese network produces just one result. It assesses how similar two input photographs are to one other and evaluates whether or not the same things are present in both images. By supplementing the loss of the objects with a binary segmentation job, this system is particularly effective for object tracking.

JDE (Joint Detection and Embedding)

JDE is a single-shot detector created to address the issue of multi-task learning. In a shared model, JDE learns target detection and appearance embedding. In order to achieve feature representation at each layer, JDE leverages Darknet-53 as the foundation. Then, these feature representations are combined using remaining links and up-sampling. The fused feature representation is then covered with the prediction heads, creating a dense prediction map. JDE produces bounding box classes and appearance embedding from the prediction head for object tracking. Employing an affinity matrix, these aspect embeddings are contrasted with embeddings of previously discovered items. In order to smooth the trajectory of the target item and to estimate its location, the Hungarian algorithm and the Kalman filter are utilised.

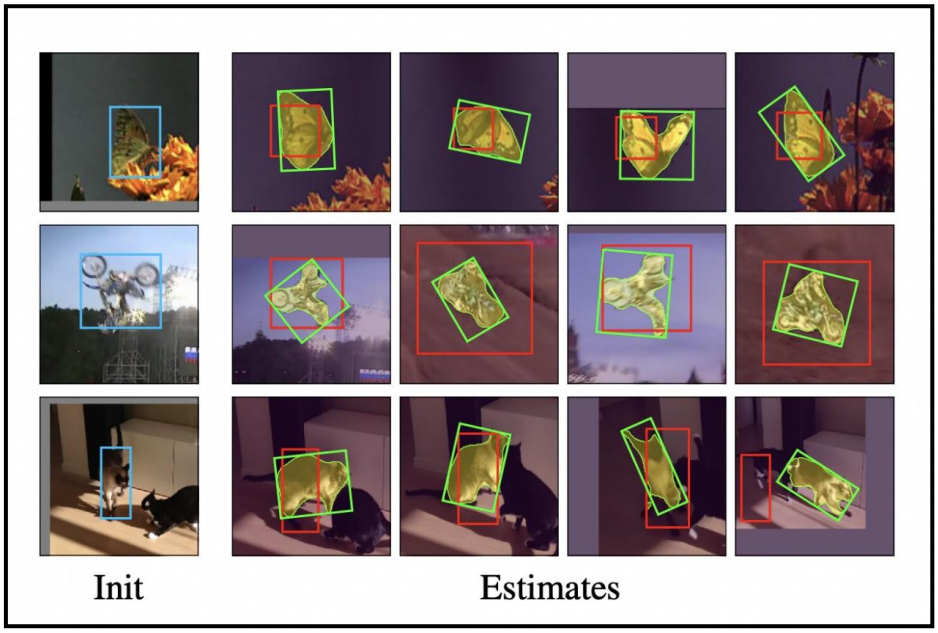

Fig.4: SiamMask[13]

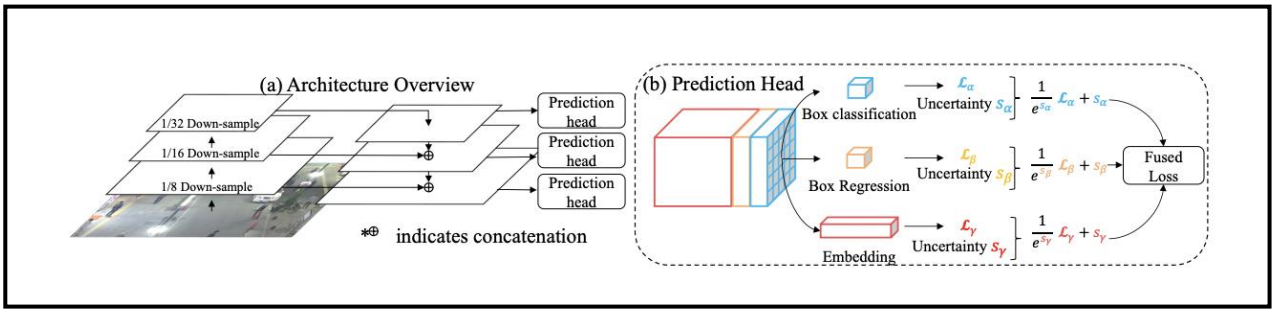

Fig.5: JDE Architecture[13]

V. COMMON OBJECT TRACKING TECHNIQUES

Point Detectors: Point detectors are employed to locate intriguing points in photos that have a certain local texture that is communicative. An interest point's resistance to variations in lighting and camera angle is a desired trait. The Moravec's detector, the Harris detector, the KLT detector, and the SIFT detector are all frequently used interest point detectors in literature.

Background Subtraction: Developing an illustration of the scene known as the backdrop model and then looking for departures from the model for each incoming frame can be used to detect objects. Background removal can be done in a number of ways, including frame differencing, region-based or geographical information, Hidden Markov models (HMM), and Eigen space breakdown.

Object Detection and Segmentation: Images are searched extensively to find and locate objects of interest. Convolutional neural networks (CNN)and some types of feature classifiers with sliding window techniques can both be used to recognise objects, albeit the latter is less frequent because of how long it takes.

It is possible to combine or perform separately the activities of object detection and creating a connection between object instances across frames. In the first scenario, an object detection method is used to determine the potential object region in each frame, and the tracker then correlates the items across frames. The object region and connection are jointly calculated in the latter scenario by continuously modifying object position and region data gleaned from prior frames.

Point Tracking: Moving objects are symbolised by their feature points during tracking in an image layout. Point tracking is a challenging issue, especially when there are object occlusions and incorrect object detections. By thresholding at the time of identification of these locations, identification can be accomplished rather easily.

Silhouette Based Tracking Approach: Simple geometric shapes are unable to precisely represent certain items due to their complicated shapes, such as hands, fingers, and shoulders. The items' shapes can be accurately described using silhouette-based approaches. Using an object model created from earlier frames, the goal of a silhouette-based object tracking is to locate the object region in every frame. capable of handling a range of object shapes, as well as object splitting and merging due to occlusion.

Kernel Based Tracking: In order to accomplish kernel tracking, a moving object that is represented by an embryonic object region is often computed from one frame to the next. Typically, the motion of the object takes the form of a parametric motion, like translation, conformal, affine, etc. These techniques differ in regards to the number of tracked objects, the presence description that is employed, and the technique for approximating object motion.

VI. CONCLUSION

The task of object detection and tracking is crucial in the field of computer vision. There are two main processes involved in object detection and tracking: object identification and object tracking. The backdrop subtraction method is used to locate objects in video images taken from a single camera with a stable background, fixing the camera.

REFERENCES

1. Amit, Kirti Bhatia, Rohini Sharma, Shalini Bhadola, International Journal for Scientific Research & Development|, An Overview on Object Tracking in Motion Pictures, Vol. 7, Issue 03, May, 2019, pp. 106-111.
2. Deepak Kumar Shrivastava, Kirti Bhatia, Shivkant, Rohini Sharma, Development and analysis of mean shift based Video object tracking tool, International journal of Innovative Research in computer and communication engineering, Vol-08, Issue-07,july 2020.
3. Zhang, Z.; Zhang, C.; Shen, W.; Yao, C.; Liu, W.; Bai, X. Multi-oriented text detection with fully convolutional networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4159–4167.
4. Chen, X.; Wei, P.; Ke, W.; Ye, Q.; Jiao, J. Pedestrian detection with deep convolutional neural network. In Asian Conference on Computer Vision; Springer: Cham, Switzerland, 2015; pp. 354–365.
5. Shailja Sharma, Princy, Kirti Bhatia, Rohini Sharma, A Study on Image Categorization Techniques, International Journal Of Multidisciplinary Research In Science, Engineering and Technology (IJMRSET), Volume 6, Issue 5, May 2023, pp. 1147-1152.
6. Kang, K.; Li, H.; Yan, J.; Zeng, X.; Yang, B.; Xiao, T.; Zhang, C.; Wang, Z.; Wang, R.; Wang, X. T-cnn: Tubelets with convolutional neural networks for object detection from videos. IEEE Trans. Circuits Syst. Video Technol. 2017, 28, 2896–2907.
7. Fan, Q.; Brown, L.; Smith, J. A closer look at Faster R-CNN for vehicle detection. In Proceedings of the 2016 IEEE Intelligent Vehicles Symposium (IV), Gotenburg, Sweden, 19–22 June 2016; pp. 124–129.
8. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. Med. Image Anal. 2017, 42, 60–88.
9. Mahum, R.; Rehman, S.U.; Okon, O.D.; Alabrah, A.; Meraj, T.; Rauf, H.T. A novel hybrid approach based on deep CNN to detect glaucoma using fundus imaging. Electronics 2021, 11, 26.
10. Bai, H.; Wu, J.; Liu, C. Motion and haar-like features based vehicle detection. In Proceedings of the 2006 12th International Multi-Media Modelling Conference, Beijing, China, 4–6 January 2006; p. 4.
11. Wei, Y.; Tian, Q.; Guo, J.; Huang, W.; Cao, J. Multi-vehicle detection algorithm through combining Harr and HOG features. Math. Comput. Simul. 2019, 155, 130–145.
12. Qin-jun, Q.; Yong, L.; Da-wei, C. Vehicle detection based on LBP features of the Haar-like Characteristics. In Proceedings of the 11th World Congress on Intelligent Control and Automation, Shenyang, China, 29 June–4 July 2014; pp. 1050–1055.
13. https://www.v7labs.com/blog/object-tracking-guide.
14. arXiv:1510.07945v2 [cs.CV] 6 Jan 2016.
15. http://davheld.github.io/GOTURN/GOTURN.html
16. arXiv:1607.05781v1 [cs.CV] 19 Jul 2016.