International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 06 | June 2022 www.irjet.net p-ISSN: 2395-0072

A Literature Review on Vehicle Detection and Tracking in Aerial Image Sequence using Deep Learning Technique

Niyazahamad S K[1], Sanjay Subramanya[2], Rishab Chandrashekar Shirur[3], Ratan Prakash Shikhari[4]

[1,2,3,4] *Department Of Computer Science And Engineering, Dayananda Sagar College of Engineering, Bengaluru*

Abstract: ***Visual multi-object tracking that is both robust and high-performing is indeed a key difficulty in computer vision, particularly with in context of drones. Small target recognition and tracking in UAV situations is problematic for standard Multi-Object Tracking (MOT) techniques based on the tracking-by-detection paradigm. We will be performing real time vehicle detection and tracking on Aerial Image Sequences using different AIML approach which comprehensively includes techniques such as Image Processing, Pattern Recognition and Computer Vision. It has a wide variety of applications encompassing Visual Surveillance, Traffic Control, Digital Forensics and Human-Computer Interaction.***

I. INTRODUCTION

Visually multi-object monitoring that is both resilient & high-performing is indeed a major difficulty in computer vision(CV), particularly in the context of UAV. Also with huge popularisation for commercialized unmanned aerial vehicles (UAVs) as well as the advancement of OpenCV & AIML technologies, drone detection methods are becoming a hot topic for researchers. Auto-navigation, campus security surveillance, & catastrophe assistance have all become easier thank to effective video image computational techniques and powerful deep neural networks.

II. SURVEY

[1] Multiple Object Tracking Performance: The CLEAR MOT Metrics

Metrics that describe, he quality and key characteristics in numerous object tracking systems must be studied and compared in accordance to carefully analyse and evaluate their performance. Regrettably, there has yet to be agreement on such a range of generally valid measures. They present two new measures for evaluating MOT systems in this paper. Multiple object tracking precision (MOTP) as well as multiple object tracking accuracy (MOTA) are suggested benchmarks that can be used for a variety of monitoring activities & permit for objective contrast of tracking systems' primary features, like accuracy at locating targets, precision at recognising target configurations, but also way to detect targets on consistent bases. They put the proposed metrics to the test in a series of global evaluation workshops to see how useful and expressive they were. The CLEAR workshops in 2006 and 2007 featured a wide range of monitoring activities whereby a big number of models were tested & evaluated. Their studies findings reveal that its suggested measures accurately reflect the numerous methods' qualities and shortcomings in a simple and direct manner, helps in easy evaluation in performance, thus relevant toward a wide range of circumstances.

[2] Fully-Convolutional Siamese Networks for Object Tracking

Traditionally, the problem of arbitrary target tracking was tackled by developing a system of the targets arrival entirely online, with only the video as training data. Despite their effectiveness, these approaches' online-only methodology limits using depth information which can be studied. Many efforts have actually been developed towards harnessing deep convolutional networks' descriptive ability. Once the target to monitor isn't determined ahead of time, Stochastic Gradient Descent online is required in adjusting the network's parameters, risking overall system's speed. For object detection in video, a basic tracking method is combined with a novel fully-convolutional Siamese network that has been trained end-to-end on the ILSVRC15 dataset. The tracker reaches state of art success in various tests with the minimal brevity. It works at fps that are faster than actual.

[3] Simple Online And Realtime Tracking

This research looks at a realistic approach for monitoring many items, with the primary objective of associating objects successfully for online and real-time operations. The study claims that recognition ability is a critical component in determining detection accuracy, with modifying its detector boosting tracking efficiency by up as 18.9 percentage. In contrast to many batch-based tracking systems, this research focuses on online tracking, where

the tracker is only shown detections from the previous and current frames. Despite just employing a simple mix of existing techniques such as the Kalman Filter and the Hungarian algorithm for the tracking components, this approach achieves tracking accuracy similar to state-of-the-art online trackers. This research looks at a realistic approach for monitoring many items, with the primary objective of associating objects successfully for online and real-time applications. The study claims that detection quality is a critical factor in determining tracking performance, with changing the detector boosting tracking performance by up to 18.9%. The tracker also updates at a rate of 260 Hz, which is nearly 20 times faster than other state-of-the-art trackers due to the simplicity of the tracking method.

[4] High-Speed Tracking-by-Detection without Using Image Information

As the performance in object detectors increases, the foundation for a tracker becomes significantly more trustworthy. The problems for a successful tracker have changed as a result of this, as well as the increased use of higher frame rates. As a result of this shift, considerably simpler tracking algorithms may now compete with more complex systems for a fraction of the processing cost. This paper outlines and illustrates such a method by conducting extensive tests with a range of object detectors. The proposed technique can easily operate at 100K fps on the DETRAC vehicle tracking dataset, beating the state-of-the-art. The notion of a passive detection filter is used to analyse a very simple tracking technique in this research. Due to its modest computing footprint, the suggested approach can serve as a basic predictive model for other trackers and provide an appraisal of the necessity of additional efforts in the tracking algorithm. It also permits reviewing tracking benchmarks to evaluate if the specific concerns they indicate (for instance, missed detections, frame rate, etc.) are within the capabilities of existing algorithms.

[5] Cascade R-CNN: Delving into High Quality Object Detection

The this study, they presented the Cascade R-CNN, a multi-stage object recognition framework for developing high-quality object detectors. Overfitting during learning and quality disparity during inference have both been demonstrated to be avoided using this design. On the hard COCO and famous PASCAL VOC datasets, the Cascade R-substantial CNN's and consistent recognition improvements show that effective object detection necessitates modelling and knowledge of several corroborating aspects. The Cascade RCNN has been shown to work with a wide range of object detection architectures. They hope it will be useful in a variety of future object detection research projects.

[6] Real-Time 'Actor-Critic' Tracking

Object tracking is one of the most important step in object detection and tracking. In this paper Actor-Critic framework been used, where the 'Actor' model seeks to infer the best option in a continuous action space, causing the tracker to move the bounding box to the object's current location. The 'Critic' model is used for offline training to create 'Actor-Critic' framework along with reinforcement learning as well as a Q-value for directing both the 'Actor' and 'Critic' deep network learning processes. Visual tracking is viewed as a dynamic search process in which the 'Actor' model outputs only one action to locate the tracked object in each frame. Offline training of better policy for finding the best result is done using reinforcement learning. Furthermore, the 'Critic' network serves as a verification system for both offline and online instruction. Using popular benchmarks, The suggested tracker gets contrasted to certain state of art trackers, as well as the stimulation finding reveal that it performs well in actual.

[7] Hybrid Task Cascade for Instance Segmentation

At instance level, segmentation process is basic computer vision job which identifies objects per-pixel. In real-world settings like automated driving and video surveillance, precise and reliable feature extraction is challenging to achieve. Cascade would be a basic but efficient design which has increased results on a wide range of workloads. A basic Cascade R-CNN and Mask R-CNN combination produces just a little boost. The secret to good instance segmentation cascade would be to completely use inverse interaction across detection as well as segmentation so that to discover a more effective technique.

They propose Hybrid Task Cascade (HTC) in this paper, that is different in two key ways: (i) it intertwines these two tasks for simultaneous multi-stage computation, rather than conducting cascaded refinement on them individually; and (ii) employs a convolutional section in order to give spatial features, that help distinguish difficult frames in cluttered background. Bounding box analysis plus masked predictions is coupled inside a multi-tasking way at each step of HTC. At certain stages, easily applicable within its masked sections are also given - the masked characteristics in every step are combined and supplied to the next. The whole design increases data flow inside the activities as well as stages, resulting in good refined predictions at all levels and more reliable forecasts overall. HTC is simple in setting up & may be programmed from beginning to end. It gained 2.6 % & 1.4 % greater masked AP than that of the Masked R-CNN as well as Cascade Masked R-CNN benchmarks, respectively, on difficult COCO benchmark.

[8] MMDetection: Open MMLab Detection Toolbox and Benchmark

CV tasks like target recognition & instance segmentation are both fundamental. Also the detection framework's pipeline is typically more complicated than that of classification jobs, and various implementation parameters might produce drastically different results. In this paper they are having the goal of providing a high-quality codebase and unified benchmark, for this they have built MMDetection model. Major features of MMDetection (1) Modular design - They split the detection framework into separate components, allowing users to quickly build a customised object detection framework by combining different modules. (2) Out-of-the-box support for multiple frameworks. Popular and current detection frameworks are supported by the toolbox. (3) High efficiency - GPUs handle all fundamental bbox and mask operations. Other codebases, such as Detectron, maskrcnn-benchmark, and SimpleDet, have training speeds that are quicker or equivalent. (4) State-of-the-art - The toolbox is based on the software created by the MMDet team, who won the 2018 COCO Detection Challenge.

[9] A Cost Effective Tracking System for Small Unmanned Aerial Systems

The paper describes a object tracking system for UAV systems is described. It is built on object recognition with a sum of absolute differences similarity measure. Every algorithm loop contains a template change that incorporates new data so that the former stays valid, allowing the track to proceed even if the object's appearance changes. It also enables the human operator to enter targets manually and receive data from other picture processing units..

The paper supports the importance of uav vision model suggesting a small, imprecise, non-gyro-stabilized architectural method for detecting & monitoring many targets in real-time by not making any rigid assertions on trajectories, depending onto limited input of their surroundings and also with avoiding necessity motion compensation, which usually necessitates the need for a inertial measurement unit (IMU). This approach eliminates the additional expense & cost of a binocular model and those of gyro-stabilized turret instead opting for small, unstable system. N notion that camera doesn't always need to be corrected for such model to work reduces the building & configuration procedure even further. Even more simplifies the assembling and setup procedure. The idea can run in real time on low-cost, low-power computer systems thanks to its ability to track several things at a cheap computational cost. It's highly crucial because automated driving vehicles with restricted payload capacity, whom it is impractical to have a substantial energy backup for sustaining the computer. Despite the fact that current advances in recent computer architecture & powerful batteries assurances that meet the issue by supplying more productive, advanced machineries and greater power intensity battery, a urge of a simpler style to vision systems integration will almost definitely stay important. This is particularly true for smaller unmanned aerial systems, such as the Maxxi Joker 2 system employed in this research.

[10] The research on visual object tracking algorithm based on an adaptive combination kernel

The research suggested a visual object tracking approach based on the Adaptive Combination Kernel to increase the resilience to intricate transitions of multiple objects and a complex backdrop picture. The object tracking approach has indeed been divided down into separate subtasks to approximate the object's details: Translation Filter as well as Scale Filter. At commence, the Translation Kernel Tracker utilizes a new Linear Kernel Filter as well as Gaussian Kernel Filter(GKF) pair. The goal function, which incorporates not just overfitting problem but also the highest amount of response output for every kernel, was used to determine the weight coefficients in both the Linear Kernel filter as well as GKF. The advantages of both the local plus global kernels are combined in the Adaptive Combination Kernel. Next, the tracker position was identified by using response result of the dynamic combination kernel correlation filter. Furthermore, the interpretation filter has been constructed with a scene-adaptive training data depending on the highest response score. The effective learning speed can be used to update the translating filters. Lastly, the item scale was estimated using a 1D magnitude filters. Compared to previous techniques, the proposed algorithm is more robust to deformation and occlusion.

[11] Dual-Channel Convolutional Neural Networks-Based Image Super-Resolution Algorithm

By using picture super-resolution technique on single network, it's really challenging in accomplishing combined high-quality pattern reconstruction and quick converging. To overcome the drawbacks of earlier approaches, the current study proposes a picture high-resolution strategy built upon dual-channel CNN (DCCNN). A deep tunnel as well as a shallow tunnel were created in the system model's novel architecture. The shallow tunnel was largely employed to maintain the original picture's general shape, whereas the deep tunnel has been used to retrieve specific feature data. To begin, during feature extraction phase, the leftover frame was altered, as well as the channel's nonlinear mapping capability was increased. After the characteristic mapping scale was reduced, the picture's effective features were recovered.

During up sampling step, deconvolutional kernel's general variables are tweaked, and high bandwidth network degradation were minimised. High-resolution texture area may get recreated iteratively employing long as well as the small data chunks in recreation steps, increasing texture information recovery even further. Second, the convolutional kernel of narrow network was adjusted for decreasing the variables, Suring that the whole contour of picture is recovered & the channel narrowed faster. Lastly, double channel error rate is simultaneously changed to increase capability to fit features in terms of achieving a most recent high-resolution picture result.

[12] Enhanced Hierarchical Principal Component Analysis Method for Detecting Saliency

A saliency identification technique formed upon Hierarchical Principal Component Analysis (HPCA) was created with in study to solve the challenges of existing important item identification approaches, such as severe environmental noise, less accuracy, and high processing performance. After transforming the RGB picture to monochrome, the monochrome image was split in eight layers and used the digital surface stratification approach. Important subject data correlates to the layered image characteristics in each picture layer. Second, the monochrome picture is reinitialised using the monochrome colour conversion technique, which uses the initial image's colour layout like a source images, resulting in a tiered picture that not only represents the initial structural characteristics and also effeciently conserves the initial picture's color information. Moreover, Principle Component Analysis (PCA) was used on layered picture to identify the structural and colour differential features of every tier with in principle component orientation.

To get a saliency layout having great resilience as well as to improve our findings further, two characteristics have been combined: recognized prviously were added to picture organisation, which may pinpoint its photograph's topic around the image's centre. Finally, the entropy computation was used to produce the idealized picture as from multilayered saliency layout; the best layout contains the lowest background as well as most clearly saliency entities compared to the others.

[13] Online Multi-Object Tracking Using CNN-based Single Object Tracker with Spatial-Temporal Attention Mechanism

A CNN-based architecture for online MOT is proposed in this research. This framework makes use of the advantages of single object trackers When it comes to altering the shape of models & searching for the objective within next frame, There are concerns with processing efficiency and occlusion-induced drifting outcomes when employing the single object tracker to MOT. The given approach achieves computational efficiency besides wanting to share characteristics and using ROI-Pooling for obtaining individual characteristic to every target. The appearance model is adjusted in every target using any target-specific CNN layers learned online. Introduction of the spatial-temporal attention mechanism (STAM) in the framework was carried out to deal with drift caused by occlusion and target engagement. A target's visibility plan was learnt and used to infer its spatial attention plan.

Characteristics are then weighted using the spatial attention map. Furthermore, the occlusion state can be evaluated using the visibility mapping, that regulates a continuous updating mechanism using weighted loss upon training samples having varying occlusion states over multiple frames. It's possible to think of it as a temporal attention process. On the rigorous MOT15 and MOT16 benchmark datasets, the suggested approach obtains 34.3 percent and 46.0 percent in MOTA, respectively.

[14] Deep Learning in Video Multi-Object Tracking

Although the target's presence is understood in advance in SOT, MOT needs the detection stage for detecting the objects which might leave but rather reenter the frame. One of most difficult aspect of monitoring many objects at once is the numerous occlusions as well as interactions amongst target, which might sometimes seem toward being identical. Tracking-by-detection is the standard strategy used in MOT algorithms: The tracking procedure is guided by a collection of detections (i.e. enclosing box denoting all objects inside picture) extracted out from image sequences.

The following steps are found in the great majority of MOT algorithms:

- Object detection stage: The target identification method analyses every input video frame using bounding boxes to discover target belonging here to target class, often called by 'detections' as in domain of MOT..

- Extraction of features prediction stage: One or many feature retrieving techniques investigate these identification plus tracklets in order to obtain information about look, movement, and interactivity. Each monitored target's upcoming location may be predicted using a motion prediction.

- Affinity stage: By giving the same ID for detections which indicate the same target, similarity metrics are utilised to correlate detections with tracklets pertaining to target.

- Association stage: By giving the same ID for identification which indicate the same object,

similarity metrics have been used to correlate identification with tracklets pertaining to the target.

[15] Joint Multi-view Face Alignment in the Wild

In this paper, Using face detection model followed with a deformable model fitting on the bounding box is the de facto approach to estimate facial landmarks. It entails two major issues:

- A detection & deformable fitting processes were done separately, as well as the detector may or may not offer the optimum initialization to fitting step,
- A look of a face varies greatly with postures, making changeable facial fitting extremely difficult, necessitating the usage of several images.

They demonstrate the first joint. To the best of their knowledge, multiview convolutional networks can handle significant postural variances across facial with in field and neatly integrate facial identification and facial landmark localisation tasks. Currently present combined facial detection & landmark positioning methods mainly evaluate limited number of landmarks. Their technique can recognise as well as match a huge proportion of markers in mid-frontal (68 landmarks) as well as profiles (39 landmarks) faces. They put their system through its paces on range of samples, namely COFW, 300W, IBUG, as well as the current Menpo Standard both for mid-frontal as well as profile faces. There is considerable enhancement over state of art methods in deformable face recognition on 300VW benchmark. Also on FDDB and MALF benchmark, they also gives state of the art face recognition results.

[16] Object Detection from Video Taken by Drone through CNN

The study used CNN to recognise objects in footage taken by the drone. They suggest employing CNNs in this work to allow drones can discriminate between different entity types like buildings, automobiles, trees, and humans. Despite the fact that convolutional neural networks are computationally expensive, they are trained with smaller picture datasets using the approach transfer learning. They used TensorFlow's sophisticated object detection API in their project, which allowed them to quickly build a new model and deploy it for detection. According to the data, the two models' detection rate for houses, trees, cars, & pedestrians is fairly good, with just an mean of more than 85 % and a peak of 99 percent. They compared the memory consumption, speed, and accuracy of two current object detectors in an experiment. In comparison to Faster RCNN models, SSD concepts give more importance to size, ratio, as well as predicted sample location, with an average frame time of 115 ms and a low target identification rate. R-CNN that is faster and precise but also discovers many items in the image; about 95% among all targets in frame can be recognised, yet every frame takes on average 140 milliseconds. In general, their experiments show that SSD outperforms the HDD model in the long term.

[17] Rich feature hierarchies for accurate object detection and semantic segmentation

According to this paper, Object detection performance has been stagnant in previous years. Complex ensembles combine several low-level picture features with high-level context from object detectors and scene classifiers to produce the best results. This research proposes a simple and scalable object detection technique that improves on the best recent findings on PASCAL VOC 2012 by 30 percent. They were able to attain this level of success because to two key observations. This one was to localise and segment objects using bottom-up region recommendations utilising high-capacity CNN. The 2nd is a strategy of teaching large CNNs when labelled training data is limited. They demonstrate that pre-training the network for an auxiliary task with a large amount of information (object classification) and afterwards tuning the system for primary goal with minimal input is very effective (detection). They believe that the "guided pre-training/domain-specific finetuning" approach will work well for a range of vision difficulties with little data. They end by emphasising how remarkable it is that they were able to obtain these results utilising a combination of traditional computer vision(CV) as well as deep learning(DL) tools. Instead being antagonistic paths of rational research, they are closely intertwined.

[18] Deep Residual Learning for Image Recognition

According to paper, hierarchical neural networks are difficult to be trained. They offer an residual learning strategy for significantly deeper training models than earlier used models. Researchers specifically reformulate the levels as training residual methods with relation to the levels inputs, rather than learning unreferenced functions. Researchers presented significant experimental proof suggesting that residual networks is simpler to implement and that increasing complexity can boost performance. Authors used the ImageNet data to evaluate residual nets having put to 152 levels of complexity, that are 8 layers deeper to VGG networks whereas have less intricacy. The aggregation of such residue nets scores 3.57 % error just on ImageNet test set. The work won 1st spot with in ILSVRC 2015 classification problem. CIFAR-10 study using 100 - 1000 levels is shown. The complexity in depictions is crucial for many image identification tasks. Only because of their exceptionally deep depiction can they accomplish an 28 % notable improvement on COCO object recognition samples. Their contributions towards ILSVRC & COCO 2015

contests used deep residual networks as the basis, and thus won 1st spot in the ImageNet identification, ImageNet localisation, COCO recognition, as well as in COCO segmentation tasks.

REFERENCES

1. Bernardin K, Stiefelhagen R (2008) Evaluating multiple object tracking performance: the CLEAR MOT metrics. EURASIP Journal on Image and Video Processing, 2008, 1–10

2. Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PH (2016) Fully-convolutional siamese networks for object tracking. In European conference on computer vision (pp. 850–865). Springer, Cham

3. Bewley A, Ge Z, Ott L, Ramos F, Upcroft B (2016) Simple online and realtime tracking. In 2016 IEEE International Conference on Image Processing (ICIP) (pp. 3464–3468). IEEE

4. Bochinski E, Eiselein V, Sikora T (2017) High-speed tracking-by-detection without using image information. In 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1–6). IEEE

5. Cai Z, Vasconcelos N (2018) Cascade R-CNN: Delving into high quality object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6154–6162)

6. Chen B, Wang D, Li P, Wang S, Lu H (2018) Real-time'Actor-Critic'Tracking. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 318–334)

7. Chen K, Pang J, Wang J, Xiong Y, Li X, Sun S ... Loy CC (2019) Hybrid task cascade for instance segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp.4974–4983)

8. Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X, ... Zhang Z (2019) MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155

9. Chen Y, Wang J, Liu S, Chen X, Xiong J, Xie J, Yang K (2019)Multiscale fast correlation filtering tracking algorithm based on a feature fusion model. Concurrency and Computation: Practice and Experience, e5533

10. Chen Y, Wang J, Xia R, Zhang Q, Cao Z, Yang K (2019) The visual object tracking algorithm research based on adaptive combination kernel. J Ambient Intell Humanized Comput 10(12):4855–4867

11. Chen Y, Wang J, Chen X, Sangaiah AK, Yang K, Cao Z (2019) Image super-resolution algorithm based on dual-channel convolutional neural networks. Appl Sci 9(11):2316

12. Chen Y, Tao J, Zhang Q, Yang K, Chen X, Xiong J, ... Xie J (2020) Saliency Detection via the Improved Hierarchical Principal Component Analysis Method. Wireless Communications and Mobile Computing, 2020

13. Chu Q, Ouyang W, Li H, Wang X, Liu B, Yu N (2017) Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4836–4845)

14. Ciaparrone G, Sánchez FL, Tabik S, Troiano L, Tagliaferri R, Herrera F (2020) Deep learning in video multi-object tracking: A survey. Neurocomputing 381:61–88

15. Deng J, Trigeorgis G, Zhou Y, Zafeiriou S (2019) Joint multi-view face alignment in the wild. IEEE Transactions on Image Processing 28(7):3636–3648

16. Fan, D. P., Wang, W., Cheng, M. M., & Shen, J. (2019). Shifting more attention to video salient object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8554–8564).

17. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580–587)

18. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778)