ISSN No:-2456-2165

Multiple Object Tracking for Video Analysis and Surveillance: A Literature Survey

Advitiya C S., Adarsh R Shenoy, Shravya A R., Abhishek Battula, Akash Raghavendra, Dr. Krishnan R
Department of Computer Science and Engineering
Dayananda Sagar College Of Engineering
Bangalore, India

Abstract:- Multiple Object Tracking (MOT) is the detection of unique objects and their movements through frames. There are various issues in Multiple Object Tracking like occlusions, similar appearance of different objects, interaction among multiple objects, etc. There are even more issues with Object Tracking in aerial image sequences due to weather conditions, small sizes of objects, distortion in proportions, etc. Even with the various techniques and methods developed for MOT, only a fraction are able to perform on aerial image sequences with the rest focusing on ground level image sequences. Research in Multiple Object Tracking has been gaining a lot of attention with an increasing trend in the number of Multiple Object Tracking research papers published each year. This paper focuses on documenting the advancements in Multiple Object Tracking in recent years, with extra attention paid to techniques that help with aerial image sequences. The various architectures and techniques are classified and are compared along with their advantages, disadvantages and performance scores, with detailed mention of the datasets generally used for various applications.

Keywords:- *Multiple object tracking, aerial images, attention networks, occlusion, pedestrian tracking.*

I. INTRODUCTION

In recent years, Multiple Object Tracking has gained a lot of attention and has become one of the more important tasks in computer vision. Object tracking is usually done in two phases where an object is first detected and uniquely identified and is then tracked as it moves through the frames in phase two. Multiple object tracking performs the detection and tracking over multiple objects in the same frames. The detection phase must be able to handle situations with deformation, varying illumination, cluttered or textured background. etc. Re-identification and association of objects is also very important and dictates the accuracy of the tracking as there are various issues like occlusion, similar appearance in different objects, interaction among multiple objects. etc. The object detection and tracking in aerial images specifically has been a problem for a long time. Even with the rise of numerous methods which deal with sequences involving people and objects, issues that arise by applying the same to aerial image sequences need to be addressed. These issues include smaller object sizes, weather conditions, change of scales, and distortion in proportions. With the increasing trend in research of MOT, more and more approaches have been looked into to counter these various issues. These methods are looked into and tabulated with more emphasis on object tracking in aerial image sequences in this paper, with a comparison of their performance scores, advantages, disadvantages and particular scenarios they excel at.

II. FUNDAMENTAL ARCHITECTURES

A. Convolutional Neural Networks (CNN)

An improvement of the artificial neural network known as the convolutional neural network (CNN) is excellent for pattern recognition in images and is mostly employed for image processing and recognition. The input layer, the hidden layer, and the output layer are the three layers that make up the CNN. The input layer receives the initial data that will be processed further. With the help of the activation function, the hidden layer uses this collection of weighted inputs to generate an outcome. The necessary output is subsequently produced by the output layer.

B. Recurrent Neural Networks (RNN)

Recurrent neural networks are a modification of feedforward neural networks with the addition of internal memory. RNNs are iterative in nature as they perform the same function for each data input, but the output for the current input depends on the last computation. Unlike feedforward neural networks, RNNs can use internal states (memory) to process a sequence of inputs. This way you keep the context remembered during training.

C. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are modified versions of RNNs but are more efficient in remembering past data in memory. The RNN vanishing gradient problem is solved here. The LSTM network has three gateways.

- Input Gate — Determines which value from the input is used to modify the memory. The sigmoid function determines the value passed.
- Forgotten Gates - Determines details that are discarded from blocks. This is determined by the sigmoid function. Forget gates help solve the vanishing gradient problem in RNNs.
- Output Gate - Uses the block's input and memory to determine the output. The sigmoid function determines the value to pass and the tanh function weights the value.

D. Attention Based

While training an image model, the model should be able to focus on key elements of an image. This can be accomplished through attention mechanisms. Attention can be described as a function mapping a query vector Q together with a key-value vector pair K, V to an output. The

weights V are computed by the softmax expression. Combining these operations and running it parallelly is termed as Multi-Head Attention The Transformer is an encoder-decoder model which uses a multi-headed attention mechanism to improve training speed. This model was designed to tackle problems in natural language processing but is now vastly used in computer vision tasks since they solve the problem of long-term dependencies. The encoder takes the inputs combined with the positional encodings to retain the order of sequential data. A layer in the encoder consists of the multi-head attention block which performs self-attention. The output is fed into a feed forward neural network. The decoder works similarlybut also receives keys and queries from the encoder. Finally, the decoder output is fed into a feed-forward and softmax layer to produce probabilities for the next item.

Ref. No.	Algorithm/Technique	Results	Datasets Used
[4]	The model has three main components- a convolutional neural network backbone to get feature representation, a transformer, and a simple feed forward network that makes the final prediction.	The main results of this paper are that DETR can achieve results comparable to an optimized Faster R-CNN network on the COCO dataset. This means that the transformer network can be used to detect objects in images with performance similar to the well-established Faster R-CNN baseline.	COCO 2017 and panoptic segmentation datasets
[5]	A network using background attention that is weakly supervised that merge various scales of feature maps	The results of this paper show that the proposed Guided Attention Network (GANet) outperforms existing methods for object detection and counting on the CARPK, PUCPR+ and UAVDT datasets. Moreover, GANet also achieved better results in terms of speed and computational efficiency.	UAVDT, CARPK,PUCPR
[6]	Semi-supervised Multi-Task Network with Self-Attention	This paper shows that by using a multi-task learning approach with self-attention, we can improve the accuracy of object detection on two traffic surveillance datasets, UA-DETRAC and UAVDT. This model could potentially help with instance segmentations at almost no cost.	UAVDT, UA-DETRAC
[7]	Dense anchor scales with large scale variance for detection. Squeeze-and-Excitation blocks are used to capture the channel dependencies. Deep association network is used after the detection module and feeds the generated hypotheses to the DeepSORT network.	The proposed model was able to accurately detect objects in drone images with a maximum of 500 detections using denser anchor scales with large scale variance, Squeeze-and-Excitation (SE) blocks, and a trained deep association network.	VisDrone2019
[8]	SMSOT-CNN, GOTURN (Tracking Using Regression Networks)	The created model is able to accurately and efficiently track multiple pedestrians and cars in drone captured images. The results of the ablation study on the dataset showed that the proposed approach achieved the highest accuracy and the shortest execution time.	KIT AIS, DLR's Aerial Crowd Dataset
[9]	Attentional Asymmetric Siamese Network and Discriminative Correlation Filter	The results of this paper show that the proposed DCF-ASN tracking framework, which combines the discriminative correlation filters (DCF) with an asymmetric siamese network (ASN), achieves the state-of-the-art performance on five popular tracking datasets.	UAV123 and UAVDT
[10]	A module for data augmentation. An algorithm to merge the tiny objects into several clusters of similar sizes. The candidate center points and the size of the clusters are found out and a network finely calculates the center points and	The network can effectively detect objects from aerial images, with precision results higher than the state-of-the-art approaches when it was implemented.	VisDrone and UAVDT

sizes of the small objects.

[11]	The TLD algorithm and the ATLD algorithm	The results of this paper show that the Appearance and Tracking Learning Detection (ATLD) algorithm outperforms the original Tracking Learning Detection (TLD) algorithm in terms of accuracy. The ATLD algorithm also performs better than other benchmark learning based algorithms.	Aerial sequences from the UCF website, TLD dataset and various classified sources
[12]	LSTM, GCNN and a siamese neural network module	The results of this paper show that the proposed AerialMPTNet method on the pedestrian datasets, outperforms all other methods before it., and achieves competitive results on the vehicle dataset. Also, the results show that adding LSTM and GCNN to the algorithm which tracks improves the tracking performance.	KIT AIS, DLR-ACD and AerialMPT
[13]	The accuracy of detection is improved by separating the ReID and detection branches into two parts. This also makes the two parts more independent. Temporal information in target detection and the ReID head.	The results of this paper show that there is an improvement of performance on the tracking of multiple objects compared to other models dataset UAV video. Additionally, the model was able to detect more targets than the baseline model and reduced the issues of false and missed detections.	VisDrone2019
[14]	Built using MMDetection and Pytorch. Different modules are used for detection and for extraction of features . Merging of detected focals is done with the help of NMS and obtaining of final predictions is done using IBS.	The results of this paper shows that the detection of object framework of two stages achieved a 42.06 AP score on the validation dataset of VisDrone, outperforming all other small detection of object methods seen in the literature.	VisDrone and UAVDT

Table 1: Summaries of papers focused on Aerial Image Sequences

Ref. No	Algorithm/Technique	Results	Datasets Used
[15]	CNN, Spatial-Temporal Attention Mechanism	The model shows the falsely classified negatives and positives, switching of identities of different objects and other errors that can occur when tracking multiple objects. The efficiency displayed by the algorithm is shown by the results of the accurate tracking of multiple objects.	MOT 15, MOT16
[16]	ResNet-50, Transformer with multihead attention	It shows the falsely classified negatives and positives, switching of identities of different objects when calculating the accuracy. TransTrack was found to be a good matchup against other methods.	MOT17, MOT20, CrowdHuman
[29]	ResNet-34, CenterNet	While the time spent on re-ID matching grows linearly with density, the time spent on joint detection and re-ID is only slightly impacted by density.	CrowdHuman, 2DMOT15, MOT16, MOT17 and MOT20
[17]	Uses convolutional block attention module (CBAM), constructed using a spatial temporal network on a ResNet-50 Backbone	uses a cost-sensitive tracking loss to concentrate on negative distractions and attention networks which consist of both temporal and spatial mechanisms are used to suppress noisy observations. These mechanisms help the algorithm to track objects in video frames better than both online and offline trackers, as indicated by the identity-preserving	MOT 16, MOT 17

		metrics.	
[18]	R-FCN architecture with an encoder decoder block and Kalman Filter	The R-FCN (Region-Based Fully Convolutional Network) is a deep learning model that performs the majority of its calculations on the entirety of the image. ReID (Person Re-Identification) features are deep-learned appearance representations that are trained on huge datasets which focus on the re-identification of people to increase the ability of identification.	MOT 16, IDRI
[28]	The main neural network used is ResNet-50. It generates multi-scale feature representation by integrating Feature Pyramid Networks.	The model takes into consideration various factors, such as false negatives, false positives, fragmentation and identity switches. The proposed model is also faster and more simple than existing methods, and it does not require any extra training data.	MOT16, MOT17
[19]	Attention based model that makes use of higher order attention mechanisms for ReIdentification.	The results of the experiments performed to validate the precedence of the MHN for person re-identification show that it outperforms a wide range of state-of-the-art methods on various datasets, which include Market-1501, DukeMTMC-ReID and CUHK03-NP.	Market1501 ,DukeMTMC-ReID and CUHK03-NP
[20]	Bidirectional recurrent neural network (Bi-RNN) which determines the assignment matrix based on the prediction1 to-ground-truth distance matrix	The results of this paper show that the proposed framework increases the performance of existing multi-object trackers, and on the MOTChallenge benchmark, it established a new state of the art score.	MOT15, MOT16, MOT17
[21]	Variational Bayesian Model, VEM algorithm	The paper's results show that the proposed OVBT (Online Variational Bayesian Tracker) algorithm performs well on the MOT 2016 dataset, with low accuracy (MOTA) but high precision (MOTP). This is likely because the algorithm is sometimes unable to detect targets or misidentifies them, leading to identity switches (ID) and missed targets (FN). The results also show that when multiple observations are considered within the visibility process, performance is improved for all sequences and most measures.	MOT 16
[22]	Joint Detection and Embedding with the DarkNet-53 backbone network. A feature pyramid network is also used.	The results of this paper show that the proposed MOT system is the first real-time MOT system, with a high running speed of about 22 to 40 frames per second. It also has a high MOTA score	ETH, CityPersons, CalTech, MOT16, PRW and CUHK-SYSU datasets
[23]	Global Context Disentanglement, Deformable Attention, Guided Transformer Encoder	The paper demonstrates the precedence of the proposed RelationTrack MOT framework. The experiments conducted on the MOT20, MOT16 and MOT17 benchmarks show that the proposed framework has established a new state-of-the-art performance and has surpassed preceding methods.	MOT16, MOT17 and MOT20
[24]	Joint-object-detection-and-tracking system(Transformer-based)	The paper presents a joint-detection-and-tracking system called PatchTrack that uses current frame of interest patches in order to infer both appearance information and object motion.	MOT16, MOT17, CrowdHuman

[25]	Uses YOLOX detector and performs association between the detection boxes and tracks	ByteTrack can significantly improve the IDF1 score. It has achieved a high MOTA performance	MOT 17, MOT 20, ETHZ
[26]	A frame-level hypothesis generation module, a track-level memory encoding module, a memory decoding module	The results of the paper show that MeMOT achieves better performance on normal as well as crowded scenarios. MeMOT also achieves good performance on the MOT Challenge benchmarks for pedestrian tracking, outperforming other methods.	MOT16, MOT17, MOT20
[27]	Encoder-decoder transformer	The model achieves high performance for both public and private detections on MOT17 and MOT20. It is able to track objects for a long period of time accurately.	MOT17, MOTS20

Table 2: Summaries of papers focused on Pedestrian Tracking

III. STRUCTURED RELATED WORK

Vaswani, Ashish, et al.[1] introduces the Transformer model which follows the encoder-decoder approach. The encoder extracts features from the input and passes it onto the feed forward neural network. The decoder receives keys and queries from the encoder block. The model is auto-regressive which means that it uses the previously generated outputs as the input for the present step. Transformer model, which is based solely on attention mechanisms, a better quality was achieved compared to RNN and CNN. Using parallelization the Transformer model took significantly less time to train.

Beheim, Tsuyoshi [2] discusses tracking of multiple vehicles in aerial image sequences. To establish a benchmark, MOT methods from different fields are applied to an aerial dataset. Based on that, several adjustments are made to examine the impact on the methods' ability to track. This proposed collection includes a motion predictor, object re-identification, and a vehicle orientation prediction module.

Jetley, Saumya, et al. [3] employed attention maps to locate and utilize the useful spatial support of visual data that CNNs use to make classification predictions. The paper features a scalar matrix that shows how important layer activations are in relation to the goal at various 2D spatial locations. When the technique was used throughout a network, the model's performance improved drastically.

Carion, Nicolas, et al. [4] proposed a novel approach where the detection of objects is treated as a direct set prediction issue. The method simplifies the pipeline by the elimination of the requirement for numerous elements created manually, like a suppression mechanism or generation of anchors that specifically convey the past information we know about the task. The key components of the proposed framework is a transformer architecture with encoder and decoder and a global loss that makes sure distinctive predictions are made using bipartite matching. Given a predetermined, constrained limited set of item queries, The transformer through its understanding of the different objects and the relationships between them, and the overall context of image, gives the final predictions as output.

Cai, Yuanqiang, et al. [5] introduce the guided attention network for the detection of objects in scenes that are captured using drones. The method is an anchorless approach where it fuses the feature maps of different scales by making use of the background attention to learn background discriminative representation and makes use of the foreground attention module to examine the local view of the object. It uses data augmentation on training data to synthesize the brightness of the images from different settings and noise to imitate different weather conditions which leads to better accuracy.

Hughes Perreault, et al. [6] implemented a strategy that makes use of multi-task learning to create a network with attention. Segmentation labels are generated for the foreground and the background in a supervised and unsupervised manner to train visual attention via background subtraction or optical flow. With the help of these labels, an object detection model is trained to generate bounding boxes and foreground/background segmentation maps while sharing the majority of model parameters. The segmentation maps are employed inside the network to weigh the feature map that led to the creation of the bounding boxes, reducing the weightage of irrelevant parts of the image. The model is trained to do multiple tasks - segmentation and bounding box detection.

Jadhav, Ajit, et al. [7] created a model for object recognition in aerial view photos. The model is built on top of the RetinaNet model. The anchor scales are adjusted for dense distributions and smaller objects. Squeeze-and-Excitation (SE) blocks are used for channel interdependencies. This contributes to large performance increases at a small additional computational cost. A custom built DeepSORT network is used for the detection of objects using the above architecture on the VisDrone2019 MOT dataset.

Bahmanyar, Reza, Seyedmajid Azimi, and Peter Reinartz [8] propose a model based on CNNs. The CNN extracts features from two consecutive frames, one from the current frame and one from the previous ones for each object. The object which is to be tracked is obtained from the previous frame and the search region is extracted from the current frame. These frames are then processed and the locations of the objects increases and facilitates parallelism

Xizhe Xue, et al. [9] proposes a tracking framework that first coarsely infers the state of the target using a discrete correlation filter module and then accurately locates the object using a trained Asymmetric Siamese Network. Tracking is done with the discrete correlation filter trackers, which can be efficiently made use of with a frequency domain transform. Using DCF instructions and the weights of the channels learnt from the annotated data, the model refines the feature representation and accurately finds the object.

Tang, Ziyang, Xiang Liu, and Baijian Yang [10] propose a network which employs a module to improve the unbalanced datasets, a detector without anchors to coarsely estimate the tiny object clusters center points, and another detector without anchors to finely increase precision. To improve the viability of detecting dense tiny objects, an adaptive merging technique is used with a hierarchical loss function that is designed to optimize classification.

Malagi, Vindhya P, Ramesh Babu DR, and Krishnan Rangarajan [11] propose a technique for tracking multiple and single objects in aerial photos termed aerial tracking learning detection that is according to the well-known algorithm TLD (Tracking Learning Detection) that tracks using both motion and appearance information. The established algorithm includes adjustments for movement of camera, algorithmic alterations that integrate appearance and motion cues for multiple object tracking and detection, and improvements of distance between objects for the enhancement of tracker performance if there are several similar objects close-by.

Azimi, Seyed Majid, et al. [12] assessed a range of conventional and Neural Networks based Single Object and Multi Object Tracking algorithms on how they handle various issues in tracking multiple pedestrians and vehicles in aerial images of high-resolution. The proposed Deep Learning system combines graphical, temporal, and appearance data utilizing a Siamese Neural Network, LSTM, and a GCNN module for more precise and steady tracking. Additionally, Online Hard Example Mining and Squeeze-and-Excitation layers are looked into to see how they affect the system's performance.

Lin Y, Wang M, Chen W, Gao W, Li L and Liu Y [13] implement a technique based on FairMOT that suggests an enhanced multiple object tracking paradigm. The detection and ReID procedures are separated in the designed structure of the model to reduce the effect one function has on the others. A temporal embedding structure is also created to enhance the representation capacity of the model. The performance of the model in object tracking and object detection tasks is enhanced by merging temporal-association structures and separating distinct processes.

Koyun, Onur Can, et al. [14] suggest a two-stage object detection system to handle the issue of small object detection. The focused zones are produced by clusters of objects in stage one, which comprises a network which is used for the detection of objects under the supervision of a probabilistic model. The second stage, also an object detection network, predicts objects in the focus zones. To get around the truncation impact of the region search methodology, a method of suppression is also put forth. After combining the predicted boxes, overlapping boxes are suppressed using the standard Non-Max Suppression (NMS) algorithm.

Chu, Qi, et al [15] apply single object tracking techniques to MOT. They make use of dynamic CNN-based spatial temporal networks with ROIPooling and shared CNN characteristics. Features are obtained from the search area provided by the motion model. These features are weighted using spatial attention. The candidate with the highest score is marked as the target state. Samples which were previously positive are also used updating the tracker.

Sun, Peize, et al. [16] introduce a simple yet effective solution to the multiple object tracking problems, proposed in this paper as TransTrack. It employs the transformer architecture. The CNN takes the image frame as input and outputs the feature map. This feature map along with the previous map is passed on to the encoder to generate common features. Two decoder blocks are used, one takes the features and determines the detection boxes, while the other generates tracking boxes from previous frames. Iou algorithm is used to associate these boxes.

Zhu, Ji, et al. [17] implement a technique based on single object tracking. The model uses a spatial attention network which determines the area of interest by feature extraction. A single object tracker monitors each target in each frame. The track-let consists of neighboring frames to preserve the trajectory of occluded objects. The temporal attention network is made up of bi directional LSTMs. Weighted average pooling is used to determine if the identified object newly entered the frame or had occluded. The target is placed in a lost state and the tracker is stopped when the tracking process becomes unreliable. The framework can handle both noisy and occluded data.

Chen, Long, et al. [18] use an encoder-decoder network with a fully convolutional architecture. One image frame can forecast the score maps for the entire image. The encoder component is a shallow convolutional backbone. Features from the encoder network are concatenated with up-sampled features in the decoding part to capture both the semantic and low-level information.

Chen, Binghui, Weihong Deng, and Jiani Hu. [19] propose a module in order to capture subtle differences among pedestrians and offer discriminative attention recommendations. It is designed to represent and exploit complex and high-order statistical information in the attention mechanism. ReID is seen as a zero-shot learning

problem that explicitly enhances the richness and discrimination of attention information by using the Mixed High-Order Attention Network.

Xu, Yihong, et al. [20] suggest a module that simulates the Hungarian matching process. In order to directly improve deep trackers, the module enables evaluating the correlation between ground truth objects and object tracks. There are currently just a few submodules that are trained using loss functions, and these loss functions frequently do not correspond with standard tracking assessment metrics like MOTA and MOTP.

Ban, Yutong, et al. [21] suggest a variational Bayesian model for tracking multiple persons. A variational expectation maximization algorithm results from this. Due to the use of closed-form equations for both the estimate of the parameters of the model and the posterior distributions of the latent variables, the proposed method is computationally efficient. The tracker can manage a variable number of people over extended time periods.

Wang, Zhongdao, et al. [22] propose a system which combines the detection model, used to isolate targets and the appearance embedding model used to link detections between frames, into one framework. The model can detect many objects at once and is faster than current methods bounding box regression, and anchor classification, and where the separate losses are automatically weighted.

Yu, En, et al. [23] propose a module which disentangles the learnt representation to give embeddings unique to detection and ReID. The implicit method provided by this module helps to balance the requirements of these subtasks. MOT techniques often use local information to link the identified objects together without taking into account the global semantic relationship. We employ a module called the Guided Transformer Encoder to get around this limitation. It combines the robust reasoning capacity of the Transformer Encoder with deformable attention.

Chen, Xiaotong, Seyed Mehdi Iranmanesh, and Kuo-Chin Lien. [24] propose the model PatchTrack, a combined tracking and detection system based on Transformers that can predict tracks using patches from the current frame of interest. The Kalman filter is used to forecast the locations of tracks that are already visible in the current frame. Patches that have been removed from the anticipated bounding boxes are sent to the Transformer decoder in order to infer new tracks. The proposed approach leverages object motion and object appearance information encoded in patches to concentrate on the regions where new tracks are most likely to develop.

Zhang, Yifu, et al. [25] propose a simple method called BYTE. The model detects objects and generates bounding boxes. These boxes are classified into two categories, high score and low score. The model first connects the tracklets to the high score boxes, if they do not match, the low score boxes are linked to the objects. To predict new locations, Kalman filters are used.

Cai, Jiarui, et al. [26] propose a model which takes sequences of frames as input and generates trajectories of the tracked objects. The model uses a memory which holds the tracked states of different objects. It uses a transformer encoder to convert the frames to embeddings which are fed to the decoder while simultaneously stored in memory. The output from the decoder along with the embeddings from the memory buffer are fused together to generate bounding boxes.

Meinhardt, Tim, et al. [27] propose TrackFormer, which uses attention to track objects. The model uses CNN to extract features from the input images which are fed to the transformer encoder. The encoder makes use of attention to focus on important features and generate bounding boxes. The decoder generates new tracks and also follows existing tracks using track queries.

Zhang, Yifu, et al. [28] propose FairMOT, which suggests a single framework which consists of a combination of multi-task learning of detection and re-ID. This approach allows joint optimization of the two tasks. It achieves great accuracy for both tasks and outperforms state-of-the-art methods. Multi-task learning is a technique that allows a single model to learn several tasks at once, sharing information between them to improve performance.

Peng, Jinlong, et al. [29] introduces a new multiple-object tracking (MOT) model called Chained-Tracker (CTracker), which is an end-to-end solution that integrates all three subtasks of detection of objects, extraction of features, and association of data into a single framework. Secondly, it also provides an open source code for the model, making it freely available for anyone to use and improve. This could allow for more efficient and accurate tracking systems to be developed, and facilitate further research into MOT. Finally, the paired attentive regression methodology used in this model could be applied to other computer vision tasks, potentially leading to more efficient and effective modelFinally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar.

IV. DATASETS

A. UAVDT

UAVDT is a massive Detection and Tracking dataset for detection, Single Object Tracking and Multi-Object Tracking.

The dataset contains about 80,000 frames extracted from 10 hours of UAV footage. The videos are recorded in various complex scenarios like in the rain, fog, night, etc. The main objects of interest in this dataset are vehicles where each frame is annotated with bounding boxes around the vehicles and are described with various attributes like weather condition, occlusion, elevation, vehicle type. etc. The videos are recorded with image resolution of 1080x540 pixels at 30 fps.

ISSN No:-2456-2165

B. VisDrone2019

VisDrone2019 is a sizable benchmark dataset that includes ground-truth that has been annotated for a variety of significant computer vision applications. Over 280 video segments totaling 261,908 frames and 10,209 static pictures make up the benchmark dataset. The dataset includes information on location, environment, items, and density, among other things. The dataset was gathered utilizing a variety of drone platforms in varied contexts, with varying levels of weather and illumination. These frames have bounding boxes carefully marked with targets of common interest, including pedestrians, vehicles, bicycles, and tricycles. The VisDrone2019 benchmark is used for single and multiple object tracking in pictures and videos.

C. UA-DETRAC

This is a difficult real-world benchmark for multi-object detection and tracking. The dataset is made up of 10 hours' worth of videos that were recorded in 24 various places in cities in China. The resolution of the videos are 960 x 540 pixels in size and are shot at 25 fps. The UA-DETRAC collection contains more than 140 thousand frames and 8250 manually annotated cars, totaling over a million labeled bounding boxes.

D. MOT 16/17

The MOT16/17 standards are often used in MOT to detect and track pedestrians. The fourteen sequences that make up MOT16 encompass a range of situations, vantage positions, camera angles, and weather conditions. In MOT16, 7 image sequences are used for training and 7 for validation. MOT17 was reconstructed from MOT16 and improved on. In comparison to MOT16, MOT17 offers more accurate ground truth and more bounding boxes for detection from a variety of detectors, including DPM, SDP, and Faster RCNN.

E. DOTA

Over 1.7 million object instances from 18 types of oriented-bounding-box annotations were collected for the planned DOTA dataset using more than 11,000 aerial photographs. Baselines encompassing 10 algorithms with more than 70 configurations are generated based on this large and well-annotated dataset. A comparison has been done on the speed and accuracy of each model and shown in the paper published by Ding, Jian, et al. [30]

V. METRICS USED

A. MOTA (Multiple Object Tracking Accuracy)

MOTA is the most popular measure for Multiple Object Tracking that is most comparable with how human vision works. Amapping is done between the predicted detected objects and ground truth detected objectsacross every frame if they are similar to compute the contingency table values. Identity Switch (IDSW), which happens when a tracker mistakenly switches the identities of the objects or when a tracker loses its target and then re-initialized with a different object, is used to quantify the relationship. Three types of tracking mistakes are measured by MOTA: ID Switch, False Positive, and False Negative.

$$MOTA = 1 - \frac{|FN| + |FP| + |IDSW|}{|gtDet|}$$

B. MOTP (Multiple Object Tracking Precision)

MOTP is a metric used for evaluating precision in MOT algorithms. It is defined as the mean distance between the estimated object location and the inferred object location. In simpler terms it is a measure of how well the exact positions of objects are estimated. As MOTP primarily measures the detector's localization accuracy, it doesn't offer much about the tracker's actual performance.

$$MOTP = \frac{1}{|TP|}\sum_{TP} \quad S$$

C. IDF1 (Identification Metrics)

IDF1 is utilized as a supplementary statistic on the MOTChallenge benchmark since it focuses on quantifying association accuracy rather than detection accuracy.. It places more emphasis on association accuracy than detection accuracy. IDF1 determines the presence of trajectories by the computation of ground truth trajectories and predicted trajectories which are mapped using a bijective function. The proportion of correctly identified detections to the average number of computed detections and ground-truth detections is measured by the IDF1 ratio. Instead of providing information regarding effective detection or association, the overall number of distinct items in a scene is shown by the IDF1 score being high. Additionally, it doesn't assess how well trackers perform localization.

$$IDF1 = \frac{|IDTP|}{|IDTP| + 0.5\,|IDFN| + 0.5\,|IDFP|}$$

VI. CONCLUSION

Multiple Object Tracking is being researched around the world by many institutions and organizations. This paper offers a comprehensive view of the recent advancements in the field consisting of the various techniques which include those that have produced state-of-the-art results. In particular, this paper pays more attention to techniques that tackle issues with relation to object tracking in aerial image sequences - with new architectures like attention models. The various techniques and algorithms are compared with their pros and cons and are tabulated. Finally, this paper goes through common datasets and benchmarks used, and potential strategies to tackle the MOT challenges.

REFERENCES

[1.] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

[2.] Beheim, Tsuyoshi. (2021). Multi-Vehicle Detection and Tracking in Aerial Imagery Sequences using Deep Learning Algorithms.

[3.] Jetley, S., Lord, N. A., Lee, N., & Torr, P. H. (2018). Learn to pay attention. *arXiv preprint arXiv:1804.02391*.

[4.] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213-229). Springer, Cham.

[5.] Cai, Y., Du, D., Zhang, L., Wen, L., Wang, W., Wu, Y., & Lyu, S. (2019). Guided attention network for object detection and counting on drones. *arXiv preprint arXiv:1909.11307*.

[6.] Perreault, H., Bilodeau, G. A., Saunier, N., & Héritier, M. (2020, May). Spotnet: Self-attention multi-task network for object detection. In *2020 17th Conference on Computer and Robot Vision (CRV)* (pp. 230-237). IEEE.

[7.] Jadhav, A., Mukherjee, P., Kaushik, V., & Lall, B. (2020, February). Aerial multi-object tracking by detection using deep association networks. In *2020 National Conference on Communications (NCC)* (pp. 1-6). IEEE.

[8.] Bahmanyar, R., Azimi, S. M., and Reinartz, P.: MULTIPLE VEHICLES AND PEOPLE TRACKING IN AERIAL IMAGERY USING STACK OF MICRO SINGLE-OBJECT-TRACKING CNNS, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-4/W18, 163–170, https://doi.org/10.5194/isprs-archives-XLII-4-W18-163-2019, 2019.

[9.] Xue, X., Li, Y., Yin, X., & Shen, Q. (2021). DCF-ASN: Coarse-to-fine Real-time Visual Tracking via Discriminative Correlation Filter and Attentional Siamese Network. *arXiv preprint arXiv:2103.10607*.

[10.] Tang, Z., Liu, X., & Yang, B. (2020, December). PENet: Object detection using points estimation in high definition aerial images. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 392-398). IEEE.

[11.] Malagi, Vindhya & Babu, Ramesh & Rangarajan, Krishnan. (2016). Multi-object Tracking in Aerial Image Sequences using Aerial Tracking Learning and Detection Algorithm. Defence Science Journal. 66. 122. 10.14429/dsj.66.8972.

[12.] Azimi, S. M., Kraus, M., Bahmanyar, R., & Reinartz, P. (2020). Multiple pedestrians and vehicles tracking in aerial imagery: A comprehensive study. *arXiv preprint arXiv:2010.09689*.

[13.] Lin Y, Wang M, Chen W, Gao W, Li L, Liu Y. Multiple Object Tracking of Drone Videos by a Temporal-Association Network with Separated-Tasks Structure. *Remote Sensing*. 2022; 14(16):3862.https://doi.org/10.3390/rs14163862

[14.] Koyun, O. C., Keser, R. K., Akkaya, İ. B., & Töreyin, B. U. (2022). Focus-and-Detect: A small object detection framework for aerial images. *Signal Processing: Image Communication*, *104*, 116675.

[15.] 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).

[16.] Sun, P., Cao, J., Jiang, Y., Zhang, R., Xie, E., Yuan, Z., ... & Luo, P. (2020). Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*.

[17.] Zhu, J., Yang, H., Liu, N., Kim, M., Zhang, W., & Yang, M. H. (2018). Online multi-object tracking with dual matching attention networks. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 366-382).

[18.] Chen, L., Ai, H., Zhuang, Z., & Shang, C. (2018, July). Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *2018 IEEE international conference on multimedia and expo (ICME)* (pp. 1-6). IEEE.

[19.] Chen, B., Deng, W., & Hu, J. (2019). Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 371-381).

[20.] Xu, Y., Osep, A., Ban, Y., Horaud, R., Leal-Taixé, L., & Alameda-Pineda, X. (2020). How to train your deep multi-object tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6787-6796).

[21.] Ban, Y., Ba, S., Alameda-Pineda, X., Horaud, R. (2016). Tracking Multiple Persons Based on a Variational Bayesian Model. In: Hua, G., Jégou, H. (eds) Computer Vision – ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science(), vol 9914. Springer, Cham.https://doi.org/10.1007/978-3-319-48881-3_5

[22.] Wang, Z., Zheng, L., Liu, Y., Li, Y., & Wang, S. (2020, August). Towards real-time multi-object tracking. In *European Conference on Computer Vision* (pp. 107-122). Springer, Cham.

[23.] Yu, E., Li, Z., Han, S., & Wang, H. (2022). Relationtrack: Relation-aware multiple object tracking with decoupled representation. *IEEE Transactions on Multimedia*.

[24.] Chen, X., Iranmanesh, S. M., & Lien, K. C. (2022). PatchTrack: Multiple Object Tracking Using Frame Patches. *arXiv preprint arXiv:2201.00080*.

[25.] Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., ... & Wang, X. (2022). Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision* (pp. 1-21). Springer, Cham.

[26.] Cai, J., Xu, M., Li, W., Xiong, Y., Xia, W., Tu, Z., & Soatto, S. (2022). MeMOT: Multi-Object Tracking with Memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8090-8100).

[27.] Meinhardt, T., Kirillov, A., Leal-Taixe, L., & Feichtenhofer, C. (2022). Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8844-8854).

[28.] Peng, J., Wang, C., Wan, F., Wu, Y., Wang, Y., Tai, Y., ... & Fu, Y. (2020, August). Chained-tracker: Chaining paired attentive regression results for end-

to-end joint multiple-object detection and tracking. In *European conference on computer vision* (pp. 145-161). Springer, Cham.

[29.] Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2021). Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, *129*(11), 3069-3087.

[30.] Ding, Jian, et al. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges.