An Efficient and Robust Multi-Object Recognition and Tracking Algorithm using Mask Region based Convolution Neural Network (R-CNN)

A. Nirmala, S.Arivalagan, R.Arunkumar

Abstract*: *Presently, Multi-Object tracking (MOT) is mainly applied for predicting the positions of many predefined objects across many successive frames with the provided ground truth position of the target in the first frame. The area of MOT gains more interest in the area of computer vision because of its applicability in various fields. Many works have been presented in recent years that intended to design a MOT algorithm with maximum accuracy and robustness. In this paper, we introduce an efficient as well as robust MOT algorithm using Mask R-CNN. The usage of Mask R-CNN effectively identifies the objects present in the image while concurrently creating a high-quality segmentation mask for every instance. The presented MOT algorithm is validated using three benchmark dataset and the results are extensive simulation. The presented tracking algorithm shows its efficiency to track multiple objects precisely.

Index Terms*: *Computer vision, Mask R-CNN, MOT, Recognition.

I. INTRODUCTION

Multi-Object tracking (MOT) offers extensive applications ranges such as behavior and pose analysis, medical image analysis, and video surveillance [1]. MOT provides other difficulties like data association when comparing with the single-object tracking that aims to relate the detections over the frames to safe the identities of the targets. For upgrading the tracking model in the frame by frame, the prior MOT attempts focus over the methods of Expectation-Maximization (EM). Those methods are error prone such as drifts which it is complex to recover. Tracking via detection methods had been projected to avoid these issues. The main focus is to execute the object detectors autonomously over every frame of input batch and then test the association of data to connect the detections into whole tracks. The association of data is designed frequently as Network Flow Programming or Conditional Random Field where every detection indicates to a graph model node. In the detection process, one of the leading tracking difficulties is computational complexity in which the phase of data association needs combinatorial optimization over a large node or detection counts that hides the real-time performance. Before executing data association, numerous methods initially group unambiguous detections in subsequent frame into a cluster to deal with this issue. Those are known as tracklets. It is considered that entire detections in tracklet distribute the similar label which it depends on the similar goal.

In a similar manner like detections, most of the previous MOT methods treat the tracklets [2]. Tracklets are considered as independent in the phase of data association excluding two physical conditions like one target should not claim higher than one tracklet, in particular, instant and every tracklet should get claimed by one target. It excludes the high informative and composite interplays like higher-order motion among the tracklets mainly for non-associable. A new MOT graph is defined where every node in the graph depends on the tracklets pair at this end. The tracklet that occur at a particular instant is known as front tracklet and the another one is known as back tracklet in every node.

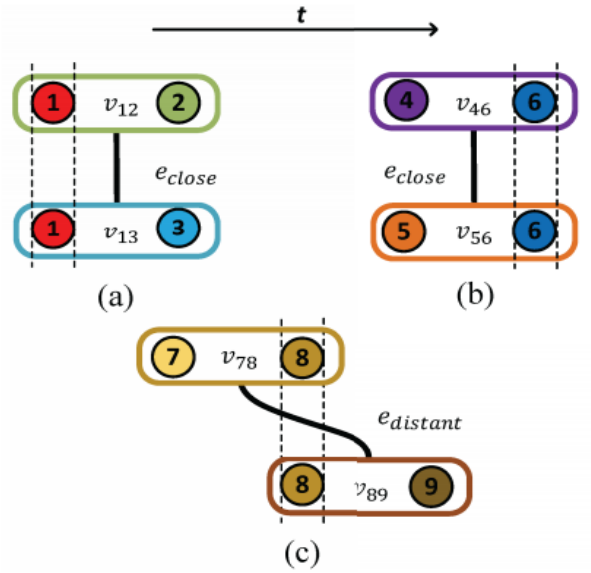

Fig. 1 A sample MOT graph.

Denoting it is chosen as a target trajectory part, every node might allocate to a binary label. With every node by edges, graph nodes might interrelate. Figure 1 demonstrates the three edge kinds; two nodes distribute the similar back tracklets, two nodes that the front tracklet of one is the back tracklet and two nodes that distribute similar front tracklet. The primary two edge types depend on nearby interaction whereas the third type depends on the distant interaction.

Revised Manuscript Received on July 05, 2019

A. Nirmala, Research Scholar, Annamalai University, Chidambaram, India.
Dr. S. Arivalagan, Assistant Professor, Annamalai University, Chidambaram, India.
Dr. R. Arunkumar, Assistant Professor, Annamalai University, Chidambaram, India.

Every individual color circle indicates a tracklet, and every eclipse including two tracklets represents a node in the graph. An edge is used to share the identical tracklet. Intrinsic associations are used among tracklets by doing this even if it might not be associable directly. Moreover, the higher-order motion appearance and smoothness consistency among non-associable tracklets are used to enhance the data association in the distant interaction case such as Tracklets 9 and 7. For first-order appearances and motions, this is in comparison with the traditional network flow models. We use conditions like it does not co-exist in close interaction case such as Tracklets 3 and 2 in one trajectory. Through considering one tracklet as a negative sample and other is a positive sample, a local appearance-based classifier is trained. Over the third tracklet, we use learning a classifier that is connectable to both, for every node to learn a superior confidence score in our graph that depends on the associable tracklets pair.

To manipulate the close interaction, we use Tracklets 1, 2 and 3 as an instance. V12 and v13 are two nodes that use unary potential. Therefore, Tracklets 3 and 2 are equally exclusive, train a local classifier by employing the detections derived from the two tracklets and Tracklet 3 as negative 2 as positive samples. Over Tracklet 1, we imply local learned classifier; it might be associated with one of Tracklets 3 and 2. The gained Tracklet 1 classification score revises the linking Tracklet 2 to 1 and Tracklet 1 to 3 in the same way v_{12} and v_{13} unary potential. Due to the function of global data association, it's confidence revision is critical and frequently fails while the tracklets are near to every other, mainly while it occlude in crowded scenes every other.

The MOT issues become a binary labeling one with those notions which the target is to allocate a 0 or 1 label for every node or an associable tracklets pair. In this paper, we introduce an efficient as well as robust MOT algorithm using Mask R-CNN. With identical primary stage, Mask R-CNN uses a similar two-step process. Along with box offset and class prediction, Mask R-CNN also gives a binary mask for every RoI in the second phase. The usage of Mask R-CNN effectively identifies the objects present in the image while concurrently creating a high-quality segmentation mask for every instance. The presented method is validated using three benchmark dataset, and the results are extensively simulation. The presented MOT algorithm shows its efficiency to track multiple objects precisely.

The remaining portions of the paper are planned here. Section 2 discusses the works related to the MOT algorithm. The presented MOT model is given in Section 3, simulated in Section 4 and concluded in Section 5.

II. RELATED WORKS

This section reviews some of the works that are related to the presented work. We describe the tracking techniques that are intended for interactions that manage the tracklet interactions explicitly. Moreover, the approaches of graph-based data association are also discussed.

A. Tracking with Interaction Modeling

We are aiming at one target at a particular instant in case of single-object tracking. We need to follow multi-targets in MOT at a time and protect the identities. The target should not as independent of every other. In one way or other, it can communicate with every other. There exist many methods at present years [3-7] which makes communications over multi-targets. The targets interaction features are learned by [4] and exhibit a superior prediction of motion. To discover concurrent and overlapping interactions, [3] inserts pairwise costs. To examine the target's potential locations, [7] projected a Relative Motion Network (RMN) by employing other communicating targets. The disappearing and appearing relationships among objects are observed and projected to examine the different interactions among various object types like car and pedestrians. To examine the interactions, [5] used human behaviors in people tracking scenes whereas the interacting model is modeled clearly for learning the people dynamics in real-time scenarios. In MOT, the peer-to-peer model has been projected by [6] to learn the composite interactions.

There exist a few existing works that manage the interactions among tracklets. The detections are linked by tracklets and used the interactions among the next ones by [7]. Through reducing the highly overlapped tracks, [8] manage the interactions in trajectory-level as it probably depends on a similar target. [9] and [10] addressed the interactions among the two overlapped or close tracklets; however, it ignores temporally the interactions as isolated tracklets. The trajectories smoothness is measured through [2]. However, the method complexity had constraints the capability to discover the close interactions. A novel tracklet confidence measure has been projected [11] by employing continuity and detect ability, and deep networks can be employed for learning the discriminative appearance model depending on that the tracklet association is done.

B. Tracking Graph

In MOT, the association of data is the main issue. The data association can be designed in the framework of tracking-by-detection as a combinatorial optimization issue over graphs. Because of the efficiency and simplicity, numerous MOT methods use the Hungarian algorithm. Through estimating the potential one among every node pairs, the Hungarian algorithm search for an optimal one-to-one mapping that might vary in cardinality. For linking the detection from two frames, it is appropriate, however, not from extensive temporal range. Network flow programming is the other extensively employed technique of data association.

In a vast spatio-temporal graph, these technique searches for the associating detections from a frame batch. Through considering every grid cell that can be engaged through one individual, [9] estimates the minimum-cost flows over discretized ground plane especially. Through using two flow variables sets, [12] uses the traditional network flow in the graph, each one of that respect to one target kind and follows the two target kinds at the same time. Many combinatorial optimization methods had also been discovered. [13] uses data association as a graph decomposition problem. [14] depends on the maximum-weight independent set. At the same time, for designing longer-term appearance and to resolve the problem of reduced cost lifted multicut optimization, lifted edges were introduced [15]. Through assuming the node as edge and hypothesis as relative among hypotheses, the entire techniques build a data association graph.

[16] also depends on the problem of 0-1 optimization. However, it aims at frame-by-frame or online tracking where the projected method focuses over batch processing whereas we discover the interaction among both non-associable and associable tracklets. Because of the online manner, [16] is greedy and it does not care for motions in higher-order. Through the model of distant tracklet, a frames batch is used and use higher-order motion explicitly to enhance the end trajectories of many objects that are advantageous. Other techniques aim at high order target motions. To develop MOT dense structure, [2] modeled a hypergraph to use the tracklets as graph nodes. To manage the targets' potential occlusions, online Conditional Random Field is used when the [17] designed a network flow program that is efficient with higher-order smoothness. The presented method addresses both distant and close interactions when comparing with the above mentioned ones at the same time.

III. PROPOSED MASK R-CNN BASED MOT ALGORITHM

A. Overview

The presented MOT algorithm uses Mask R-CNN to identify the existence of target objects precisely and robustly.The process involved in the presented MOT model is shown in Fig. 2.Initially, the given video input is segregated into a set of frames. Then, the feature extraction process takes place intending to improve the results of the tracking process. The feature extraction process will assign class labels to every object along with its structured mask. Then, Mask-RCNN is applied to train the model. Once the model is trained, new test video sequences can be applied for tracking the presence of multiple objects in the frame. Similar to the training process, the video sequences will be segregated to a set of frames before applying it to the test model. Once the frames are tested, a mask will be produced at every object present in the frame along with its given label.

B. Faster R-CNN

The Faster R-CNN comprises of two phases. The initial phase is known as Region Proposal Network (RPN), projects a candidate object bounding boxes. The next is essence Fast R-CNN [18] that derives features by employing RoIPool from every candidate box and carry out bounding-box regression and classification. For rapid inference, the features of both phases can be distributed.

C. Mask R-CNN

Mask R-CNN [19] is theoretically simple: For every candidate object, Faster R-CNN comprises of two outputs such as bounding-box offset and a class label. Then, a subsequent branch is inserted that gives the object mask. Mask R-CNN is an intuitive and natural concept. From the box and class outputs, the extra mask output is unique that needs delicate spatial object layout extraction. We introduce subsequently the main Mask R-CNN elements that involve pixel-to-pixel alignment, which is missing in the Fast/Faster R-CNN missing piece as shown in Fig. 3.

With identical primary stage, Mask R-CNN uses a similar two-step process. Along with box offset and class prediction, Mask R-CNN gives a binary mask for every RoI in the second phase. This is very much contrasting with the modern system wherever the classifications are based on mask predictions. The method follows the Fast R-CNN motivation which gives regression and bounding-box classification at the same time.

We describe a multi-task loss while training over every sample RoI $L = L_{cls} + L_{box} + L_{mask}$. L_{box} is the bounding-box and L_{cls} is classification loss that are identical. For every RoI, the mask branch comprises of Km^2dimensional output that the encode K binary masks of m × m resolution for every K class. We employ per-pixel sigmoid and describes average binary cross-entropy loss as L_{mask}. L_{mask} is described over k-th mask for RoI given ground-truth class k.

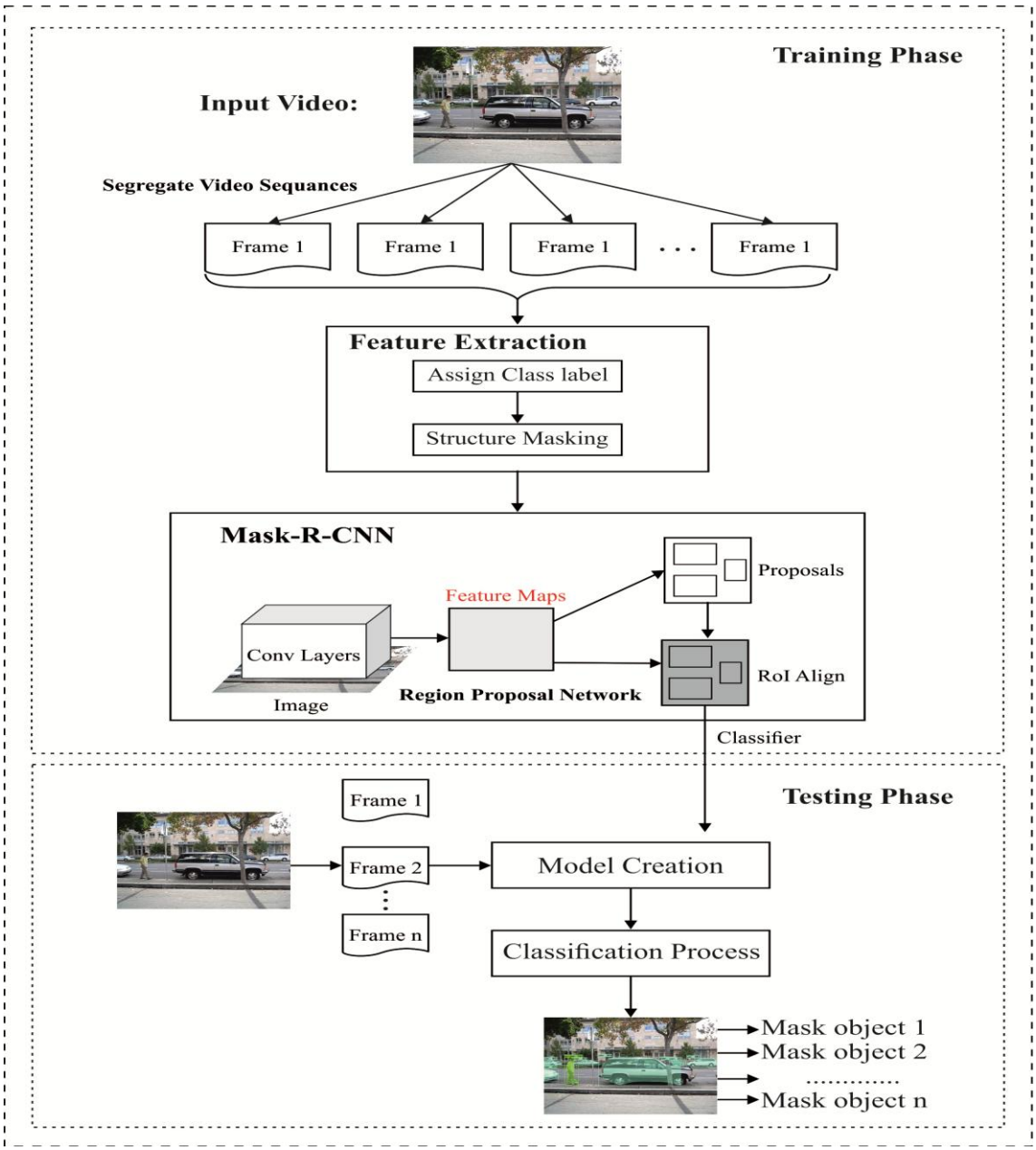

Fig. 2 Overall process of the proposed MOT algorithm

With competing over the classes, the L_{mask} definition enables the network to produce masks for each class. For class label prediction, we depend on the dedicated classification branch employed to choose the output mask. This decouples class and mask prediction. While using FCNs [20] towards semantic segmentation, this varies from the usual doings that especially employs a multinomial cross-entropy loss and per-pixel softmax. The masks over classes competes each for other with a binary loss and per-pixel sigmoid. We demonstrate that this is a main thing for superior segmentation results.

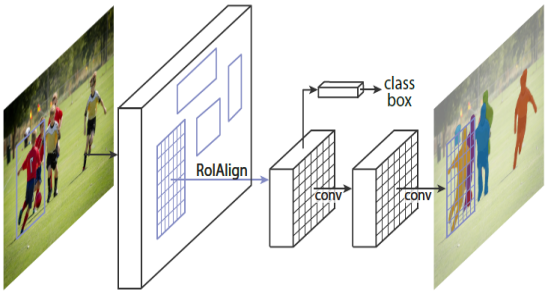

Fig. 3 Mask R-CNN

IV. PERFORMANCE VALIDATION

To validate the performance of the presented Mask-RCNN based MOT model, it is tested against a set of three benchmark image dataset and the results are validated under different aspects. In the following sections, dataset employed and the attained results are discussed in detail.

A. Dataset used

For comparing the results of the presented method, a benchmark video is used [21]. The employed bird dataset holds a set of 99 frames from the 3s second.

B. Results analysis

Fig. 4 illustrates the multiple objects detected by the presented MOT algorithm on the sample of 12 frames from the Bird dataset. The figure indicates that the Mask R-CNN efficiently identifies the existence of multiple objects in the frame. Also, it is apparent that the mask structure is provided to each object present in the frame. Additionally, the detection rate of every individual object identified in the image is also given.

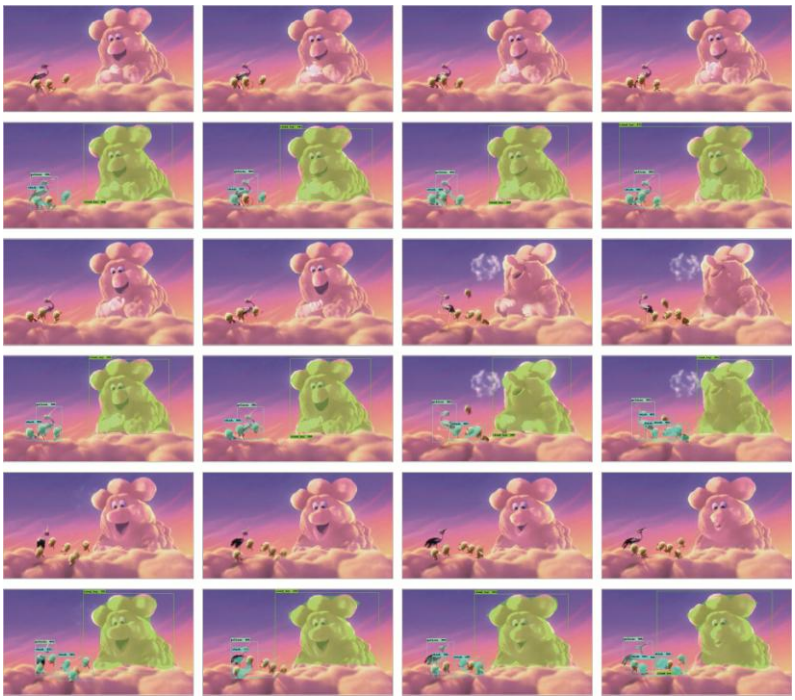

Fig.4 MOT results on the applied Bird dataset(odd row-original frame, even row-detected frame)

Table 1. Accuracy of Objects per frame on Bird Dataset

Frame Number	O-1	O-2	O-3	O-4	O-5
07	99	99	99	-	-
09	99	99	99	-	-
10	99	99	99	-	-
11	99	99	97	-	-
12	99	99	99	-	-
13	99	99	99	-	-
83	99	96	99	-	-
84	99	85	96	99	-
94	99	92	96	99	-
95	98	99	99	-	-
96	99	79	99	99	-
97	99	93	95	99	99

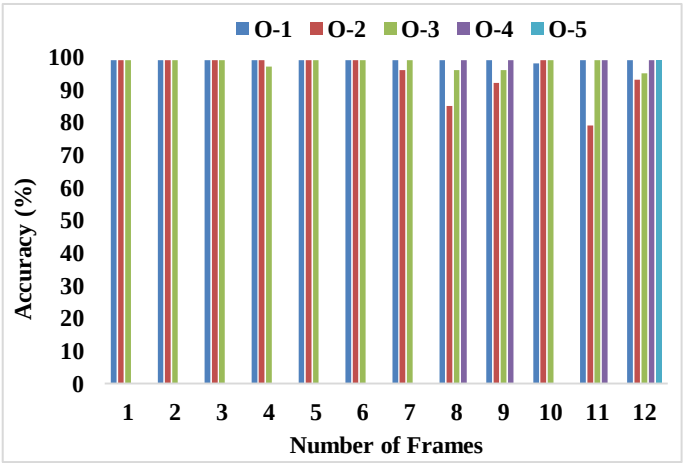

Fig. 5 Accuracy of Objects for Bird Dataset

Table 2. Comparative Accuracy of Objects per frame on Bird Dataset

Frame Number	O-1	O-2	O-3	O-4	O-5
07	99	99	99	-	-
09	99	99	99	-	-
10	99	99	99	-	-
11	99	99	97	-	-
12	99	99	99	-	-
13	99	99	99	-	-
83	99	96	99	-	-
84	99	85	96	99	-
94	99	92	96	99	-
95	98	99	99	-	-
96	99	79	99	99	-
97	99	93	95	99	99

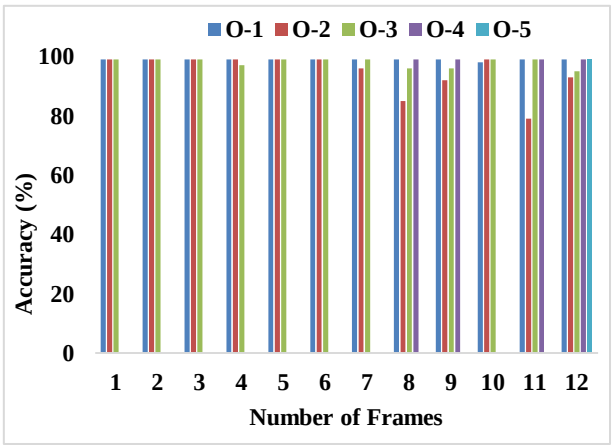

Fig. 6 Comparative Accuracy of Objects per frame on Bird Dataset

Table 1 and Fig. 5 the detected accuracy rate of all the objects identified in the applied frames. Since the frames in the dataset contain a minimum of three and maximum of five objects, the accuracy rates of all the objects tracked in the corresponding frame are given. For instance, for the applied frame number 07, the presented MOT algorithm precisely identifies multiple objects O-1, O-2 and O-3 with the detection rate of 99%. Likewise, in the frame numbers 9 and 10, maximum tracking accuracy of 99% is attained for all the objects presented in the frame. Similarly, on the applied frame number 97, the five objects O-1, O-2, O-3, O-4 and O-5 are tracked with the accuracy rate of 99%, 93%, 95%, 99% and 99%. In line with, the presented MOT algorithm correctly identifies the presence of all objects in the applied frames. A comparison of different methods against the presented MOT algorithm are employed to different frames and the achieved performance are presented in Table 2 and Fig. 6.

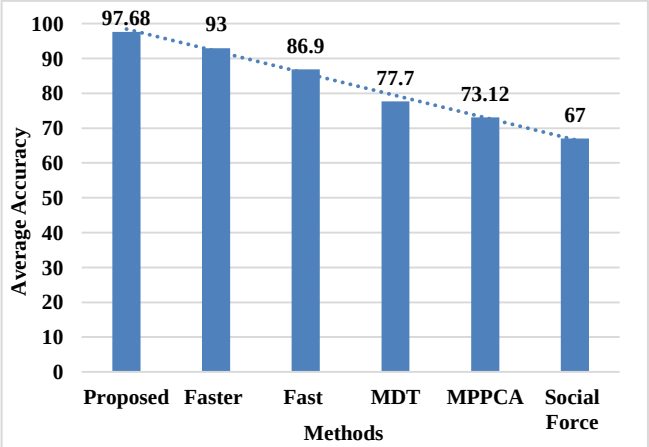

Fig. 7 Comparative Average Accuracy of Objects per frame on Bird Dataset

As shown in the table, for the applied frame number 07, it is observed that the SF obtains lower accuracy of 65.27 whereas MPPCA, MDT, Fast R-CNN and Faster R-CNN obtains the accuracy of 70.28, 70.35, 89.02 and 94.34 respectively. However, the presented model shows maximum performance with higher accuracy of 99%. For the applied frames 9 and 10, the presented MOT model shows higher accuracy of 99 and 99 respectively. Also, the SF exhibited poor performance with minimum accuracy values of 62.10 and 63.28 respectively. On continuing, the proposed MOT algorithm attained maximum results with an enhanced accuracy rate on the employed frames.

Figure 7 provides the average accuracy results obtained by various approaches on the employed test frames. The values shown in the figure indicated that maximum performance with a higher accuracy of 97.68. Also, SF depicted poor performance with lower accuracy of 67. At the same time, the MPPCA tries to handle well with lower accuracy of 73.12 and MDT exhibits sufficient accuracy of 77.7. The Fast R-CNN and Faster R-CNN model obtains an average accuracy of 86.9 and 93 respectively. However, the presented model shows maximum accuracy of 97.68 over the compared methods.

V. CONCLUSION

In this paper, we have introduced an efficient as well as robust MOT algorithm using Mask R-CNN. The usage of Mask R-CNN effectively identifies the objects present in the image while concurrently creating a high-quality segmentation mask for every instance. The presented method is validated using three benchmark dataset and the results are extensively simulation. The presented tracking algorithm shows its efficiency to track multiple objects precisely. The presented model shows a maximum average accuracy of 97.68 over the compared methods. In the future, the presented model can be implemented in real time applications.

REFERENCES

1. V. Belagiannis et al., "Parsing human skeletons in an operating room," Mach. Vision Appl., 2016, vol. 27, no. 7, pp. 1035–1046.
2. L. Wen, Z. Lei, S. Lyu, S. Z. Li, and M.-H. Yang, "Exploiting hierarchical dense structures on hypergraphs for multi-object tracking," IEEE Trans. Pattern Anal. Mach. Intell., Oct. 2016, vol. 38, no. 10, pp. 1983–1996.
3. V. Chari, S. Lacoste-Julien, I. Laptev, and J. Sivic, "On pairwise costs for network flow multi-object tracking," in Proc. Conf. Comput. Vis. Pattern Recognit., 2015, pp. 5537–5545.
4. L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese, "Learning an image-based motion context for multiple people tracking," in Proc. Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 3542–3549
5. S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in Proc. Conf. Comput. Vis. Pattern Recognit., Sep./Oct. 2009, pp. 261–268.
6. A. Sadeghian, A. Alahi, and S. Savarese. (2017). "Tracking the untrackable: Learning to track multiple cues with long-term dependencies." [Online]. Available: https://arxiv.org/abs/1701.01909
7. J. H. Yoon, M.-H. Yang, J. Lim, and K.-J. Yoon, "Bayesian multi-object tracking using motion context from multiple objects," in Proc. WACV, 2015, pp. 33–40.
8. B. Yang and R. Nevatia, "An online learned CRF model for multi-target tracking," in Proc. Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 2034–2041.
9. A. Milan, K. Schindler, and S. Roth, "Detection- and trajectory-level exclusion in multiple object tracking," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2013, pp. 3682–3689.
10. A. Milan, K. Schindler, and S. Roth, "Multi-target tracking by discretecontinuous energy minimization," IEEE Trans. Pattern Anal. Mach. Intell., , Oct. 2016, vol. 38, no. 10, pp. 2054–2068.
11. S.-H. Bae and K.-J. Yoon, "Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 3, pp. 595–610, Mar. 2017
12. X. Wang, E. Türetken, F. Fleuret, and P. Fua, "Tracking interacting objects optimally using integer programming," in Proc. Eur. Conf. Comput. Vis., 2014, pp. 17–32.
13. A. Milan, L. Leal-Taixé, I. D. Reid, S. Roth, and K. Schindler. (2016). "MOT16: A benchmark for multi-object tracking." [Online]. Available: https://arxiv.org/abs/1603.00831
14. W. Brendel, M. Amer, and S. Todorovic, "Multiobject tracking as maximum weight independent set," in Proc. IEEE Int. Conf. Comput. Vis., Jun. 2011, pp. 1273–1280
15. S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 3539–3548
16. L. Lan, D. Tao, C. Gong, N. Guan, and Z. Luo, "Online multi-object tracking by quadratic pseudo-Boolean optimization," in Proc. IJCAI, 2016, pp. 3396–3402.
17. A. A. Butt and R. T. Collins, "Multi-target tracking by Lagrangian relaxation to min-cost network flow," in Proc. Conf. Comput. Vis. Pattern Recognit., Jun. 2013, pp. 1846–1853.
18. R. Girshick. Fast R-CNN. In ICCV, 2015
19. He, K., Gkioxari, G., Dollár, P. and Girshick, R., 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961-2969).
20. J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015
21. Dataset http://cvlab.hanyang.ac.kr/tracker_benchmark/datasets.html