Literature Review Report – Group 10

Title: Improving Multi-Object Tracking Using Depth - A Comprehensive Literature Review

Authors: [Allan Cheboiwo #73661290], [Tarek Alkabbani #84930239], [Haoyu Wang #42343871], [Vanessa Laurel Hariyanto #72484546]

Course: COSC 444/544: Computer Vision

Date: February 24, 2024

---

1. Abstract

Multiple Object Tracking (MOT), which enables machines to interpret visual input, is a significant application of computer vision. Despite being widely employed in surveillance, self-driving automobiles, and traffic monitoring, MOT has limitations such as occlusion which happens when objects obstruct one another and ID switching. While deep learning models provide excellent accuracy but need substantial processing resources, traditional tracking techniques like the Kalman Filter are still useful because of their effectiveness.

This method focuses on tracking pedestrians, a challenging MOT problem because of the unpredictability of human mobility. We evaluate the advantages and disadvantages of the Kalman Filter and suggest a better method that makes use of depth information to lessen occlusion problems. Our research seeks to optimize MOT for practical uses, particularly in object identification, by striking a compromise between accuracy and efficiency.

2. Introduction

Computer vision is an interdisciplinary scientific field focused on enabling computers to automatically extract meaningful information from images or videos to understand or represent the real world. It aims to develop systems that can achieve high-level visual understanding, similar to human perception [1][2][3]. Computer vision is crucial because it enables machines to understand and analyze visual data, thereby facilitating advancements in fields such as healthcare, autonomous driving, and industrial automation.

Multiple Object Tracking (MOT) plays an important role in computer vision. The task of MOT is largely partitioned into locating multiple objects, maintaining their identities, and yielding their individual trajectories given an input video [4]. By tracking multiple objects in the video accurately, MOT can enhance the ability to analyze movement patterns, improve safety in transportation systems, optimize traffic flow management, and strengthen security monitoring. These abilities make it a foundation in a variety of real-world scenarios that require dynamic object analysis.

When designing MOT methods, two key aspects must be considered: 1) determining an effective way to measure the similarity between objects within frames and 2) developing a method to restore identity information by utilizing similarity measurement algorithms across consecutive frames [5]. This introduces two significant challenges of MOT, which are occlusion and ID switching. Occlusion occurs when one object partially or fully obstructs another in the same video frame, making it difficult for the tracker to maintain a continuous trajectory, especially when using only camera-based tracking without additional sensor data. ID switching happens when the tracker incorrectly assigns a new identity to an object that should retain its original ID in the entire video, often due to temporary disappearance from the frame or inaccurate motion prediction. These issues are critical in practical applications such as monitoring and autonomous driving, as these scenarios require robust solutions to ensure robust tracking performance [5].

Although deep learning methods (such as convolutional neural network (CNN), Long Short-Term Memory (LSTM), and transformer) have shown outstanding performance in MOT tasks [6], their drawbacks like high computational cost, reliance on large-scale datasets, and slow real-time inference speed make traditional methods such as Kalman filtering still valuable for research. For example, in resource-constrained environments, such as embedded systems and edge computing devices, these methods can provide stable tracking performance with significantly lower computational overhead. Therefore, gaining a thorough understanding of these classical methods is crucial for recognizing their limitations and guiding the development of more accurate and efficient tracking solutions.

In this review, we mainly focus on research related to pedestrian tracking. There are three key reasons for choosing this application scenario. First, unlike many other objects around us, pedestrians are non-rigid and dynamic, which makes them a representative application scenario for studying MOT problems with many aspects that can be studied and improved [4]. Second, pedestrian videos are widely used in the real world, which makes them of great commercial value and practical significance. Third, pedestrian video datasets are relatively easy to obtain, which provides rich data for the development and evaluation of tracking algorithms.

Building on this focus, we specifically chose the use of Kalman Filter (KF) and Hungarian Algorithm for pedestrian tracking. KF is still a widely used approach in MOT due to its efficiency in state estimation and motion prediction. However, challenges such as occlusion and ID switching persist, making it necessary to explore enhancements [7]. We realize that these problems arise with a reliance on two-dimensional data and a lack of understanding of three-dimensional environment, so our approach will take into account third dimensional information, such as depth information, to overcome the challenges of occlusion and ID switching, enhancing the overall accuracy and robustness of the tracking process. The rest of this report is organized as follows: Section 3 provides a review of related work, emphasizing traditional computer vision techniques for MOT and their pros and cons. Section 4 presents our

proposed approach, detailing its core concept and initial considerations. Finally, Section 5 offers a conclusion and discusses potential directions for future work.

3. Related Works

3.1 Traditional Tracking Methods

3.1.1 Kalman Filter with Hungarian Algorithm (SORT)

Bewley et al. [8] proposed Simple Online and Realtime Tracking (SORT). It is a tracking method that combines Kalman filtering and the Hungarian Algorithm. The Kalman filter is used to predict the state of a detected object in the next frame. The predictions are made on the assumption that the object will move in a linear way and at a constant velocity. Also, gaussian noise is added to the predictions to handle random deviations. The Hungarian algorithm is then used to match the predictions from the Kalman filter with measurements (object detections) from the next frame. These matches are then used to update the Kalman filter. This process is then repeated recursively until the last frame is completed

Strengths: Since SORT is simple and requires lower computation power compared to other methods, it is suitable for real-time tracking

Weaknesses: Since SORT assumes that the object will move in a straight line and with a constant speed, its predictions can be way off when objects deviate from the path abruptly as happens in the real world. Also, since it does not include depth information, it makes it harder for it to handle occlusions which results in label switching even if it is the same object when the object reappears.

3.1.2 Multiple Hypothesis Tracking (MHT)

Multiple Hypothesis Tracking (MHT) introduced by Reid [9] uses the Kalman filter to predict the locations of objects in the next frame. It then tries to match the predictions with the detected objects from the next frame using a likelihood score. If it is unsure, with a low probability below a certain threshold, it delays the decision to match and creates a tree of possible scenarios. It then eliminates branches that are unlikely to occur in the tree to prevent it from becoming too big and computationally expensive to process.

Strengths: MHT excels in handling occlusions and complex motion because it keeps track of all possible matches for a prolonged period.

Weaknesses: Its combinatorial complexity grows exponentially as more and more objects and frames are processed especially in densely populated scenarios. This makes it impractical for real-time applications. Like SORT, it lacks depth information, which results in occlusions which lead to label switches.

3.2 Deep Learning-Augmented Tracking Methods

3.2.1 Deep SORT with Appearance Features

Wojke et al. [10] extended SORT with DeepSORT. DeepSort integrates a CNN-trained appearance descriptor into the tracking framework. In addition to the Kalman Filter and Hungarian Algorithm, DeepSORT uses the distance between the appearance features to improve association. This complements the Intersection of Union costs.

Strengths: Appearance features reduce ID switches in occlusion-heavy pedestrian scenes. This improves its robustness over SORT.

Weaknesses: The CNN descriptor increases computational overhead, and its 2D focus misses depth cues which might help reduce occlusion issues.

3.2.2 FairMOT: End-to-End Tracking

The research of Zhang et al. [11] presented FairMOT as a fully connected multi-object tracking (MOT) system which unifies detection with tracking through deep learning. The system implements CenterNet as its backbone network to discover object centers while generating re-identification feature outputs.

Strengths: The end-to-end structure of FairMOT improves tracking precision in 3D pedestrian environments through successful detection in combination with association.

Weaknesses: One major drawback of this system is its need for substantial computing power and large datasets which reduces its usefulness for edge devices. Also, its limitation to 2D dimension operations prevents this system from understanding depth which produces inaccuracies whenever an object becomes occluded.

3.3 Comparison and Gaps

The SORT method is simple and less computationally expensive compared to other methods. This makes it suitable for real-time tracking especially when compute is limited. MHT is better than SORT at handling occlusions but this comes at a cost of time and compute as it has to keep track of a lot of states and possibilities. The deep learning methods used in DeepSort and FairMOT make them more accurate in data association but you would require a large dataset

and compute to train the models compared to the traditional approaches. The methods we have discussed all lack depth consideration in their execution. This makes them always susceptible to occlusion problems that arise from 3D environments. To better deal with this, we have proposed extending SORT, due to its simplicity and suitability to real-time tracking, to include depth information. This will help us better handle occlusions and reduce label switching in our pedestrian tracking application.

4. Proposed Method

4.1 Problem Statement

As mentioned earlier, ID switching due to occlusion is a prevalent problem in multi-object tracking. It becomes difficult for the algorithm to maintain track of a labeled object when another object comes in between the object and the camera frame. This in turn causes the algorithm to either switch the IDs of the two objects, if applicable, or label the occluded object with a new ID when the obstruction is removed. This issue stems from the 2D dimensional representation of our data and objects. This dimension reduction renders the current models mathematically incapable of accounting for the specified type of occlusion. This is due to a lack of understanding of the three-dimensional environmental factors causing the occlusion.

4.2 Core Idea

We propose the introduction of depth information into the tracking process of existing algorithms. Specifically, for Kalman Filtering and Hungarian Algorithm used in SORT [8]. The assumption is that using depth data will give a better understanding to our tracking algorithm of where objects are. Occlusions due to an object coming in between the tracked object and the frame would be expected and would allow for the tracking algorithm to account for them. Additionally, ID switching due to occlusions can be mitigated for intersecting objects, as each object's depth is now a differentiating factor. For longer instances of occlusion, we can also incorporate a memory system. Thus, once the body that was hindering the site of the tracked object moves away again, the original depth adds another metric for comparison. This approach would be great for applications such as autonomous driving, where there are LIDAR sensors that allow for the extraction of depth data using sensor fusion. This idea can also be expanded to include traditional 2D images using existing depth estimation algorithms and models. The current advancements in neural processing units and deep learning, allows for the use of light weight depth estimation models within our pipeline.

4.3 Technical Details

Algorithm:

1. Object Detection:
    - Input: Video Frame
    - Find bounding boxes for detected objects using existing object detection models.
    - Extract feature descriptors within each bounding box. These descriptors capture the local image structure of the detected object.
    - Output: Location and descriptors of detected objects.
2. Depth Estimation:
    - Input: Video Frame
    - Compute the depth map of the image using existing depth estimation algorithms.
    - Integrate the depth data into our existent detected objects and their corresponding descriptors. This is done by concatenating our original descriptors with the new depth data.
    - Output: Depth modified location and descriptors of detected objects.
3. Kalman Filter:
    - Input: Object descriptors and previous Kalman Filter weights
    - Predict what the new states are based on the previous frames.
    - Output: Predicted state of previous detected objects.
4. Hungarian Algorithm:
    - Input: Predicted state (label 10 states) from the Kalman Filter in step 3 and the extracted descriptors (no label 10 descriptors) from step 2.
    - Use the Hungarian Algorithm to match the predicted states and the actual states.
    - Set detected object IDs based on the results of the matching.
    - Output: Matched object labels
5. SIFT [12] (Optional for Testing):
    - Input: Unmatched IDs and Object Images
    - When no match is found, use the image from the previous found IDs, and apply SIFT on them.
    - Apply SIFT on the unmatched new detected object.
    - Attempt feature matching between the two sets.

- If a match is found, use the old ID. Else, assign a new ID.
- Output: Matched and new object labels for remaining objects.

6. Kalman Filter Update:
    - Input: Matched objects to previous labels.
    - Update the states based on the new object descriptors.
    - Output: Updated Kalman Filter

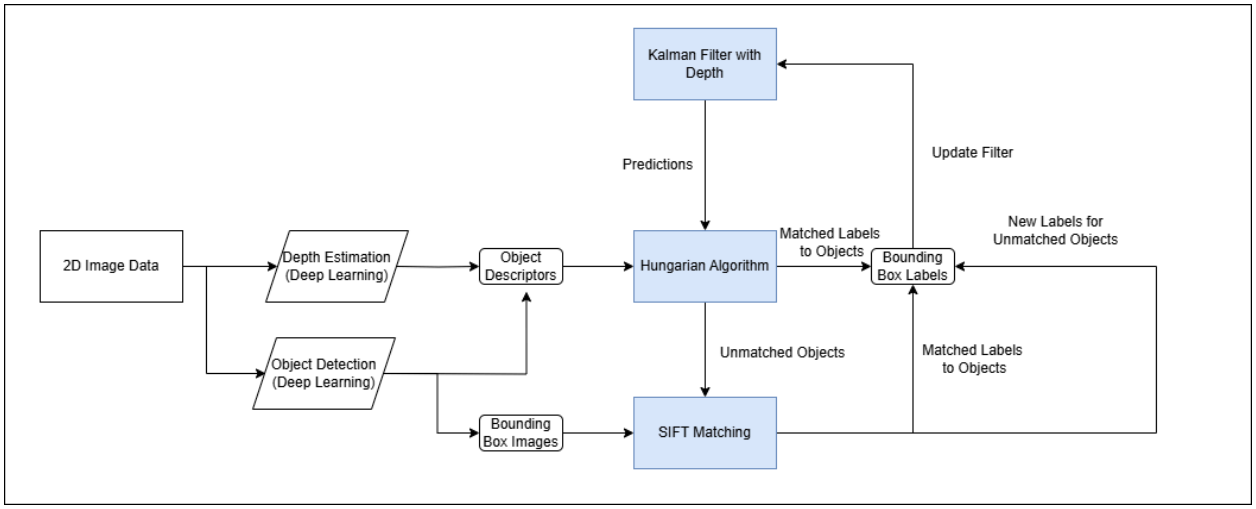

Figure 1: Algorithm Diagram

5. Conclusion and Future Work

This literature review overviews the challenges of Multiple Object Tracking (MOT), in occlusion and ID switching, which significantly impact tracking accuracy in real-world applications such as surveillance and autonomous driving. Although deep learning models have significantly improved performance, their high processing requirements and reliance on massive datasets make them less useful in resource-constrained contexts. Because they are effective, traditional techniques like the Kalman Filter are still frequently employed. However, they frequently fail to handle occlusion and maintain object identity, particularly in pedestrian tracking where human movement is unpredictable and ever-changing.

To address these challenges, the next step that we propose is an approach that integrates depth information into MOT systems to enhance object association across frames and reduce ID switching caused by occlusion. By incorporating three-dimensional spatial awareness, our proposed method aims to improve both accuracy and robustness while maintaining computational efficiency. In the next phase of our research, we will implement and evaluate depth-aware tracking algorithms, comparing them against existing methods through controlled experiments. Main challenges include optimizing performance for real-time applications,

maintaining seamless system integration, and refining feature-matching techniques using Scale-Invariant Feature Transform (SIFT) to improve object re-identification in heavily occluded scenes. These efforts will help develop a more reliable and adaptable MOT system for practical deployment.

References

[1] Klette, Reinhard. *Concise computer vision*. Vol. 233. London: Springer, 2014.

[2] Linda G. Shapiro; George C. Stockman (2001). *Computer Vision*. Prentice Hall.

[3] Tim Morris (2004). *Computer Vision and Image Processing*. Palgrave Macmillan.

[4] Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., & Kim, T. K. (2021). Multiple object tracking: A literature review. *Artificial intelligence*, *293*, 103448.

[5] Park, Y.; Dang, L.M.; Lee, S.; Han, D.; Moon, H. Multiple Object Tracking in Deep Learning Approaches: A Survey. *Electronics* 2021, *10*, 2406.

[6] Z. Shagdar, M. Ullah, H. Ullah and F. A. Cheikh, "Geometric Deep Learning for Multi-Object Tracking: A Brief Review," *2021 9th European Workshop on Visual Information Processing (EUVIP)*, Paris, France, 2021, pp. 1-6, doi: 10.1109/EUVIP50544.2021.9484040.

[7] Khodarahmi, M., & Maihami, V. (2023). A review on Kalman filter models. *Archives of Computational Methods in Engineering*, *30*(1), 727-747.

[8] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 3464–3468, doi: 10.1109/ICIP.2016.7533003.

[9] D. B. Reid, "An algorithm for tracking multiple targets," *IEEE Trans. Autom. Control*, vol. 24, no. 6, pp. 843–854, Dec. 1979, doi: 10.1109/TAC.1979.1102177.

[10] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 3645–3469, doi: 10.1109/ICIP.2017.8296962.

[11] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 1–17, doi: 10.1007/978-3-030-58580-8_1.

[12] D. G. Lowe, “Object recognition from local scale-invariant features,” *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999. doi:10.1109/iccv.1999.790410