Exploring the use of depth estimation to improve multi-object tracking in the context of ice hockey analytics

Ken M. Nsiempba[1]* **John Zelek**[1] **David Clausi**[1]
[1]Vision and Image Processing Group, System Design Engineering, University of Waterloo
{kmnsiemp,jzelek, dclausi}@uwaterloo.ca

Abstract

This study presents a novel approach to improve multi-object tracking within the field of Ice Hockey Analytics. By harnessing depth estimation, the goal is to tackle the common challenges related to tracking multiple objects in a 3D scene that is projected onto a 2D screen. The methodology encompasses acquiring video sequences, performing depth estimation for all frames, and subsequently conducting multi-object tracking on the resulting depth images. This study opens the door to more initiatives to enhance multi-object tracking in the field of hockey analytics.

1 Introduction

Generally, Ice Hockey analytics focus on puck-centric events such as shots and goals[1]. However, puck-centric events are not sufficient for understanding elements like global team strategy, which plays a crucial role in long-term team success. Generating comprehensive Ice Hockey analytics involves grasping simultaneous actions of players on the same team and how they contribute to a global strategy. This, in turn, requires understanding individual player actions and the evolution of their pose over time. Achieving this necessitates reliable simultaneous identification and tracking of players over substantial periods.

This article investigates methods of enhancing the tracking accuracies for the tracking of hockey players. It does so through the following timeline: firstly, it discusses the current state of tracking in ice hockey, with a focus on depth estimation and multi-object tracking. Subsequently, the method section provides detailed information about the workflow, the datasets used, and the data pre-processing methods. Finally, the results section will present qualitative and quantitative findings, including a discussion of the evaluation metrics employed.

2 Background

2.1 Ice Hockey Analytics

Improving the tracking of players in Ice Hockey faces multiple obstacles. Notably, the regular occlusions of players, the moving camera, the speed of the players etc...

In pursuit of this goal, various approaches have been explored. Vats et al. [2] introduced a transformer network designed for recognizing players through their jersey numbers in broadcast National Hockey League (NHL) videos. This transformer takes temporal sequences of player frames as input and outputs the probabilities of jersey numbers present in the frames. The same author has adopted a more comprehensive approach by identifying both the team and the player's jersey in a separate article [3]. Furthermore, other researchers have continued to focus on jersey recognition. In their work, Balaji et al. [4] proposed a robust keyframe identification module that extracts frames containing essential high-level information about the jersey number. A spatio-temporal network is then employed to model spatial and temporal context and predict the probabilities of jersey numbers in the video.

However, for our present work, we do not perform identification; we focus solely on tracking. The field of multi-object tracking is expansive, and the next subsection will delve into it further.

*https://botengu.github.io/portfolio/

2.2 Multi-object tracking

Multi-object tracking is the process of simultaneously detecting and tracking multiple objects in an image or a video sequence, with diverse applications, including surveillance, autonomous vehicles, robotics, and more.

When tracking is conducted frame by frame with real-time observations, the process is referred to as online. On the other hand, when tracking is performed on the entire sequence, considering all frames, it is known as offline. Multi-Object Tracking involves two primary tasks: object detection can be initialized either through object detection methods or manually initialized by manually labeling the objects [5, 6]. Once objects are located and identified in individual frames, tracking algorithms maintain the object identities across the frames. For the scope of this project, we will focus on offline object detection-based tracking methods.

Multi-Object Trackers (MOTs) can be implemented in multi-view or single-view (monocular) settings. Single-view analysis, such as broadcast views in hockey, is the most common as it requires less setup, but it also poses significant challenges for MOTs, including occlusions, scale variations, object interactions, appearance changes, and cluttered scenes. Addressing these challenges is a major focus in the literature.

Common approaches for tracking algorithms include more conventional methods like the Kalman filter, the particle filter, and the Hungarian algorithm. However, recent advances in deep learning have significantly enhanced the accuracy of Multi-Object Trackers (MOTs). Popular tracking/detection algorithms include the YOLO (You Only Look Once) trackers [7]. The YOLO trackers are based on the 'unified' concept, enabling the simultaneous prediction of multiple bounding boxes and class probabilities, thereby improving both speed and accuracy. This technique redefines object detection as a regression challenge, mapping it to spatially distinct bounding boxes and corresponding class probabilities. The advantage lies in optimizing the entire detection pipeline as a single network, enabling end-to-end optimization for enhanced detection performance. YOLO-v8 [8] represents the latest version of this algorithm.

More recently, the introduction of the transformer model [9] has led to the development of transformer MOTs [10–12], further contributing to accuracy improvement.

However, despite these advancements, the accuracies, as measured using the multi-object tracking accuracy (MOTA) benchmark [13], still stagnate between 70% and 80%. One of the leading impediments to good accuracies in tracking models is the occurrence of ID switches. For every tracked object, an ID is assigned, and objects follow specific paths through the frames, known as tracklets. ID switches happen during mismatches between objects and tracklets.

Since trackers are predominantly used on 2D, unprocessed images, leveraging elements in images such as depth estimation could provide trackers with additional information about the third dimension, potentially enhancing final accuracies.

Several studies have demonstrated the benefits of employing new representations to enhance multi-object tracking [14–16]. Liu et al. [15] developed a depth cascading method that calculates the minimum and maximum pseudo-depth values for both detection and trajectory sets in a given RGB image. The algorithm transforms depths into sparse target subsets and conducts data association by prioritizing objects from near to far, dividing the interval between them into depth intervals. It associates trajectories and detections at the same depth level using Intersection over Union association. The promising results shown by Liu et al. [15] underscore the potential of utilizing depth estimation techniques. Sparsetrack has performed admirably; however, the nature of the data used (MOT) differs qualitatively from ours, making it not necessarily generalizable. In other words, the

Sparsetrack algorithm works with images where the depth images have not been added with RGB images.

To allow for more flexibility, we have chosen to perform tracking and depth estimation separately. Additionally, breaking down the workflow affords us greater control over parameters such as the type of depth estimation, tracking method, and concatenation approach for the final output.

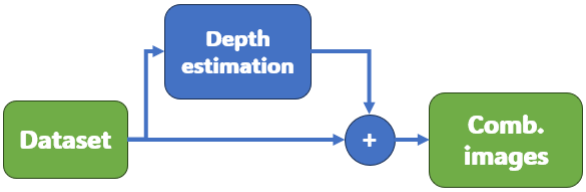
2.3 Depth estimation

Depth estimation is crucial in computer vision and 3D scene understanding, finding applications in robotics, augmented reality, autonomous vehicles, and more. It involves methods to infer the depth or distance of objects in a scene from 2D images or image sequences. Depth information can be acquired using depth sensors and stereo vision, but for monocular depth estimation (MDE), the focus is on estimating depth from monocular views which is what we will use in our scenario since we are dealing with broadcast views. Recently, transformers have proven useful for MDE [17]. Additionally, depth estimation can benefit from cross-modal information, such as incorporating semantic details from RGB images, enhancing the accuracy of depth prediction.

Diffusion models have been used for depth estimation. They operate by destroying training data through the successive addition of Gaussian noise and then learning to recover the data by reversing this noising process. Ke et al. [18] have opted for latent diffusion models to perform MDE. They have used a frozen variational autoencoder to put both the image and its depth map into a hidden space for training their denoiser that works conditionally. For most papers, their understanding of the world relies on training data. However, with stable diffusion, understanding can be drawn from priors.

3 Methods

(a) Workflow to obtain depth images



(b) Workflow to obtain combined images

Fig. 1: Workflow

Our approach is two-fold. Firstly, we generate depth maps for selected images in our dataset using the method proposed by [18]. Instead of considering the monochromatic colormap, we have opted for a spectral one. Performing piecewise addition with spectral colormaps has a more pronounced effect, resulting in an image different enough to impact the tracking stage. Subsequently, the generated depth images are utilized to reconstruct a video, which is then input into the YOLO-v8 model. The model outputs the location and properties of the final bounding boxes and corresponding IDs for all the frames.

3.1 Datasets

The ice hockey sequences were obtained from the McGill Hockey Player Tracking Dataset (MHPTD) [19]. The dataset follows a format similar to the popular MOT challenge dataset used for pedestrian tracking. Each entry in the dataset represents an instance of a hockey player. The key distinction lies in the assignment of identity: MHPTD assigns identity at a personal level, while the MOT challenge assigns identity at a tracklet level. In MHPTD, the same identity is assigned to a person who exits and re-enters the field of view, resulting in multiple tracklets sharing the same identity.

The dataset comprises 25 high-definition NHL (National Hockey League) gameplay video clips, each capturing one shot of the gameplay from an overhead camera position. A "shot" is defined as a series of frames that run without interruption, without cut or camera switch. To accommodate different NHL broadcast video frame rates (60 fps and 30 fps), half of the video clips have a frame rate of 30 fps, and the other half have a frame rate of 60 fps. The annotation of the videos was performed using the Computer Vision Annotation Tool (CVAT), an open-source video annotation tool [20].

3.2 Data representation

The various video clips belong to different competitions. For this experiment, we have selected the initial 10 seconds (300 frames - 30 fps) of the first (001) clip from the All-Star competition. The corresponding depth images for these frames were captured and utilized to create three datasets with distinct data representations (refer to Figures 1 and 2).

- The first dataset taken was the original(referred to as regular) broadcast frames.
- The second type of data was the depth images.
- The third dataset was obtained by superposing both the weighted depth images onto their corresponding weighted original frames. The respective weights were 0.7 and 0.3. The superposition was done using piecewise addition instead of concatenation. This is because the tracker was not adapted for 4-layer "RGBD" images.

4 Results

During the post-process operation, bounding boxes were extracted from the ground truth labels for the McGill Hockey dataset as well as for the YOLO algorithm. The existing ground truth labels provided by the McGill Hockey dataset were utilized. The ground truth labels were reformatted into a dictionary data structure format, where the keys of the dictionary corresponded to the frame number. For each frame, there were two lists: the first list contained the object IDs, and the second list contained sublists which each contained three items, representing the bottom left point coordinates, width, and height of each bounding box. The order of the IDs in the first list corresponded to the order of the sublists containing the bounding box information in the second list. The predicted bounding boxes for the different data representations followed the same structure.

4.1 Qualitative results

Figure 3 shows the superposition of the predicted tracked bounding boxes for the various data representations as well as the ground truths. All these bounding boxes are displayed on top of the original first frame. Despite it only being the first frame, the results suggest a low bounding box count for the depth images, where the only detected people are outside of the rink. The bounding boxes for the combined and regular experiments appear to detect most people on the rink correctly and align well with the ground truth labels.

4.2 Quantitative results

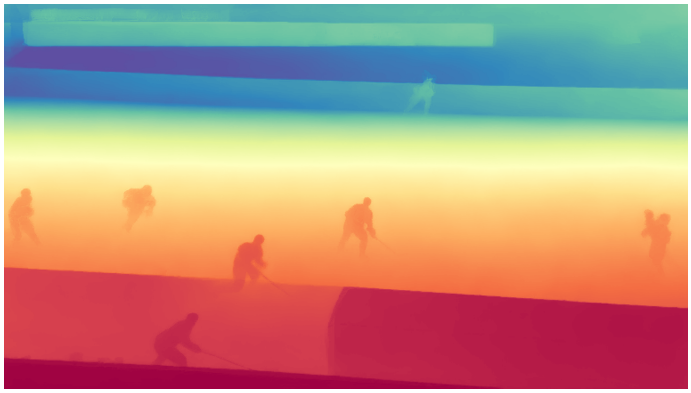
4.2.1 Dice score

The dice score is a metric used to quantify the similarity between two sets, commonly employed in the evaluation of segmentation tasks, particularly in the field of medical image analysis and computer vision [21].

The Dice score is defined as:

$$Dice = 2 \cdot \frac{|P \cap G|}{|P| + |G|}$$

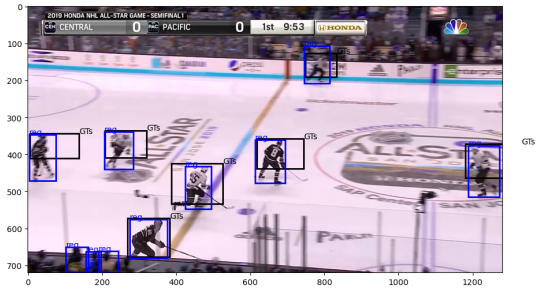
(a) First frame of original footage



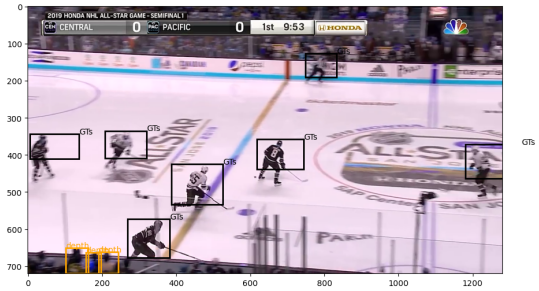
(b) Equivalent depth map for original frame

(c) Combined original frame and depth map

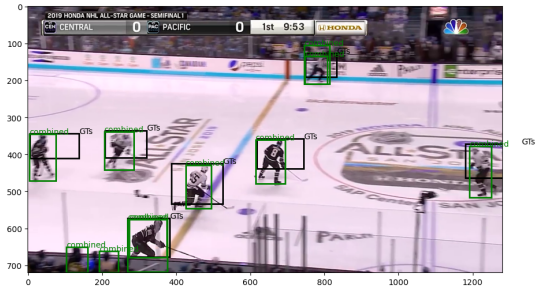
Fig. 2: Data representations



(a) Ground truths' bounding boxes and predicted bounding boxes for the regular frames



(b) Ground truths' bounding boxes and predicted bounding boxes for the depth images



(c) Ground truths' bounding boxes and predicted bounding boxes for the combined images

Fig. 3: Qualitative results

The numerator represents the number of elements that are common to both sets (P and G), while the denominator refers to the sum of the elements in the sets. In the context of image segmentation, these sets typically correspond to the pixels in the predicted segmentation mask and the ground truth segmentation mask. The Dice score ranges from 0 to 1, where 0 indicates no overlap between the sets (no agreement between the predicted and ground truth segmentation), and 1 indicates perfect overlap.

Initially, we performed box matching and compared each dataset with the ground truths. For each key in the dictionary, we identified the boxes with the most overlap with the ground truths, and then we calculated the mean Dice score for the frame. The Dice score coefficient reflects the overlap between the matched boxes. For display purposes, we have grouped the scores across multiple frames.

Figure 4 shows that while the regular and combined representations have similar scores (with the Dice score of the combined representation being slightly smaller). The Dice score for the depth images on the other hand is considerably smaller.

4.2.2 ID switches

To detect ID switches in the different tracked footage, we have developed a specific procedure. Initially, we create an ID-tracking table where each entry contains bounding box information. The row number corresponds to the ID of the tracked object, and the column corresponds to the frame number for which the object is tracked (see Figure 5). Each row in this table represents the evolution of the bounding box properties (size and coordinates) over time.

To estimate the number of ID switches for a particular prediction, we create two ID-tracking tables: one for the ground truths and one for the predicted case using a specific data representation. Let's denote the ID-tracking tables of the ground truth and the specific data representation as T_{gt} and T_{d-rep}, respectively. We create an empty table, $T_{results}$, which has the same dimensions as T_{gt}.

Entry ij refers to the bounding box properties for an object with ID i at frame j. For every frame j in T_{gt}, the bounding box of object i is compared to all the bounding boxes present in that frame for T_{d-rep} (the entire column associated with a particular frame). We search for the bounding box in the table T_{d-rep} that overlaps the most with the bounding box of object i in T_{gt}. We record the ID of that bounding box and place it in entry ij in $T_{results}$, as shown in Figure 5

To avoid any ID switches, every entry in a row of $T_{results}$ should have the same entry number corresponding to the specific ID from table T_{d-rep}. The number of changes reflects the amount of ID switches. The number of ID switches for every object from the original ground truths is added to give the total ID switches.

The results are shown in Table 1:

The depth representation method shows fewer ID switches than

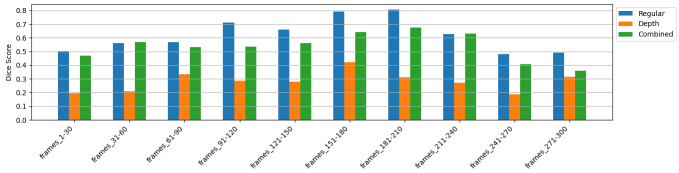


Fig. 4: Mean dice score for different frame intervals

	frame_1	frame_2	frame_3
ID_1	GT_bbox_11	GT_bbox_12	GT_bbox_13
ID_2	GT_bbox_21	GT_bbox_22	GT_bbox_23
ID_3	GT_bbox_31	GT_bbox_32	GT_bbox_33

	frame_1	frame_2	frame_3
ID_1	DR_bbox_11	DR_bbox_12	DR_bbox_13
ID_2	DR_bbox_21	DR_bbox_22	DR_bbox_23
ID_3	DR_bbox_31	DR_bbox_32	DR_bbox_33
ID_4	DR_bbox_41	DR_bbox_42	DR_bbox_43

	frame_1	frame_2	frame_3	#Switches
ID_1	3	3	1	1
ID_2	1	1	4	1
ID_3	2	2	2	0

Fig. 5: Example of tables required for the ID-switch method. The top left one stores the boxes info for the ground truths (T_{gt}), the bottom left stores info related to the predicted boxes for the specific data representation(T_{d-rep}). The right one, $T_{results}$, tracks the ID switches.

	Regular	Depth	Combined
ID Switches	183	155	390

Table 1: Table showing ID switches

the other method, but that is also because there are fewer boxes, as can be seen in the original Figure 3. However, the regular data representation performs better than the combined one. This indicates that the addition of the depth map may have negatively impacted the tracker's performance.

5 Conclusion and future work

The experiments have indicated that depth maps, or at least the current method of integration, are not sufficient to improve tracking. Despite the interesting initiative, further validation is required to establish depth maps as viable solutions. The depth map has the potential to be extremely informative and could aid in preventing occlusions for ice hockey analytics. Additional research is needed on how to appropriately integrate the depth map with the original image before feeding it into the multi-object tracking system (i.e. exploring 4-channel RGBD data representation). Furthermore, we are also planning to combine regular frames and depth maps within the latent space [22] to further enhance the tracking process.

References

[1] U. Johansson, E. Wilderoth, and A. Sattari, "How analytics is changing ice hockey," in *Linköping Hockey Analytics Conference*, 2022, pp. 49–59.

[2] K. Vats, W. McNally, P. Walters, D. A. Clausi, and J. S. Zelek, "Ice hockey player identification via transformers and weakly supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3451–3460.

[3] K. Vats, P. Walters, M. Fani, D. A. Clausi, and J. S. Zelek, "Player tracking and identification in ice hockey," *Expert Systems with Applications*, vol. 213, p. 119250, 2023.

[4] B. Balaji, J. Bright, H. Prakash, Y. Chen, D. A. Clausi, and J. Zelek, "Jersey number recognition using keyframe identification from low-resolution broadcast videos," in *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, 2023, pp. 123–130.

[5] Y. Ming, X. Meng, C. Fan, and H. Yu, "Deep learning for monocular depth estimation: A review," *Neurocomputing*, vol. 438, pp. 14–33, 2021.

[6] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T.-K. Kim, "Multiple object tracking: A literature review," *Artificial intelligence*, vol. 293, p. 103448, 2021.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[8] [Online]. Available: https://www.ultralytics.com/

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[10] X. Zhou, T. Yin, V. Koltun, and P. Krähenbühl, "Global tracking transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8771–8780.

[11] P. Chu, J. Wang, Q. You, H. Ling, and Z. Liu, "Transmot: Spatial-temporal graph transformer for multiple object tracking," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4870–4880.

[12] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "Trackformer: Multi-object tracking with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8844–8854.

[13] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.

[14] C.-J. Liu and T.-N. Lin, "Det: Depth-enhanced tracker to mitigate severe occlusion and homogeneous appearance problems for indoor multiple-object tracking," *IEEE Access*, vol. 10, pp. 8287–8304, 2022.

[15] Z. Liu, X. Wang, C. Wang, W. Liu, and X. Bai, "Sparsetrack: Multi-object tracking by performing scene decomposition based on pseudo-depth," *arXiv preprint arXiv:2306.05238*, 2023.

[16] S. Yan, J. Yang, J. Käpylä, F. Zheng, A. Leonardis, and J.-K. Kämäräinen, "Depthtrack: Unveiling the power of rgbd tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 725–10 733.

[17] A. Masoumian, H. A. Rashwan, J. Cristiano, M. S. Asif, and D. Puig, "Monocular depth estimation using deep learning: A review," *Sensors*, vol. 22, no. 14, p. 5353, 2022.

[18] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," *arXiv preprint arXiv:2312.02145*, 2023.

[19] K. C. Yingnan Zhao, Zihui Li, "A method for tracking hockey players by exploiting multiple detections and omni-scale appearance features," *Project Report*, 2020.

[20] B. Sekachev, N. Manovich, M. Zhiltsov, A. Zhavoronkov, D. Kalinin, B. Hoff, TOsmanov, D. Kruchinin, A. Zankevich, DmitriySidnev, M. Markelov, Johannes222, M. Chenuet, a andre, telenachos, A. Melnikov, J. Kim, L. Ilouz, N. Glazov, Priya4607, R. Tehrani, S. Jeong, V. Skubriev, S. Yonekura, vugia truong, zliang7, lizhming, and T. Truong, "opencv/cvat: v1.1.0," Aug. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4009388

[21] J. Bertels, T. Eelbode, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 2019, pp. 92–100.

[22] K. Genova, F. Cole, A. Sud, A. Sarna, and T. Funkhouser, "Local deep implicit functions for 3d shape," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4857–4866.