Research Article

pISSN 1225-6161 • eISSN 2287-9307
Korean Journal of Remote Sensing 2024, 40(1), 115-122
https://doi.org/10.7780/kjrs.2024.40.1.11

Korean Journal of Remote Sensing
KJRS

Multi-Class Multi-Object Tracking in Aerial Images Using Uncertainty Estimation

Hyeongchan Ham[1]*, Junwon Seo[1], Junhee Kim[2], Chungsu Jang[2]

[1]Researcher, Advanced Defense Science & Technology Research Institute, AI Autonomy Technology Center, Agency of Defense Development, Daejeon, Republic of Korea
[2]Senior Researcher, Advanced Defense Science & Technology Research Institute, AI Autonomy Technology Center, Agency of Defense Development, Daejeon, Republic of Korea

Abstract: Multi-object tracking (MOT) is a vital component in understanding the surrounding environments. Previous research has demonstrated that MOT can successfully detect and track surrounding objects. Nonetheless, inaccurate classification of the tracking objects remains a challenge that needs to be solved. When an object approaching from a distance is recognized, not only detection and tracking but also classification to determine the level of risk must be performed. However, considering the erroneous classification results obtained from the detection as the track class can lead to performance degradation problems. In this paper, we discuss the limitations of classification in tracking under the classification uncertainty of the detector. To address this problem, a class update module is proposed, which leverages the class uncertainty estimation of the detector to mitigate the classification error of the tracker. We evaluated our approach on the VisDrone-MOT2021 dataset, which includes multi-class and uncertain far-distance object tracking. We show that our method has low certainty at a distant object, and quickly classifies the class as the object approaches and the level of certainty increases. In this manner, our method outperforms previous approaches across different detectors. In particular, the You Only Look Once (YOLO)v8 detector shows a notable enhancement of 4.33 multi-object tracking accuracy (MOTA) in comparison to the previous state-of-the-art method. This intuitive insight improves MOT to track approaching objects from a distance and quickly classify them.

Keywords: Detection, Tracking, Uncertainty, Deep learning

Received: February 14, 2024
Revised: February 22, 2024
Accepted: February 25, 2024
Published: February 28, 2024

Corresponding author:
Hyeongchan Ham
E-mail: hyeongchan@add.re.kr

1. Introduction

Multi-object tracking (MOT) is mostly used in computer vision tasks such as military, surveillance, robotics, and smart cities. The goal of the MOT is the surveillance to recognize and classify distant objects and to identify potential threats. The MOT model tracks objects based on the association between previously tracked objects and newly detected objects. Most MOT models (Bewley et al., 2016; Wojke et al., 2017; Aharon et al., 2022; Du et al., 2021) have utilized the Hungarian algorithm (Kuhn et al., 1955) with the Kalman filter (Kalman et al., 1960) and the Re-Identification (ReID) appearance model (Luo et al., 2019; He et al., 2021) to establish the association. The Hungarian algorithm utilizes a cost matrix to establish associations between tracked and detected objects. The position or appearance information of the objects composes the cost matrix which is formed by the Kalman filter and the ReID appearance model. The Kalman filter updates and predicts positional information to associate between tracked and detected objects. The ReID appearance model can be applied to compare the appearance between tracked and detected objects for the association. This tracking process is generally used for single-class datasets like MOT17 (Milan et al.,

2016), which consists of pedestrians.

The rapid identification and classification of the potential threats that are approaching from a distance are crucial in the surveillance (Araujo et al., 2019). While traditional MOT methods have shown effectiveness in single-class MOT tasks like the MOT17 dataset, they encounter considerable challenges in multi-class MOT tasks. Quasi-Dense tracking (QDTrack) (Fischer et al., 2022) demonstrated that a naïve tracking approach cannot handle misclassified detections for the multi-class MOT task. The most straightforward approach is to extend the single-class MOT process (Wang et al., 2020; Zhou et al., 2020; Zhang et al., 2021) to each class.

The extension of the single-class MOT process iterates through each class and associates tracks and detections within each class. This class-by-class approach predicates the perfect classification with an oracle detector to consistently track the objects. Previous works enhance the feature extraction of the object detector to improve the detection performance in the aerial images (Zhao et al., 2023; Zhu et al., 2023). Nonetheless, the object detection models still accompany errors in the classification. In particular, distinguishing object classes of distant and tiny objects is more challenging (Wang et al., 2021), as depicted in Fig. 1.

The VisDrone dataset (Chen et al., 2021), which is collected from a high altitude and captures a far field of view, suffers from these problems. Wrongly classified detections cannot be matched to the tracks of another class. Thus inaccurate classification results in both false positives and false negatives. Track fragmentation and redundant tracklets occur after this misclassified frame (Fischer et al., 2022). The threat level classification is not successful until the object reaches a proximity that allows for its clear identification. To track approaching objects from a distance, this problem needs to be solved. Therefore, the tracker must address the classification uncertainty to quickly and adaptively recognize the objects.

The existing multi-class MOT methods are sub-optimal when it comes to addressing uncertain classification issues in the object detector. In previous works, voting methods (Aharon et al., 2022; Du et al., 2021) have been adopted to address the classification errors that occur when detecting small objects at a distance. These voting methods ensemble classification results from the detection model output of each frame.

Each of the following voting methods infers the most appropriate category for the current timestamp based on their strategy. The hard voting method (Aharon et al., 2022) infers the current class by counting the number of classes during the current frame and choosing the majority class. The soft voting method (Du et al., 2021) differs from the hard voting method in that it works using the detection probability for each frame. It uses the confidence score obtained from the object detection model for the weighted voting method. The confidence score obtained from the object detector indicates the probability of the presence of an object within the bounding box (Redmon et al., 2016).

In other words, the confidence score reflects the objectness of detection rather than the class uncertainty score, making it an unsuitable metric for measuring class confidence level. The voting method using objectness scores results in the accumulation of inaccurately predicted class information since the objectness score which is used as weights for voting is irrelevant to the classification. Even if the correct classification is made, it will take time to compensate for the previous errors. As shown in Fig. 1,

Fig. 1. In the depicted figure, objects located at a distance appear significantly tiny and are challenging to distinguish, a common characteristic observed in aerial images. This results in a notable class uncertainty for these distant objects. As the sequence progresses, the drone camera approaches the object and leads to an increase in its apparent size and clarity. The enhanced resolution and detail as the camera approaches confirm the object as a car, concurrently reducing the associated uncertainty.

https://doi.org/10.7780/kjrs.2024.40.1.11

distant small objects have low-class probabilities even though it is certain that they exist. Therefore, these objects can be detected with the correct bounding box and the wrong classification. Given the aforementioned considerations, It is important to propose a new approach for rapid class adaptation.

In this paper, we resolve the class uncertainty problem in a multi-class MOT task using an uncertainty-weighted voting approach. The multi-class association method is employed to maintain the tracklet even though the detection results in an incorrect classification. We also propose the class update module that incorporates classification uncertainty as a voting weight. By applying the proposed method, the classification score collected from the uncertain classification has a small weight for voting, whereas the definite classification has a high weight for voting. Experiments conducted on the VisDrone-MOT dataset (Chen et al., 2021) show that weighted voting with classification uncertainty leads to improved tracking accuracy. Furthermore, our method has rapid adaptation for the uncertain classification.

2. Materials and Methods

The objective of our study is to address the challenge of uncertain classification in multi-class MOT by incorporating an extra class update module in addition to the classification of the existing detection model (Jocher et al., 2022; 2023). We first formulate the single-class MOT task (Aharon et al., 2022). Thereafter, we expand it to multi-class MOT by introducing a multi-class association method, and multi-class update module. We compare our class update method with the previous multi-class MOT methods (Aharon et al., 2022; Du et al., 2021). The overall pipeline is shown in Fig. 2.

2.1. Revisiting Multi-Object Tracker

Most single-class MOT components are similar to those of multi-class, as depicted in Fig. 2(a), except for the association module component. Both multi-object trackers process successive frames. The input frame passes an off-the-shelf object detector, which returns detection results consisting of bounding boxes, classification scores, and confidence scores. The association module computes the similarity cost matrix between these detection results and the current tracking objects. Subsequently, a Hungarian assignment (Kuhn et al., 1955) is conducted to match the tracks and detection results using the similarity cost matrix. There are three types of outputs after the matching between the tracks and detection results; matched track-detections, unmatched detections, and unmatched tracks. Matched track-detections are assigned

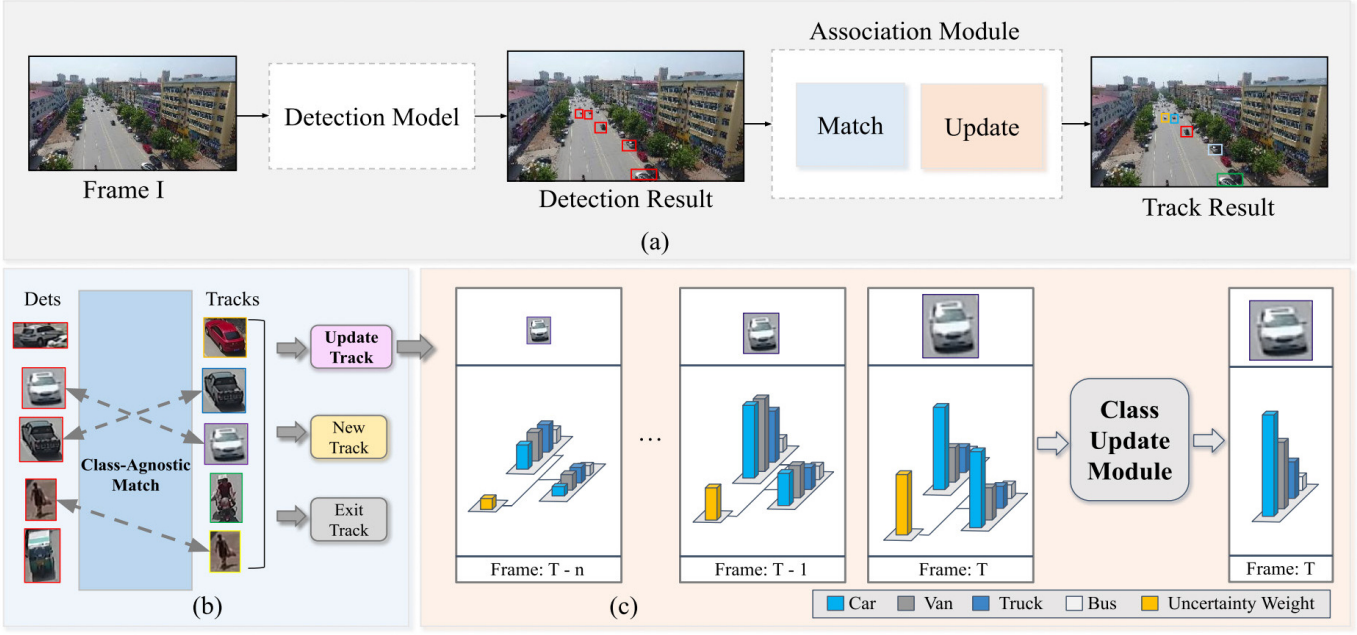

Fig. 2. Illustration of the proposed multi-class multi-object tracking model. (a) describes overall architectures. (b) shows the matching stage of the association module and its results are matched track-detections (update track), unmatched detections (new track), and unmatched tracks (exit track). For the matched tracks, the update module is applied with matched detections as shown in (c). One box represents the detection result of each frame with the probability of each class, and T indicates the current frame. The class update module computes the weighted probability for each frame, considering the uncertainty weight.

with the high similarity cost matrix and this matched detection will extend the track after the update step. The update step includes position, appearance, and class information updates.

Unmatched detections are the detections except the matched detections. They are considered as the emergence of new objects and become new tracks. Unmatched tracks are the tracks except the matched tracks. They indicate that the tracks are no longer tracked and disappeared so they will be removed in the exit step. Our method focuses on the class update of the matched track detections. The information of the matched track detections is updated by the detection results obtained from the current frame. While association and update within the same class for single-class MOT are naïve, they become non-trivial in the context of the multi-class MOT. Therefore, we introduce a new association method.

2.2. Multi-Class Association Method

In contrast to the single-class MOT, the multi-class MOT utilizes the class-agnostic association methods. In most MOT tasks, position and appearance information are used as similarity costs to match tracked and detected objects in the current frame. The single-class MOT does not need to consider classes for the association because every detection and track has the same class. However, association within the same class is not optimal for the multi-class MOT. The classification of the detector always involves errors and these errors lead to the wrong track creation. Therefore, we use a class-agnostic association method to give a chance for wrong classification cases, which is depicted in Fig. 2(b). This class-agnostic method utilizes only position and appearance information for association. The misclassified object does not create a false positive track because misclassification does not affect association.

2.3. Track Class Update Method

The concept of class uncertainty is utilized to update tracking information of associated objects. After the association module, there are three types of results: matched tracks, unmatched detections, and unmatched tracks. For the matched tracks, they need to update track information with the detection result of the current frame. As shown in Fig. 2(c), the track class update method is used to update class information in addition to the position and appearance information from the new frame. The uncertainty method, proposed in this paper, is weighted voting with classification uncertainty estimation. As shown in Fig. 1, the classification of distant small objects can be challenging.

In this case, the classification result is unreliable; therefore, we should slightly consider the detection class for the track class. This unreliable detection class has a small weight. Conversely, as the object comes closer and can be recognized clearly, its weight should be increased. The entropy quantifies this uncertainty and it becomes smaller as it is more certain. The reciprocal of entropy (Joshi et al., 2009) can be used to represent class uncertainty as a tracking class weight. The following formula represents uncertainty weight w,

$$w = \frac{1}{E} = -\frac{1}{\Sigma_c p_c \log(p_c)} \tag{1}$$

where p_c is the probability for class c, and the entropy E is the negative value of the weighted sum of these probabilities. According to information theory (Bishop, 2006), entropy is maximized when the probability is uniform and minimized when its distribution is sharp. When the object is confused as multiple classes, it has a high uncertainty. Whereas when it is confidently classified as a certain class, it has a low uncertainty. Therefore it is appropriate to consider the reciprocal of entropy as a voting weight when determining the class of a track. The classification probability obtained from each frame is used to compute the uncertainty weight, and predicts the class of the current track as follows:

$$C = \underset{c}{argmax} \left(\sum_{t}^{T} w_t \cdot p_{c,t} \right) \tag{2}$$

where $p_{c,t}$ is the probability of the object for class c at frame t and w_t is the uncertainty weight at frame t. T is the current frame.

3. Results

3.1. Dataset Description

The proposed method is evaluated on the VisDrone-MOT2021 (Chen et al., 2021) dataset. The VisDrone dataset is mostly used as a benchmark for unmanned aerial vehicle (UAV) surveillance models. The data are captured using UAV drones and are publicly available. It includes 96 video sequences, 56 training datasets with 24,201 frames, 7 validation datasets with 2,819 frames, and 33 test datasets with 12,968 frames. The dataset covers 10 categories; pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor. This experiment focuses on the classification ability of a track. To reduce class imbalance problems with other classes, we sampled four classes, car, van, truck, and bus, which are in vehicle categories with similar

https://doi.org/10.7780/kjrs.2024.40.1.11

appearances, thus classification ambiguity may occur.

3.2. Evaluation Metrics

To assess the effectiveness of MOT, the evaluation metrics employed are Multi-Object Tracking Accuracy (MOTA) and IDentification F1 score (IDF1) (Bernardin et al., 2008). MOTA measures the accuracy of the detected bounding box and is penalized by the detection error and tracking ID switches. MOTA is computed as follows:

$$\text{MOTA} = 1 - \frac{FP + FN + IDs}{GT} \tag{3}$$

where false positive (FP) and false negative (FN) are the number of unmatched predictions and unmatched ground truth. Ground-truth (GT) is the number of labeled objects in the datasets and ID switch (IDs) is the number of track ID switches for each object. A higher MOTA indicates a more accurate tracker and MOTA value is decreased as the error occurs.

A negative value of the MOTA means it has more errors than the number of ground-truth. For multi-class MOT, the total MOTA is computed by class-agnostic measurement. We reported the MOTA for each class and the average MOTA value of all classes (Avg). Also, the IDF1 metric is mostly used for the MOT task. Compared to the MOTA metric, which focuses on detection accuracy, IDF1 more focuses on the identification of tracking objects. IDF1 is computed as follows:

$$\text{IDF1} = \frac{2 \cdot IDPrecision \cdot IDRecall}{IDPrecision + IDRecall} \tag{4}$$

3.3. Implementation Details

The object detection model in the tracker used You Only Look Once (YOLO)v5L (Jocher et al., 2022) and YOLOv8L (Jocher et al., 2023) models. The association algorithm followed Aharon et al. (2022). To compensate for the tracking error from camera movement, we used the camera motion compensation module (Aharon et al., 2022). The training optimizer for the detection model is stochastic gradient descent (SGD) with an initial learning rate of 1e-4 and 10 times descent at 60 epochs. The training data was from the VisDrone-MOT and VisDrone-DET training splits. The training input images were resized to 1,580 pixels to detect small objects. The inference input images were resized to 2016 and test-time augmentation was not applied. We used four NVIDIA 2,080ti GPUs with batch size four. We implemented 100 training epochs.

3.4. Experiment Results

3.4.1. Effectiveness of the Proposed Module

We applied some class update modules for the same detector and tracker in the VisDrone-MOT2021 dataset. As shown in Table 1, our uncertainty-based class update module with the YOLOv5L detector had a 3.04 higher MOTA compared to the soft-voting-based class update module. Our uncertainty-based method gives less weight to high uncertainty and more weight to low uncertainty for the track class update. In contrast to the soft-voting method incorporating the objectness score as a weight, the uncertainty-based method can adaptively track with the correct category. The MOTA of van, bus, and truck categories are increased by 70.27, 6.31, and 10.41 each.

However, the MOTA of the car category decreased to 5.59. The VisDrone dataset seems to suffer from a class imbalance problem, and our class update module reduced this problem. The average MOTA is increased by 20.35. Also, the overall IDF1 decreased to

Table 1. Tracking results on VisDrone-MOT2021 test-dev dataset (%)

Model	Method	MOTA						IDF1					
		Car	Van	Bus	Truck	All	Avg	Car	Van	Bus	Truck	All	Avg
YOLOv5L	History-Centric	32.79	-123.27	-3.97	5.85	15.53	-22.15	65.58	28.45	54.87	49.61	58.85	49.63
	Observation-Centric	36.35	-132.46	8.39	21.71	19.41	-16.50	66.53	30.18	63.02	57.67	60.44	54.35
	Hard-Voting	37.70	-111.62	11.84	27.80	23.03	-8.57	68.04	33.39	63.57	59.19	62.45	56.05
	Soft-Voting	37.81	-90.58	16.99	28.33	25.11	-1.86	68.43	35.86	64.65	59.57	63.40	57.13
	Uncertainty (Ours)	32.22	-20.31	23.30	38.74	28.15	18.49	61.51	38.88	65.49	61.30	59.37	56.80
YOLOv8L	History-Centric	46.21	-72.22	-26.88	28.28	32.84	-6.15	71.07	36.92	57.89	49.87	64.66	53.94
	Observation-Centric	47.70	-74.45	-12.89	42.50	35.53	0.72	71.20	36.24	61.45	57.96	65.40	56.71
	Hard-Voting	48.92	-59.40	1.36	47.94	38.60	9.71	72.35	39.13	64.19	59.81	67.09	58.87
	Soft-Voting	49.07	-57.24	5.20	49.29	39.10	11.58	72.46	39.60	65.12	60.31	67.32	59.37
	Uncertainty (Ours)	49.01	-11.46	27.86	49.46	43.43	28.72	69.23	41.24	70.81	58.71	65.46	60.00

Table 2. Comparison of the duration required to achieve accurate classification

Method	Hard-voting	Soft-voting	Uncertainty (Ours)
Duration (%)	34.97	34.68	32.99

4.03. This degradation also comes from the majority number of car classes, and IDF1 scores of other classes increased using our method. In a surveillance scenario where numerous objects are tracked, maintaining the overall tracking ID with a minimum number of ID switches is more important than the IDF1 score maintaining one track ID for a long time. We showed the superiority of the proposed method by keeping the average IDF1 similar to the SOTA method and increasing the average MOTA.

This enhancement in our method comes from the ability of the rapid adaptability of our class update module. To demonstrate rapid adaptation, an experiment is conducted to measure the duration time required for the correct classification. To quantify the correctly classified time, the first correctly classified time is divided by the total observation time and gets the average across all the tested objects. As shown in Table 2, it is evident that our method can adapt faster compared to other methods.

3.4.2. Visualizations

We present a visual explanation of the white car tracking mentioned in Fig. 1. The car object has a similar appearance to the van. Classification score change for each frame is shown in Fig. 3. When the object is located at a far distance (frame 0 to 100), it is observed as small in size and is difficult to distinguish whether it is a car or a van. Therefore both car and van class probability (blue and green line) are small. As it is difficult to distinguish the category, it has a high uncertainty score (blue transparent bar), which means it is uncertain. With time, the

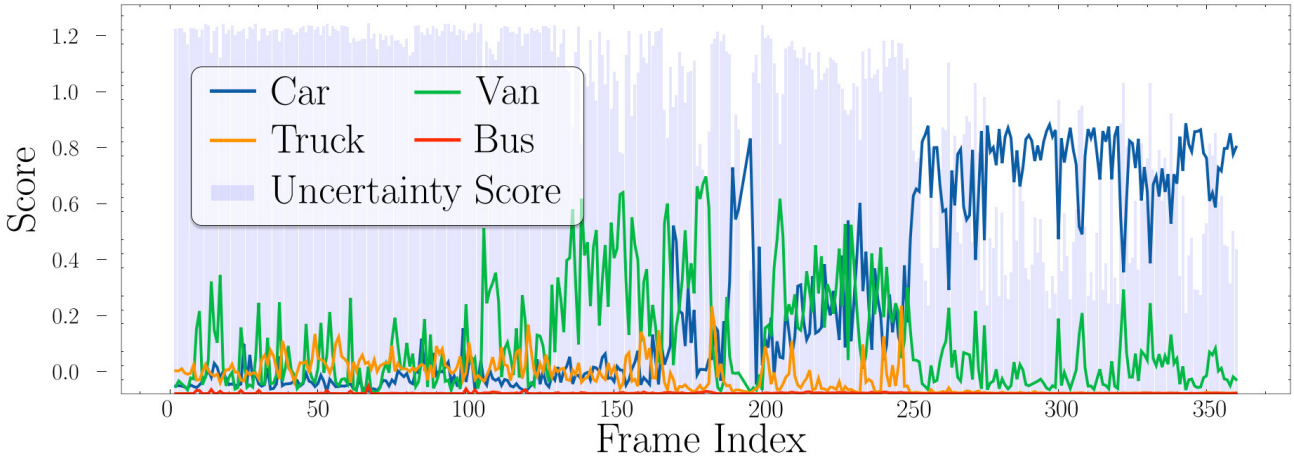

Fig. 3. Class probability scores and uncertainty scores change in time. On the x-axis is the frame index and on the y-axis is the probability (uncertainty) score. This object is depicted in Fig. 1 and the true class of this object is the 'Car'.

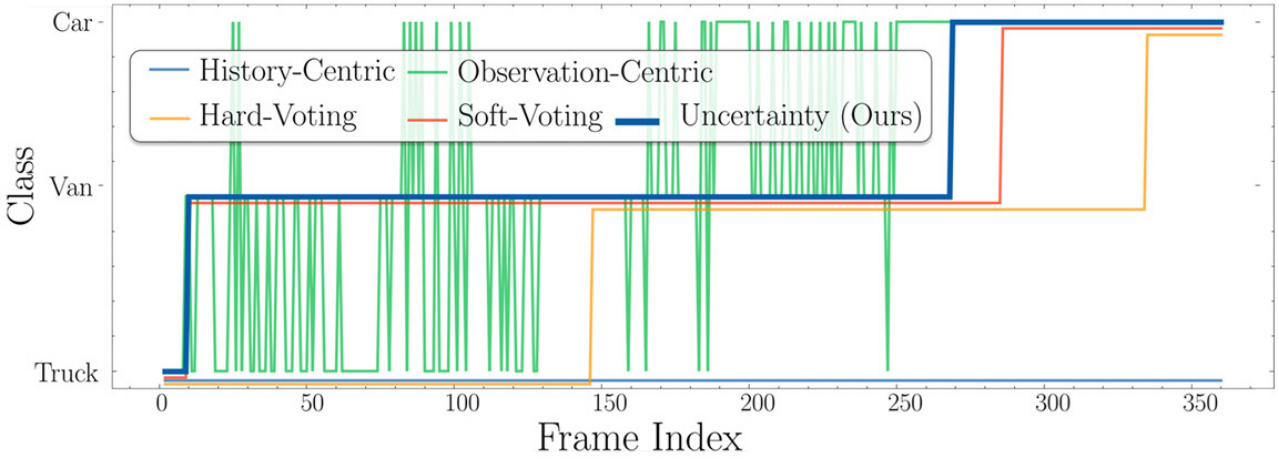

Fig. 4. The track classification results of the same object in Fig. 3. The true class of this object is the 'Car'.

https://doi.org/10.7780/kjrs.2024.40.1.11

object gradually approaches the camera, increasing in size (frame 250 to 350). As a result, it is feasible to distinguish between a car and a van, and the class distribution has a high probability for the car class. With increased car class confidence, the uncertainty has a lower score compared to the initial state. By considering this uncertainty score as a weight for the class update method in multi-class MOT, it can achieve faster and more adaptive target classification compared to other methods.

The graph of the result of each class update method in the same object is shown in Fig. 4. The history-centric update method is constant since it has been assigned to a certain class. As a result, there is a high probability of misclassification. The observation-centric update method predicts the class based on the detection result of each frame. In other words, the highest classification probability of each frame is selected as a track class. This method can immediately adapt as the object becomes recognizable, but it is unstable, and incorrect classification would occur as soon as it is unrecognizable again. Both hard-voting and soft-voting methods have the potential to adaptively predict the class of the track, but they have limitations. The hard-voting method depends on the number of votes; in other words, the number of frames. The more incorrectly classified frames that are observed, the longer it takes to recover the correct class.

In addition, for each frame, hard voting only considers the highest probability class and ignores the other classes (even though they are similar but smaller). Soft-voting considers all these class probabilities, allowing it to consider all classes. However, its weight, which is the objectness score, is not appropriate for the adaptive class update module as it has nothing to do with class distinction. Our uncertainty-based class update module demonstrated enhanced performance in predicting track class adaptively compared to the other methods. Our method correctly classifies the tracking object 18 frames faster than the soft-voting method, which is 750 ms for the 24FPS input. Considering the neglectable computation overhead of our method, it is worth for the surveillance field which needs prompt respond from the threat.

4. Conclusions

In this paper, we propose a class update module concerning class uncertainty to improve the classification ability of the multi-class multi-object tracker. This module enables the estimation of class uncertainty from the object detector. The tracker only reflects the detection results for each frame to the extent that it trusts. Uncertain objects at a distance are less reliable, but when these objects become more trustworthy, our module is capable of rapid adaptation to this change. We demonstrated that our method improves multi-class MOT performance in the VisDrone-MOT dataset. The main advantage of using uncertainty in tracking is enabling the tracking of objects in extensive fields such as remote sensing even when they exceed the model's capabilities by reflecting associated uncertainties. In the future, this approach can be applied to tasks requiring the reliability of inferences from models such as multi-view or multi-modal data.

Acknowledgments

This work was supported by the Agency for Defense Development of the Korean government.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

Aharon, N., Orfaig, R., and Bobrovsky, B. Z., 2022. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*. https://doi.org/10.48550/arXiv.2206.14651

Araujo, P., Fontinele, J., and Oliveira, L., 2019. Multi-perspective object detection for remote criminal analysis using drones. *IEEE Geoscience and Remote Sensing Letters, 17*(7), 1283–1286. https://doi.org/10.1155/2008/246309

Bernardin, K., and Stiefelhagen, R., 2008. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing, 2008*, Article ID 246309.

Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B., 2016. Simple online and realtime tracking. In *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA, Sept. 25–28, pp. 3464–3468. https://doi.org/10.1109/ICIP.2016.7533003

Bishop, C., 2006. *Pattern recognition and machine learning*. Springer.

Chen, G., Wang, W., He, Z., Wang, L., Yuan, Y., Zhang, D. et al., 2021. VisDrone-MOT2021: The vision meets drone multiple object tracking challenge results. In *Proceedings*

of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, Oct. 11–17, pp. 2839–2846. https://doi.org/10.1109/ICCVW54120.2021.00318

Du, Y., Wan, J., Zhao, Y., Zhang, B., Tong, Z., and Dong, J. 2021. GIAOTracker: A comprehensive framework for MCMOT with global information and optimizing strategies in VisDrone 2021. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Montreal, BC, Canada, Oct. 11–17, pp. 2809–2819. https://doi.org/10.1109/ICCVW54120.2021.00315

Fischer, T., Huang, T. E., Pang, J., Qiu, L., Chen, H., Darrell, T. et al., 2023. QDTrack: Quasi-dense similarity learning for appearance-only multiple object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(12), 15380–15393. https://doi.org/10.1109/TPAMI.2023.3301975

He, L., Liu, W., Liang, J., Zheng, K., Liao, X., Cheng, P. et al., 2021. Semi-supervised domain generalizable person re-identification. *arXiv preprint arXiv:2108.05045*. https://doi.org/10.48550/arXiv.2108.05045

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Michael, K. et al., 2022. *Ultralytics/YOLOv5: v7.0 – YOLOv5 SOTA realtime instance segmentation*. Zenodo.

Jocher, G., Chaurasia, A., and Qiu, J., 2023. Ultralytics YOLO (Version 8.0.0) [Computer software]. Available online: https://github.com/ultralytics/ultralytics (accessed on Feb. 26, 2024).

Joshi, A. J., Porikli, F., and Papanikolopoulos, N., 2009. Multi-class active learning for image classification. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, June 20–25, pp. 2372–2379.

Kalman, R. E., 1960. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, 82*(1), 35–45. https://doi.org/10.1115/1.3662552

Kuhn, H. W., 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly, 2*(1–2), 83–97. https://doi.org/10.1002/nav.3800020109

Luo, H., Gu, Y., Liao, X., Lai, S., and Jiang, W., 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Long Beach, CA, USA, June 16–20.

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., and Schindler, K., 2016. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*. https://doi.org/10.48550/arXiv.1603.00831

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A., 2016. You only look once: Unified, real-time object detection. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 27–30, pp. 779–788. https://doi.org/10.1109/CVPR.2016.91

Wang, J., Yang, W., Guo, H., Zhang, R., and Xia, G. S., 2021. Tiny object detection in aerial images. In *Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, Jan. 10–15, pp. 3791–3798. https://doi.org/10.1109/ICPR48806.2021.9413340

Wang, Z., Zheng, L., Liu, Y., Li, Y., and Wang, S., 2020. Towards real-time multi-object tracking. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. M. (eds.), *Computer vision – ECCV 2020*, Springer, pp. 107–122. https://doi.org/10.1007/978-3-030-58621-8_7

Wojke, N., Bewley, A., and Paulus, D., 2017. Simple online and realtime tracking with a deep association metric. In *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP)*, Beijing, China, Sept. 17–20, pp. 3645–3649. https://doi.org/10.1109/ICIP.2017.8296962

Zhang, Y., Wang, C., Wang, X., Zeng, W., and Liu, W., 2021. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision, 129*, 3069–3087. https://doi.org/10.1007/s11263-021-01513-4

Zhao, M., and Zhao, K., 2023. A self-adaptive object detection network for aerial images based on feature enhancement. *IEEE Geoscience and Remote Sensing Letters, 20*, 1–5. https://doi.org/10.1109/LGRS.2023.3291572

Zhou, X., Koltun, V., and Krähenbühl, P., 2020. Tracking objects as points. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. M. (eds.), *Computer vision – ECCV 2020*, Springer, pp. 474–490. https://doi.org/10.1007/978-3-030-58548-8_28

Zhu, J., Chen, X., Zhang, H., Tan, Z., Wang, S., and Ma, H., 2023. Transformer based remote sensing object detection with enhanced multispectral feature extraction. *IEEE Geoscience and Remote Sensing Letters, 20*, 1–5. https://doi.org/10.1109/LGRS.2023.3276052