Journal of Database Management
Volume 34 • Issue 3

Video Object Counting With Scene-Aware Multi-Object Tracking

Yongdong Li, Guangdong Industry Polytechnic, China
Liang Qu, North China Sea Environmental Monitoring Center, State Oceanic Administration, China
Guiyan Cai, Guangzhou Medical University, China
Guoan Cheng, School of Information and Electrical Engineering, Qingdao Harbour Vocational and Technical College, China
Long Qian, School of Computer Science and Technology, Ocean University of China, China
https://orcid.org/0000-0002-7548-3743
Yuling Dou, School of Computer Science and Technology, Ocean University of China, China
Fengqin Yao, School of Computer Science and Technology, Ocean University of China, China
Shengke Wang, School of Computer Science and Technology, Ocean University of China, China*

ABSTRACT

The critical challenge of video object counting is to avoid counting the same object multiple times in different frames. By comparing the appearance and motion feature information of the detection results, the authors use the multi-object tracking method to assign an independent ID number to each object. From the time the ID tag is obtained until the end of the video, each object is counted only once. However, even minor amounts of image noise can cause irreversible changes in feature information, resulting in severe tracking drifts. This paper introduces the concept of scene awareness and addresses unreasonable ID assignment caused by unreliable feature matching in the context of region division. Through the macro analysis of the scene, the authors define the region (called the transition region) where the number of objects can increase or decrease and require that all ID assignments for new objects and ID deletions for existing objects take place only in the transition region. Because the actual number of objects in the non-transition region is constant, they rematch unmatched objects with existing IDs in the region (called ID relocation) because changes in object ID are caused by feature matching failure. In this paper, the authors create algorithms for dynamically generating transition regions, detecting object increases and decreases, and relocating object IDs. Experimental results show that the method effectively improves the accuracy of video object counting.

KEYWORDS

Multi-Object Tracking, Region Division, Scene-Aware, Video Object Counting

INTRODUCTION

With the development of object detection algorithms (Lin et al., 2017; Ayoun et al., 2010; Li et al., 2017; Zheng et al., 2019) and the innovation of reidentification technology (Wu, Wu et al., 2018; Lu et al., 2019; Zhang et al., 2019; Wu et al. 2018; Farenzena et al., 2010; Choe et al. 2019), tracking-by-

DOI: 10.4018/JDM.321553 *Corresponding Author

detection has become one of the mainstream approaches for video analysis problems. It has been widely studied and has produced remarkable results in recent decades, demonstrating significant application value in vehicle navigation, intelligent monitoring, human-computer interaction, crowd counting, and other fields. We design a video object counting algorithm based on multi-object tracking based on the unique correspondence between the object and the ID number in the tracking algorithm. However, due to the presence of many uncertain factors in real-world scenes, such as occlusion, lighting change, scale change, and camera motion, maintaining the algorithm's accuracy and robustness remains a difficult task.

In recent years, most studies have concentrated on obtaining more abundant and discriminative characteristics to improve tracker performance. Feature information is also regarded as the primary basis for determining the identity of the object. However, in practical applications, the object feature is not always dependable, and even minor interference can cause the feature information to be completely modified. A slight image jitter, particularly in the field of UAVs, can cause changes in the apparent and moving characteristics of all the objects in the current scene, and even if we can accurately obtain the features of these mutations, it is difficult to relate them to previously obtained feature information. As shown in Figure 1, feature information is highly susceptible to external manipulation, and it is simple to generate redundant IDs, resulting in a considerably greater number of items counted by the algorithm than there are real objects. More significantly, tracking errors are compounding. If the ID inaccuracy in each frame increases, our counting findings will be severely affected if we run a video sequence at 20 frames per second for less than 10 seconds. As a result, raising the quality of object features alone will not substantially alter the change in object ID, allowing for strong multi-object tracking and counting. In fact, the present multi-object tracking algorithm's detection and feature extraction accuracy is already high, but it is insufficient for work in complicated contexts. As a result, it is critical to gather scene information, which is a critical step in dealing with tracking-based counting challenges. We cannot guarantee that the object always matches its original ID, but we can find the region with the same total number of objects. Even if there is mistake matching, the result of our count will not be affected as long as the number of IDs in this region remains constant.

The feature information is only adequate for the scene when viewed from a stable observation perspective, and its anti-interference capabilities is frequently insufficient in practice. As a result, we chose to provide new reference information to improve the network's robustness, which must be

Figure 1.
The necessity of scene awareness. (1) Occlusion and deformation result in the complete change of the object's appearance information and the loss of the original identity, which eventually leads to the statistical change of the number of objects. (2) We use the scene-aware method to determine the region where no object enters or leaves. Limit the increase in redundant IDs due to feature changes in this area to avoid an object being counted twice

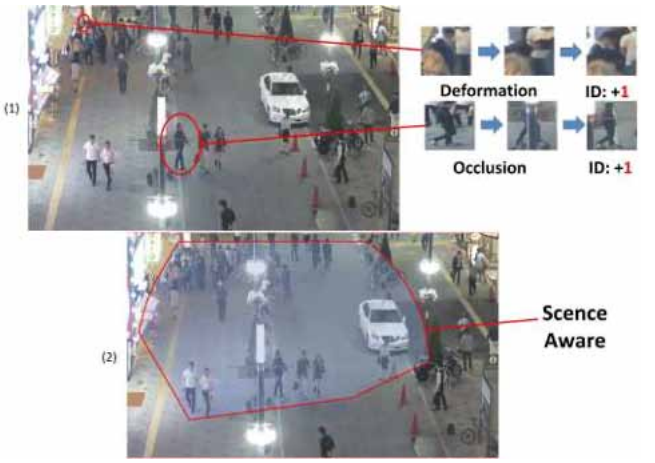
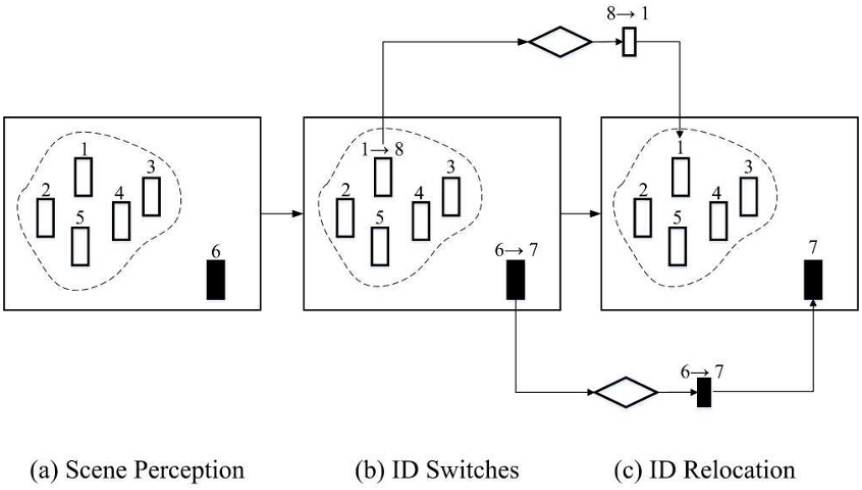


Figure 2.
(a) We use Gaussian filtering and other methods to cluster all the objects. White boxes represent the objects in the non-transition region, and black boxes represent the objects in the transition region. (b) We deal with the object with an ID switch according to its region. (c) We allow the number growth of IDs in the transition region and retrieve lost IDs for objects in the non-transition region



tightly tied to the tracking process, dependable, and simple to access. By studying the working scene, we can always discover a zone where the internal items are known and, in most circumstances, the overall number of objects remains constant. In a setting like a corridor, pedestrian entry and exit only happen near the entrance or exit. Regardless of whether the passageway is blocked, the quantity and identities of pedestrians remain constant. We call the region similar to the interior of the corridor the non-transition region, and the rest is the transition region. By restricting the number of IDs in non-transition regions, we may bring the number of effective tracks closer to the number of genuine objects, reducing interference between new and known items. Figure 2 depicts an overview of our method.

In this paper, we take full advantage of scene information to tackle unreliable feature information. The following three points are the key contributions of this study. First, we investigate why object features are unreliable and propose incorporating scene information into the tracking method via region division. Second, we examine region division and object region attribution methods in various contexts, and we construct correlation methods for transition and non-transition regions. Finally, we apply our method to various image sequences, and the experimental results demonstrate that our method efficiently decreases the number of redundant IDs while also improving the counting algorithm's accuracy.

RELATED WORK

Multi-object tracking is one of the research hot spots in the field of visual tracking. In this section, we briefly introduce several tracing algorithms focusing on different directions and the trend of multi-object tracking.

Sort (Bewley et al., 2016) proposes a "two-stage" algorithm framework of detection and tracking and improves the running speed of detection and tracking to the level of industrial application. In the detection phase, sort uses other detection algorithms to process the video sequence and obtain the detection results. In the tracking phase, sort uses the Kalman filtering algorithm to predict the position of the object in the next frame and determines the identity of the object by calculate the IOU distance between the prediction and the detection. This simple and ingenious design enables it to run at a speed of 200 FPS and provides directions for the follow-up tracking algorithm.

Deepsort (Wojke et al., 2017) is improved on the basis of Sort. Introducing the depth feature information in the matching process greatly reduces the number of identity switches, improves the tracking accuracy, and maintains a fast speed. Deepsort provides a simple and feasible scheme for tracking multiple objects at the same time.

JDE (Wang et al., 2019) realizes the transformation of object tracking from 'two-stage' to end-to-end. In the process of detection, JDE synchronously carries out feature extraction and data association and completes detection and tracking tasks in a network, which truly meets the requirements of real-time tracking. Most of the latest multi-object tracking methods adopt this method.

Fairmot (Zhan et al., 2020) is the latest achievement of end-to-end multi-object tracking, and it is also a breakthrough in JDE performance. It analyzes the internal cause of tracking failure and points out that the anchor used in the detection phase is an important factor causing ambiguity. By combining the anchor free object detection algorithm with the lightweight re ID, a heuristic and helpful evaluation baseline is proposed.

METHODOLOGY

Tracking-by-detection is usually regarded as the combination of object detection and Re-ID. We integrate and improve the tracking process and introduce scene awareness by region division and object clustering. Our framework consists of four consecutive tasks: first, object detection and feature extraction of video sequences; second, region division of the whole scene; third, clustering analysis for all objects; and fourth, data association between object and ID number based on feature information. In this section, we introduce the details of the design concept and workflow of our algorithm.

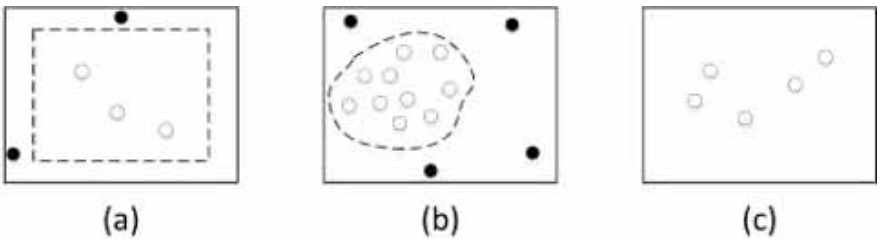
Detection and Feature Extraction

The first stage in tracking is detection, which serves as the foundation for all following work. In practice, mistakes due to false detection and missing detection build over time, and the objects are typically densely dispersed and extremely similar. This establishes two requirements for the tracking detection algorithm: one is to detect all objects as much as feasible, and the other is to effectively eliminate overlap between detection boxes. The classic anchor-based detection algorithm readily creates the occurrence of a detection bounding box including many objects, which is unsuitable for current tracking development. In the tradition of CenterNet, we display the item by determining its center point, then change the detection task into a typical key point estimation problem and effectively handle the above challenges. The goal of feature extraction is to establish a foundation for subsequent data association. Two distinct models, known as two-stage tracking, perform detection and feature extraction in conventional tracking. Two-stage tracking is not only inefficient in model training, but the algorithm's operation mode prevents real-time tracking from being realized. We employ the most recent end-to-end approach to exchange the object's feature information, concurrently detecting and extracting features in a single workflow, substantially simplifying the tracking procedure.

Scene-Aware

The goal of the multi-object tracking task is to assign different numbers (IDs) to different objects in an image sequence in order to obtain the track set of those objects. Existing multi-object tracking algorithms consider all objects that do not match any feature to be emerging objects and assign new ID numbers to them. This concept raises two serious issues. One is that crowded scenes frequently result in the phenomenon of one object continuously corresponding to multiple IDs. The new objects can also appear anywhere in the image, which is obviously illogical. As a result, the image region must be divided based on the scene's characteristics. To divide the field of vision into different regions, we must first identify the clear border between the regions. Figure 3 depicts the common work scene and its regional division, named scene-aware module. The white points represent non-transition objects, the black points represent transition objects, the light blue border represents the scope of the field of

Figure 3.
Scene-aware module



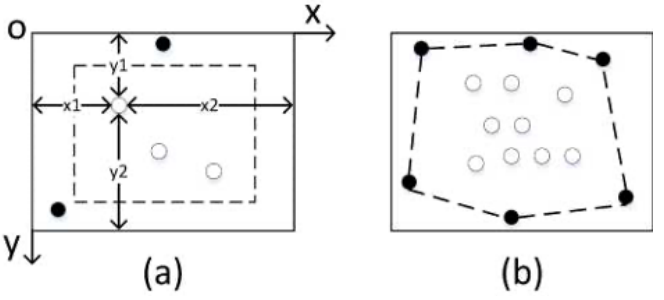
vision, and the orange border represents the border between the transition and non-transition regions. Scene (a) is the region division under the vision of the UAV. In a moving top view, new objects can only enter or leave from the edge of the field of view, so the transition regions are distributed along the edge of the field of view. scene (b) is the camera field of view under the environment of streets and scenic spots. In this kind of environment, because of the angle of the camera and the environment occlusion, the new object does not necessarily appear from the boundary of the field of view. In view of the relatively dense objects in this scene, we regard the outermost edge of the crowd as the boundary of the transition region. scene (c) is the field of view of a surveillance camera set up in a closed area such as a classroom. The walls, the doors and windows of this kind of scene form a natural border. During class time, we can default the whole field of vision as a no-transition region.

Object Clustering

In the actual work scene, it is difficult to keep all the object IDs unchanged due to the interference of object occlusion, intense motion, missed detection and the failure of the motion model. However, we can make the total number of IDs closer to the real number of objects by constraining the number of IDs of some objects unchanged. The significance of object clustering is to analyze the distribution and movement of the object, determine the regional attribution of the object, and then determine the object group to be constrained.

In general, we can determine the area of the object according to the distance between the center of the object and the natural edge of the field of view. If the distance is less than a specific threshold, it is regarded as located in the transition area. However, in practical applications, the ratio between the size of the detection frame and the size of the field of view is not fixed, and the size of the detection frames is not necessarily the same that is, a dynamic threshold is needed to determine the border. To solve this problem, we take the width W and the height H of the detection box as the threshold of region division. As shown in the figure, if the center point (x, y) of the detection frame meets any of the following conditions: x1 >W, x2 >W, yl >H, y2 >H, then it can be determined that the detection object is located

Figure 4.
Determining the object region



in the non-transition region. In the case of relatively dense objects, we can construct a set with the center points of all detection frames and perform convex hull operations on this set. If the center of an object appears in the interior of the convex hull, it can be determined that it is in the non-transition region.

ID Assignment

The process of determining one's identity is divided into two stages. The first is to associate all of the objects in the image with the IDs, and the second is to relocate the objects in the non-transition region that are irrelevant to any ID.

Data association is equivalent to a lightweight re-ID network, which is the core content of the tracking task. In the data association stage, we first use the Kalman filter to generate the track according to the coordinate information of the object detection box, that is, to predict the position and size of the tracking object in the next frame image. Then, we calculate the cosine distance of the apparent features between the predicted trajectory and the object to obtain a group of similar trajectories and objects. Then, the cosine distance of the appearance information between the predicted trajectory and the object is calculated to obtain a group of similar trajectories and objects. Finally, delete the track object pairs that are far away from each other in space and make the remaining objects inherit the track ID number.

Due to the existence of overexposure, camera jitter, illumination change and other factors, it is impossible to guarantee the same quality of the object in each position in the captured image. At the same time, the detection algorithm will inevitably have false detection and missing detection. Therefore, it is difficult to keep the object ID unchanged, and it is meaningful to retrieve the missing ID. In the scene with a sparse object distribution, the proximity rule is a simple and efficient ID relocation method. Instead of assigning a new ID number to an object that has lost its own ID, we make it inherit the ID of the track closest to it. After inheriting the ID, we delete this track to avoid interference with subsequent matches. In the scene with a dense distribution of objects, we restrict the generation of new tracks in the non-transition region. At the end of each frame association, we perform the data association again for the unmatched objects and tracks in this region.

EXPERIMENTS

Datasets and Metrics

For MOT Challenge task, in order to evaluate the performance of our scene-aware method, we conduct 4 groups of comparative experiments on the MOT20 (Lu et al., 2019) dataset, which is the latest benchmark on MOT Challenge. This dataset contains 8 challenging video sequences captured in unconstrained scenes. We sorted 4 groups of video sequences in the MOT20 training set as our verification set. We strictly follow the mot challenge benchmark to evaluate the tracking performance of the algorithms.

In addition, we add the comparative and ablation experiments on the VisDrone dataset. 10 sets of image sequences with different image scales and video lengths are selected as the training set and 5 sets of image sequences as the test set for the tracking algorithm in this dataset. Furthermore, we use the evaluation metric of MOT challenge to analysis the results.

MOTA (Multiple Object Tracking Accuracy): tracking accuracy, which is the most important performance metric for evaluating tracking algorithms. MOTA is calculated as follows:

$$MOTA = 1 - \frac{\sum_t \left(FP_t + FN_t + IDs_t\right)}{\sum_t GT_t}$$

where FP_t denotes the number of False Positive targets at time t of tracking start, FN_t denotes the number of False Negative targets at time t, IDs_t denotes the number of switching target IDs at time t,

and GT_t denotes the number of real targets in the image at time t. IDs, ID Switches, identity switching, which refers to the total number of times the target's ID number has changed. IDF1, ID F1 Score, a combination of accuracy and recall to measure the accuracy of ID information in the trajectory.

For the object counting task, we consider all detection results as pedestrian targets, and calculate the number of people obtained by counting the number of errors with the labeled data.

The average error in real-time people counting, MAE_{num}, which describes the absolute error between the real-time people counted by the algorithm and the real number of people:

$$MAE_{num} = \frac{\sum_{i=1}^{n} \left| num_i - gt_i \right|}{n}$$

where n is the length of the video sequence, num_i is the statistical value in real-time at frame i, and gt_i is the real number of people.

Cumulative counting accuracy A_{total}, which is the ratio of the cumulative object number counted by the algorithm to the real total number of person gt:

$$A_{total} = \frac{total}{gt} * 100\%$$

Cumulative counting average error MAE_{total}, which describes the rate of accumulation of the counting deviation of the algorithm for the total number of persons over time:

$$MAE_{total} = \frac{\sum_{i=1}^{n} \left| total_i - gt_i \right|}{n}$$

where n is the length of video sequence, $total_i$ is the statistical value in frame i of the cumulative counting.

In addition, we also introduce two new metrics, counting precision and divergence speed (DIVS), to test the counting ability. Counting precision is defined as follows: where GT (ground truth) is the number of test images and SQ is the statistical counting number of people by the algorithm:

$$Precision = 1 - \frac{\left| GT - SQ \right|}{GT}$$

Different from single image counting, the deviation of video object counting will accumulate over time, so the influence of video sequence length on the counting effect cannot be measured only by precision. We propose the concept of the divergence rate (DIVR) of counting deviation:

$$DIVR = \frac{1}{GT} \cdot \frac{\lambda \left| GT - SQ \right|}{\sum_{1}^{n} t \cdot f}$$

By calculating the ratio of the accumulated deviation value per frame to the real number, we obtain the ability of the algorithm to maintain the counting precision in a period of time. We set n as the number of different videos used in the test, t as the video duration, f is frame rate. The lower the value of DIVR is, the higher the performance. If the algorithm has no deviation, DIVR is 0.

Implementation Details

We employ CenterNet as the backbone of our method. In dense scenes, the center of the object is occluded seriously, and the features do not have strong robustness. We raise the learning center point by 0.22 positions (from the abdomen to approximately the chest to the shoulder), which has a certain anti-occlusion ability. The proportion of learning places was 0.23:0.77. To test the performance difference of the algorithm fairly, we use the ctdet coco dla 2x.pth of CenterNet on the coco2017 (Lin et al., 2014) dataset as the pretraining weight of the detection algorithms and retrain on MOT17 to learn the center point of the object. We set the dimension of the Re-ID branch to 128, selected the cross entropy loss, and adjusted the size of the image to 1088608 as the input of the network. We train networks for comparison testing on 4 GTX2080Ti with a learning rate of 1e-4 and a batch size of 28 for 30 epochs. The detection box was filtered to a certain extent, and the threshold value of NMS was 0.4. We filter out results where the area size is less than 100 or the ratio of height to width is less than 1.6.

Analysis on Validation Set

To test the effectiveness of the scene-aware strategy (α), we set up 4 groups of contrast tests, independently comparing the upshift of the center point the impact of scene awareness on tracking and counting. The tracking performance of the algorithm is shown in Table 1. Our scene-aware method (ours+ α) is efficient in reducing ID switches and improving IDF1 (Ristani et al., 2016), which has been successful in limiting redundant ID growth and maintaining good tracking performance. The counting performance of the algorithm is shown in Table 2. Our scene-aware method achieves the most accurate counting results on all video sequences, which proves that our counting strategy is effective. Our DIVS is much lower than other methods, which reflects the important value of the scene-aware strategy in long-term counting.

We also experimented on the visdrone dataset and added the FPS parameter to measure the computational speed. As shown in Table 3.

In this paper, ablation experiments on the scene-aware module denoted as **α** were conducted on five sets of video sequences selected from the VisDrone dataset.

As shown in Table 4, the average error MAE_{num} of real-time person count relies mainly on the detection capability of the algorithm for the target, and its value does not change after the addition of the sensing module. The cumulative person count accuracy A_{total}, on the other hand, relies on the association capability of the algorithm. Even if the detection results of the target are obtained more accurately, it is still a challenging task to make an accurate association of the target. It can be seen

Table 1.
Tracking results on MOT20

Tracker	MOTA(%)↑	MOTP(%)↑	IDF1(%)↑	IDR(%)↑	MT(%)↑	ML(%)↓	FP↓	FN↓	IDS↓
JDE	43.7	73.1	36.4	28.7	13.6	22.7	**70825**	553025	14603
Fairmot	44	**76.8**	44.1	35.7	20	23.7	96124	530640	8656
Ours	**45**	75.9	45.2	**37.5**	**20.8**	**22.7**	114627	**500598**	9289
Ours+ α	43	75.8	**45.9**	**37.5**	19.5	23.7	109544	525927	**7694**

Table 2
Counting results on MOT20

	Precision%↑	DIVR%↓
JDE	19.28%	46.90%
Fairmot	17.21%	53.98%
Ours	17.60%	52.46%
Ours+ α	**28.07%**	**28.70%**

Figure 5.
Visualization results display. We show the counting results of JDE, fairmot and our algorithm at frames 1, 214 and 428 on MOT20-01

Table 3.
Tracking results on VisDrone

	MOTA↑	IDF1↑	IDs↓	FPS↑
JDE[44]	59.0	52.7	1247	15.6
FairMOT[45]	60.1	55.5	960	**16.0**
Ours	61.3	57.3	910	**16.0**
Ours(α)	**62.3**	**57.5**	**901**	**16.0**

that the counting accuracy is positively correlated with the tracking accuracy, but all tracking algorithms assign more than twice the number of IDs in the video sequence than the actual number

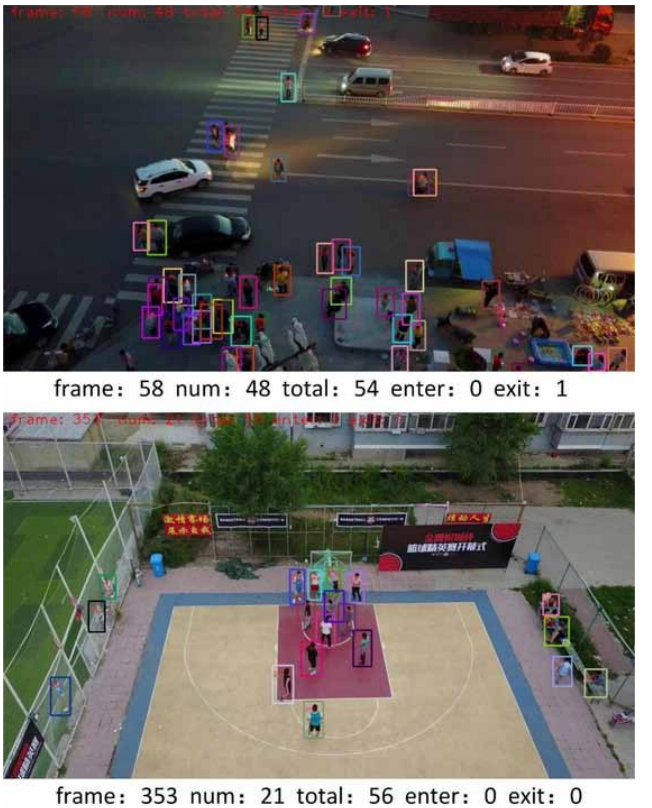
of targets, all tracking algorithms have A_{total} over 200%, so it is difficult to complete the person counting task in the UAV scenario by using a single tracking algorithm, and it must be combined with other auxiliary algorithm modules to reduce the impact of redundant IDs on the counting.

After adding the perception module of this paper, the counting results of all algorithms are closer to the real number of people (A_{total} is close to 100%), while the counting algorithm Ours + α of this paper achieves optimal results in all counting metrics and has a running speed of 16.0 FPS to meet the demand of real-time people counting in VisDrone scenarios. Some of the experimental results of the counting algorithm Ours + α on VisDrone are shown in Figures 6.

Table 4.
Ablation study on VisDrone dataset

	MAE_{num}↓	MAE Total↓	A_{total}	FPS↑
JDE[44]	6.9	-	231.0%	15.6
FairMOT[45]	5.8	-	214.6%	16.0
Ours	**5.1**	**-**	204.6%	**16.0**
JDE +α	6.9	7.0	90.6%	15.6
FairMOT +α	5.8	6.4	92.4%	**16.0**
Ours+α	**5.1**	**5.7**	**93.2%**	**16.0**

Figure 6.
A demonstration of the effect of people counting on the VisDrone dataset



CONCLUSION

In this paper, we propose a new video object counting algorithm with scene-aware module, which takes full advantage of object features and the macro scene information and effectively reduces the problem of object loss caused by feature change and improves the robustness of the counting algorithm. Object features contain precise information and detailed descriptions that aid in determining the objects precise location, size, and ID number. Scene information is unaffected by local changes in objects, and it can help to retrieve the lost IDs of objects when feature matching fails. Experimental results show that this method can significantly reduce the loss of object ID and improve the accuracy of counting. The idea of scene-aware module proposed can be used as a new processing mechanism to deal with image jitters and illumination changes. It can be combined with other multi-object tracking algorithms to deal with the complex problem of sharp changes in the appearance and scale of the object, and it is also possible to choose a new method baseline for the experiment, which is also the focus of our future study.

ACKNOWLEDGMENT

This work was partially supported by the Research Foundation of Guangdong Industry Polytechnic under Grant No. KJ2021-18 (Research on UAV Smart Airport Cluster Management Technologies), the National Key Research and Development Program of China under Grant No. 2018AAA0100400, the Natural Science Foundation of Shandong Province under Grants No. ZR2020MF131 and No. ZR2021ZD19, and the Science and Technology Program of Qingdao under Grant No. 21-1-4-ny-19-nsh.

REFERENCES

Ayoun, A., & Smets, P. (2010). Data association in multitarget detection using the transferable belief model. *International Journal of Intelligent Systems*, *16*(10), 1167–1182. doi:10.1002/int.1054

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). *Simple online and realtime tracking*. IEEE. doi:10.1109/ICIP.2016.7533003

Choe, C., Choe, G., Wang, T., Han, S., & Yuan, C. (2019). Deep feature learning with mixed distance maximization for person reidentification. *Multimedia Tools and Applications*, 2.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., & Cristani, M. (2010). Person reidentification by symmetry-driven accumulation of local features. Proc IEEE Conference on Computer Vision Pattern Recognition, 2360–2367.

Lin, T., Pintado, F., Corchado, J. M., & Bajo, J. (2017). Multisource homogeneous data clustering for multitarget detection from cluttered background with misdetection. *Applied Soft Computing*, *60*, 436–446. doi:10.1016/j.asoc.2017.07.012

Lin, T. Y., Dollar, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). *Feature pyramid networks for object detection*. IEEE Computer Society. doi:10.1109/CVPR.2017.106

Lin, T. Y., Maire, M., Belongie, S., Hays, J., & Zitnick, C. L. (2014). *Microsoft coco: Common objects in context*. Springer International Publishing.

Lu, J., He, Y., Liu, T., & Chen, X. (2019). Centralized and clustered features for person reidentification. *IEEE Signal Processing Letters*, *26*(6), 933–937. doi:10.1109/LSP.2019.2913020

Ristani, E., Solera, F., Zou, R. S., Cucchiara, R., & Tomasi, C. (2016). *Performance measures and a data set for multi-target, multi-camera tracking*. Springer. doi:10.1007/978-3-319-48881-3_2

Wang, Z., Zheng, L., Liu, Y., & Wang, S. (2019). *Towards real-time multi-object tracking*. Academic Press.

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple Online and Realtime Tracking with a Deep Association Metric. *2017 IEEE International Conference on Image Processing (ICIP)*, 3645–3649. doi:10.1109/ICIP.2017.8296962

Wu, D., Zhang, K., Cheng, F., Zhao, Y., Liu, Q., Yuan, C. A. & Huang, D. S. (2018). *Random occlusion-recovery for person reidentification*. Academic Press.

Wu, D., Zheng, S. J., Yuan, C. A., & Huang, D. S. (2018). A deep model with combined losses for person reidentification. *Cognitive Systems Research*.

Zhan, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2020). *A simple baseline for multi-object tracking*. Academic Press.

Zhang, Z., Huang, M., Liu, S., Xiao, B., & Durrani, T. (2019). Fuzzy Multilayer clustering and fuzzy label regularization for unsupervised person reidentification. *IEEE Transactions on Fuzzy Systems*, (99), 1–1.

Zheng, Y., Zhang, X., Wang, F., Cao, T., Sun, M., & Wang, X. (2019). Detection of people with camouflage patterns via a dense deconvolution network. *IEEE Signal Processing Letters*, *26*(1), 29–33. doi:10.1109/LSP.2018.2825959

Guiyan Cai graduated from south China university of technology, deparment of computer science and application. Research interests include computer application technology. Now they work in Guangzhou Medical University, School of Biomedical Engineering.

Yuling Dou received her bachelor's degree in network engineering from Qufu Normal University in 2021. In 2021, she became a graduate student at Ocean University of China. At present, she is still a postgraduate student at Ocean University of China. She is certified as a Senior Network Engineer and has received several competitive MEDALS. Her research and exploration in the field of computer vision, and her efforts to learn the relevant scientific and cultural knowledge of this major. Through the guidance of her tutor and her own hard work, she has preliminarily mastered solid professional knowledge and can flexibly apply it to practical work.

Fengqin Yao acquired the B.S. degree in Digital Media Technology from Shandong University of Finance and Economics in 2019. Currently, she is pursuing a Ph.D. degree in School of Computer Science and Technology of Ocean University of China. Her research interests include Computer Vision, Image Processing and Robotics.

Shengke Wang is an associate professor in the computer science and technology department of the Ocean University of China. He acquired the B.S. degree in Computer Science from University of Jinan, China, in 2000 and received his Ph.D. in Computer Science from South China University of Technology, China, in 2005. His research interests include computer vison, machine learning, and image processing and document image analysis.