Using YOLOv5 with OC-SORT for MOT in Thermal Videos (CV-C0)

Ekaterina Sysoykova
JKU
Linz, Austria
k11842415

Kevin Serrano
JKU
Linz, Austria
k12215919

Marco Badici
JKU
Linz, Austria
k11804695

Dimitrios Koletsis
JKU
Linz, Austria
k12216668

Lukas Renoth
JKU
Linz, Austria
k12215903

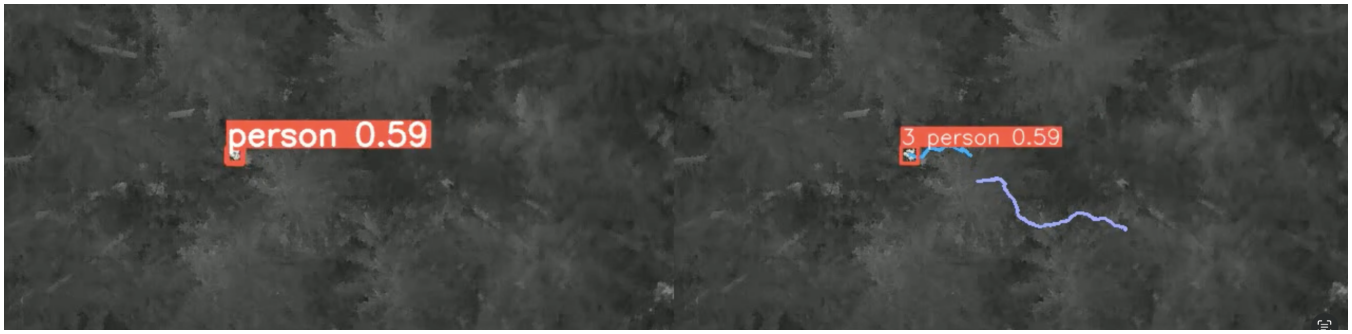

Figure 1. Side-by-side comparison of the YOLOv5 output (left) and the YOLOv5 + OC-SORT tracker output (right).

Abstract

In this project, we aimed to achieve multiple object tracking on thermal videos using a combination of the YOLOv5 network for object detection and the OC-SORT tracker for association. To overcome the issue of using pre-trained models on RGB images, we performed transfer learning on the YOLOv5 network to adapt it to our dataset of thermal grayscale videos. Our results showed that by using the OC-SORT with CIoU association, we were able to achieve acceptable results in terms of both detection and tracking accuracy.

Keywords: datasets, neural network, MOT, thermal imaging

1 Introduction

Multiple object tracking (MOT) is a challenging task in the field of computer vision, where the goal is to track multiple objects in a video sequence. This task is especially challenging when working with thermal videos, as the lack of color information and the presence of noise can make it difficult to accurately detect and track objects.

In this project, we aimed to address this challenge by using a combination of the YOLOv5 [3] network for object detection and the OC-SORT tracker for association. We chose YOLOv5 because it is a state-of-the-art object detection network that has been shown to achieve high accuracy on a variety of datasets containing infrared data [4] [2]. To overcome the issue of using pre-trained models on RGB images, we performed transfer learning on the YOLOv5 network to adapt it to our dataset of thermal grayscale videos.

We chose OC-SORT (Observation-Centric SORT) with CIoU (Complete IOU) for association. It remains simple, online and real-time (SORT) but provides improved robustness over occlusion and non-linear motion [1].

The goal of this report is to present our approach, the results we obtained, and the lessons we learned during the course of this project.

2 Pipeline

Our pipeline is a typical workflow for training and validating supervised DeepLearning methods.

2.1 Data Preparation

For this project, we used a dataset of thermal grayscale videos for multiple object tracking. The dataset consisted of 28 videos ranging between 18 and 80 seconds, where only one of the videos contained 2 objects and the rest only had a single object. The videos were taken by a drone and showed a top perspective of a forest where a person or persons could be seen walking around. The videos were captured at 30 fps.

To prepare the data for the project, we performed annotation on the dataset, the annotation format used was CVAT [6]. This allowed us to exploit interpolation and tracking

capabilities of the CVAT open-source tool in order to speed up the annotation process.

Additionally, we converted the videos to images since YOLOv5 works with images and not videos. We used FiftyOne [5], an open-source tool for building datasets and computer vision models. This tool allowed us to import the CVAT formatted annotations, convert from a video dataset to an image dataset and export it as YOLOv5 format. The videos were sampled at 3fps.

2.2 Detection

For object detection, we used the YOLOv5 network, a state-of-the-art object detection model. However, since the model was pre-trained on RGB images and our dataset consisted of thermal grayscale videos, we performed transfer learning to adapt the model to our specific task.

The transfer learning process involved performing data augmentation and training the detector with a small learning rate. We split the dataset in 5 folds and fine-tuned 5 different models for cross validation.

To evaluate the detection performance, we make use of the following metrics: recall, precision, mAP 0.5 and mAP 0,5:0.95 which me take using the validation set of each cross-validation iteration.

2.3 Tracking

For object tracking, we used the OC-SORT algorithm, a simple but effective online multiple object tracking algorithm. We also used CIoU association which was able to handle occlusions better.

The OC-SORT algorithm uses a probabilistic data association algorithm to match the detections from the current frame with the tracks from the previous frames. The data association algorithm uses a combination of the distance between the detections and the tracks, the motion of the tracks, and the size and shape of the detections to determine the best match. We set the parameters of the algorithm such that we could handle 2-second occlusions if re-identification via CIoU is successful.

3 Results

Fold	Precision	Recall	mAP @ 0.5	mAP @ 0.5:0.95
1	0.82	0.63	0.66	0.23
2	0.66	0.54	0.53	0.15
3	0.77	0.59	0.62	0.20
4	0.65	0.47	0.47	0.13
5	0.42	0.37	0.23	0.06
Mean	**0.66**	**0.52**	**0.50**	**0.15**
Std	**0.14**	**0.09**	**0.15**	**0.06**

Table 1. Cross validation results

Overall, YOLOv5 was able to detect objects in the scene with a decent accuracy and the OC-SORT tracker was able to overcome certain occlusions and filter out false popsitives coming from the YOLOv5 detections on our dataset of thermal grayscale videos.

There is a noticeable variance in the cross validation results due to the fact that the length of each video varies, leading to unbalanced cross-validation datasets where some contain more training data than others, and similarly with validation data.

4 Advantages and Limitations

One of the main advantages of our approach is the use of convolutional neural networks (CNNs) for object detection. CNNs are superior to hand-crafted feature extractors because they are able to learn and automatically extract features from the data. This allows them to achieve higher accuracy and robustness compared to traditional methods.

However, our approach also has some limitations. One limitation is that the training is only as good as the quality of the training data. In our project, we found that our annotations could be improved greatly to further increase the performance of the model.

Another limitation is that, even though YOLOv5 is a real-time object detector, it relies most of the time on a GPU to process videos at 30fps. Therefore, deploying it on a drone might not be straight forward. This is an important consideration when looking to use this approach in real-world scenarios.

Overall, while our approach has several advantages, it also has some limitations that should be taken into account when considering its use in real-world applications

References

[1] Jinkun Cao, Xinshuo Weng, Rawal Khirodkar, Jiangmiao Pang, and Kris Kitani. 2022. Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking. https://doi.org/10.48550/ARXIV.2203.14360

[2] Shuangjiang Du, Baofu Zhang, Pin Zhang, Peng Xiang, Hong Xue, and Yin Zhang. 2021. FA-YOLO: An Improved YOLO Model for Infrared Occlusion Object Detection under Confusing Background. *Wirel. Commun. Mob. Comput.* 2021 (jan 2021), 10 pages. https://doi.org/10.1155/2021/1896029

[3] Glenn Jocher. 2020. *YOLOv5 by Ultralytics.* https://doi.org/10.5281/zenodo.3908559

[4] Mate Krišto, Marina Ivasic-Kos, and Miran Pobar. 2020. Thermal Object Detection in Difficult Weather Conditions Using YOLO. *IEEE Access* 8 (2020), 125459–125476. https://doi.org/10.1109/ACCESS.2020.3007481

[5] B. E. Moore and J. J. Corso. 2020. FiftyOne. *GitHub. Note: https://github.com/voxel51/fiftyone* (2020).

[6] Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov, Priya4607, Rush Tehrani, Seungwon Jeong, Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. 2020. *opencv/cvat: v1.1.0.* https://doi.org/10.5281/zenodo.4009388