Large scale Real-world Multi-Person Tracking

Bing Shuai, Alessandro Bergamo, Uta Buechler
Andrew Berneshawi, Alyssa Boden, Joseph Tighe

AWS AI Labs
https://github.com/amazon-research/tracking-dataset

1 Video Sourcing

1.1 Query keywords

A pre-defined set of search keywords are used to query videos in stock video services (Fillerstock [1], Pexels [2], Pixabay [3]). The complete list of keywords is as follows: 1), People walking in the street; 2), People walking in the mall; 3), People walking in the city; 4), People walking near market; 5), People walking inside mall; 6), People walking inside shops; 7), People crossing the street; 8), People walking in open space. We select this set specifically to query videos that include moving people rather than static people for the interest of tracking applications.

1.2 Videos Manual Selection

This section adds details regarding the video manual selection process introduced in Sec. 3 in the main paper. More in detail, our team of experts selected the videos to be part of the dataset by taking into account the following criterias:

- **Application Aligned.** We look for videos that appear to come from fixed connected home or city level cameras. We discard videos captured from television programs that contain advertisement interruptions, or videos with subtitles or that have been otherwise edited by software.
- **Moving crowds.** We try to strike a balance between (1) avoiding trivial videos with one or two people only, and (2) having over-crowded videos with hundreds of people, which we believe would have posed challenges at annotation time and lowered the ground truth accuracy. In particular, we discard videos with less than 5 people, and videos where more than half of the scene contained crowds, with less than 10% of each person visible. We also favor the selection of videos containing moving people rather than static ones (sitting or standing), given the fact that the latter represents a trivial scenario for tracking.
- **Occlusion.** We also favor the selection of videos containing high degrees of occlusions, where people become occluded for one or more seconds, and then reappear in the scene. This is done in order to promote the development of algorithms that can handle such situations. We include both partial and full occlusions, as well as person-to-person and person-to-object occlusions.

Data source	Number of videos
Fillerstock [1]	101
Meva [4]	16
PathTrack [7]	26
Virat [9]	9
Pexels [2]	78
Pixabay [3]	6
Total	236

Table 1: Composition of our dataset: number of videos per data source

- **Background variability.** Our selection process also ensures diversity in the background scene. For example, Virat [9] contains 315 videos collected from the same 11 cameras at the same time of the day. Videos from the same camera are highly redundant in both visual appearance as well as in motion and occlusion statistics. For these reasons, we select at most 3 videos per camera.
- **Static vs moving cameras.** Our dataset focuses on the static-camera use-case, therefore we discard videos captured by moving cameras.
- **Camera position.** We select videos from a large variety of camera positions and angles (from bird's-eye view to low-angle view).
- **Environment conditions.** We select videos recorded at different times of the day (day/night) and different weather conditions (sunny/rain/snow/cloudy). In particular, we ensure a proper representation of videos captured at night or in foggy conditions which will challenge even the best state-of-the-art object detector.

1.3 Details on dataset composition

Tab. 1 shows a detailed break-down of the number of videos per data source. Note that only 20% of the videos were sourced from existing public datasets. In particular, we selected only 25 videos from Virat [9] and Meva [4] with at most 3 videos per unique camera.

2 Annotation Pipeline

The definition of the track-level tag is as follows.

- **Sitting/standing still person**. A person is sitting/standing without moving in the entire video, such as a person that is eating in a restaurant or is resting on the airport bench.
- **Person in vehicle**. A person is sitting in a moving or static vehicle and it is usually occluded. Vehicles include buses, cars, trams and trains.

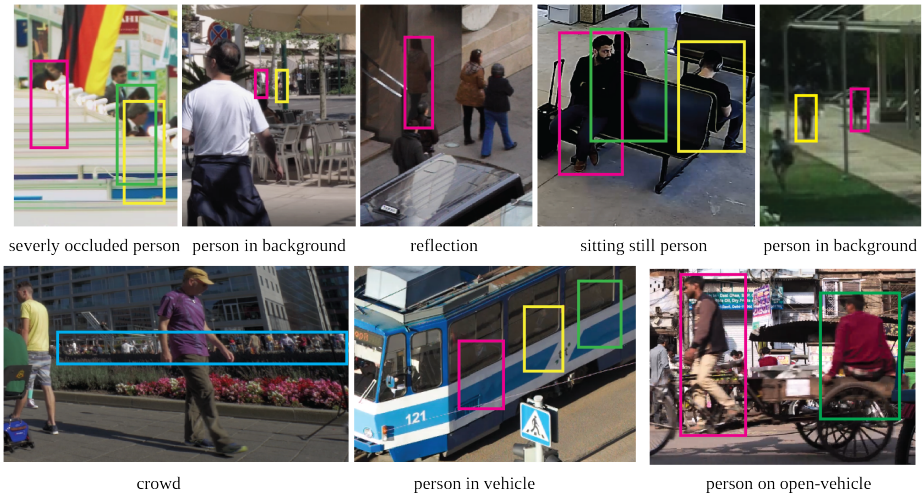


Fig. 1: Visualization of person bounding boxes with special tags. Note that we only show the person boxes that have a special tag, which means that the persons that are not enclosed with a bounding box in the given examples are tagged with "foreground person".

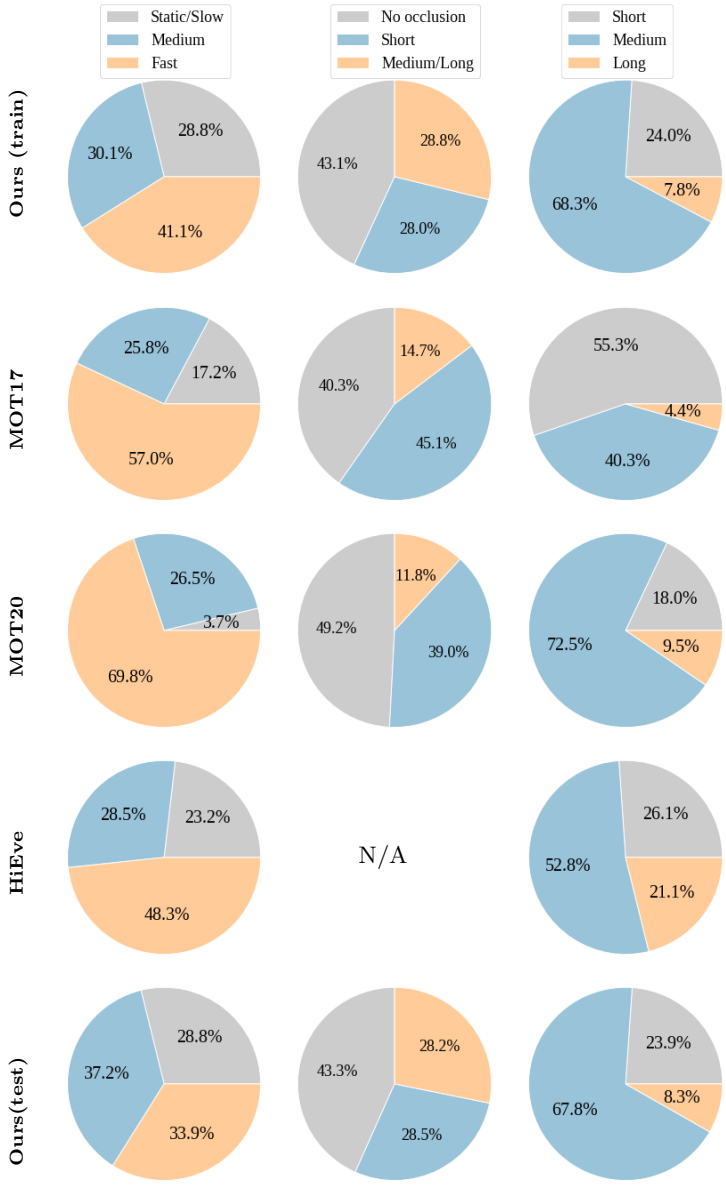
- **Person on open-vehicle**. A person/baby is sitting on an open-vehicle whicn includes bicycles, motorbikes, rickshaw charts or stroller. They are usually fully visible.
- **Reflection**. A person is reflected in a surface such as a mirror or window.
- **Severely occluded person**. A person is severely occluded (less than 20% visible) during the full duration of the video.
- **Person in background**. A person is far away from camera and its identity is barely distinguishable without context. In this case, the person can appear within a `crowd` region, or it is severely occluded, or its size is too small.
- **Foreground person**. A moving person in the foreground which might get occluded from time to time and whose identity is recognizable most of the time. We primarily target tracking these person tracks.

In Fig. 1, we also show representative visual examples for the special tag.

3 Dataset

3.1 Statistics of train/test split

In Tab. 2, we show the key statistics of our full dataset as well as its train and test splits. In Fig. 2, we extend the Fig. 4 in the main paper to show the track-level statistics for our dataset (train / test split), MOT17 [8], MOT20 [5] and HiEve [6]. Note that we are unable to estimate the occlusion duration



(a) Track speed (b) Occlusion duration (c) Track duration

Fig. 2: Comparison of track-level statistics of person tracks among different datasets. Note that we are unable to estimate the occlusion duration on HiEve dataset [6] as meta data related to occlusion of each person box is not available.

Dataset	#Videos	Length (secs)	#Annotated Frames	#Person Tracks	Min Res.
Ours (All)	236	8,334	200,769	12,150	720x480
Ours (Train)	138	4,736	118,685	7,096	720x480
Ours (Test)	98	3,598	82,084	5,054	720x480

Table 2: Key statistics of our dataset. Annotated Frames refer to the frames that are manually annotated and those that are automatically interpolated and then manually verified.

on HiEve dataset as meta data related to occlusion is not provided. In both MOT17 [8] and MOT20 [5], we define a person as fully occluded when the provided `visibility` score (between 0 and 1) is below 0.05.

Overall, our dataset includes higher-proportion of medium-to-long occlusion cases, which we think are of significant interest in real-world person tracking scenarios.

4 Experiments

4.1 Ignore person tracks

During training, we ignore person tracks that are tagged with '`person in vehicle`', '`severely occluded person`' and '`person in background`', as the corresponding person is barely recognizable. During evaluation, on top of those person tracks, we further ignore person tracks that are tagged with '`reflection`' and '`person on open-vehicle`' as we find those person tracks are less interesting in real-world applications. In addition, we also ignore person bounding boxes that are included in a '`crowd`' region during evaluation.

4.2 Implementation details

CenterTrack [13]. We use the official model made available by the authors, which was trained on the CrowdHuman dataset [10] using an input resolution of 512×512. We then fine-tune the model on our dataset using a learning rate of $1.25e^{-4}$ for 10 epochs using mini-batch size 32, increasing the input resolution to 960×544. We set $\lambda_{fp} = 0.1$, $\lambda_{fn} = 0.4$ at training time, and $\theta = 0.5$ and $\tau = 0.5$ at evaluation time, and enabled both the input tracking heatmap as well as the amodel box training and inference, which was found to improve the tracking metrics by the original authors.

SiamMOT [11] . We start with the official CrowdHuman-pretrained model made available by the authors, and we fine-tune it on our dataset by using a learning rate of 0.01 for $10K$ iterations with SGD with momentum. The learning rate is decreased by a factor of 10 after $6K$ and $8K$ iterations, respectively. All training

details remain the default as suggested by the paper [11]. During inference, we set the linking confidence $\alpha = 0.4$, the detection confidence $\beta = 0.6$ and keep a trajectory active until it is unseen for $\tau = 30$ frames. We resize the image such that its shorter side has 800 pixels while its longer side does not exceed $1,500$ pixels.

FairMOT [12]. We finetune the official CrowdHuman-pretrained model on our dataset using a learning rate of 0.001 for 20 epochs with a batch size of 32. The learning rate is decreased by a factor of 10 after 15 epochs. All other training and inference details remain the same as in the official FairMOT github repository.

References

1. Fillerstock, https://fillerstock.com/ 1, 2
2. Pexels, https://www.pexels.com/ 1, 2
3. Pixabay, https://pixabay.com/ 1, 2
4. Corona, K., Osterdahl, K., Collins, R., Hoogs, A.: Meva: A large-scale multiview, multimodal video dataset for activity detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1060–1068 (January 2021) 2
5. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L.: Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003 (2020) 3, 5
6. Lin, W., Liu, H., Liu, S., Li, Y., Qian, R., Wang, T., Xu, N., Xiong, H., Qi, G.J., Sebe, N.: Human in events: A large-scale benchmark for human-centric video analysis in complex events. arXiv preprint arXiv:2005.04490 (2020) 3, 4
7. Manen, S., Gygli, M., Dai, D., Gool, L.V.: Pathtrack: Fast trajectory annotation with path supervision. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 290–299 (2017) 2
8. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016) 3, 5
9. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.C., Lee, J.T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., et al.: A large-scale benchmark dataset for event recognition in surveillance video. In: CVPR 2011. pp. 3153–3160. IEEE (2011) 2
10. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018) 5
11. Shuai, B., Berneshawi, A., Li, X., Modolo, D., Tighe, J.: Siammot: Siamese multi-object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12372–12382 (2021) 5, 6
12. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision **129**(11), 3069–3087 (2021) 6
13. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: European Conference on Computer Vision. pp. 474–490. Springer (2020) 5