Volume xx (200y), Number z, pp. 1–3

NOVA: Rendering Virtual Worlds with Humans for Computer Vision Tasks (Supplementary Material)

Abdulrahman Kerim[1,3†], Cem Aslan[1,†], Ufuk Celikcan[1], Erkut Erdem[1] and Aykut Erdem[2]

[1]Hacettepe University, Department of Computer Engineering, Ankara, Turkey
[2]Koç University, Department of Computer Engineering, Istanbul, Turkey
[3] Lancaster University, School of Computing and Communications, UK

This document presents an extension to the data and the analysis of the tracking experiments provided in the main manuscript.

1. Additional Results on the Evaluation of Trackers Using Synthetic Sequences

In our main evaluation, we provide the overall precision plots for the six correlation-filter based trackers and evaluate their performances under challenging cases. Table 1 reports expected average overlap (EAO) scores of the evaluated trackers on our VirtualPTB1 dataset. In the following, we provide an expansion with an analysis of all attributes tested.

Scene Crowdedness. We study the effect of the level of scene crowdedness on the trackers' performances. As can be seen from Fig. 1, trackers' performances degrade when the scene contains many persons. However, the trackers perform fairly good at the scenes containing fewer number of persons. This is indeed expected since having many persons moving in view increases the likelihood of occlusions and causes some of the trackers to lose track of the person of interest. Camera altitude has a similar kind of effect on the trackers.

Camera Altitude. As demonstrated in Fig. 2, cameras located at middle altitudes allow the tracker to obtain a better overview of the persons, making it easier to track these persons, whereas the sequences captured by cameras at low altitudes are more likely to be subject to occlusions of the moving persons, resulting in a higher degree of difficulty for the trackers. At the same time, high camera altitude causes the object of interest to look small in the camera view. Thus, more difficulty in distinguish and tracking this object as it moves in the scene.

Time of Day and Weather Condition. We additionally analyze the influence of different times of day and weather conditions. As shown in Fig. 3, most of the trackers have difficulty in tracking persons in the sequences captured at night. On the other hand, they mostly succeed at tracking at sunset, sunrise and midday. Similarly, as demonstrated in Fig. 4, weather conditions have a direct effect on the trackers' performances. Fog and lightning storms might cause

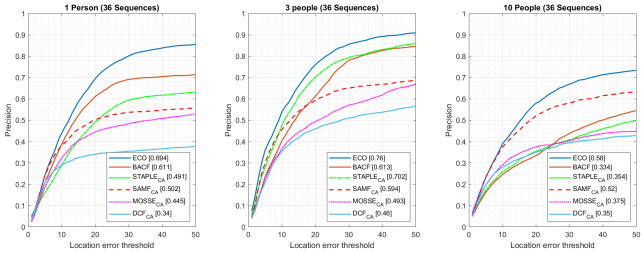

Figure 1: Precision plots for different crowdedness levels. In general, trackers perform fairly good at the scenes containing fewer number of people. However, their performance degrades as the scene becomes more crowded.

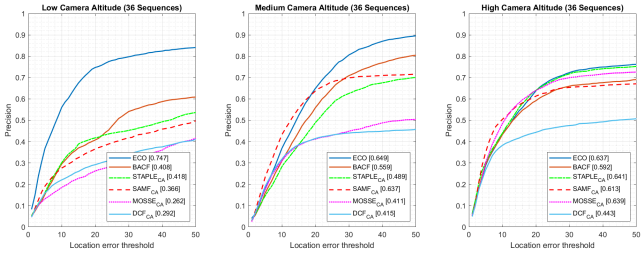

Figure 2: Precision plots for different camera altitudes. Generally, higher altitude cameras causes the trackers to perform slightly worse as compared to low and middle camera altitude.

sudden changes in the appearance of persons being tracked, and hence pose challenges for the trackers.

Level of Occlusion and Scale Change. Apart from the scene characteristics, we also examine the effect of the level of occlusions and large scale variations. Basically, if the ratio of the initial bounding box to at least one subsequent bounding box is less than 0.5 or greater than 2, then the sequence is denoted as having a large scale variation. It is clear from Fig. 5 that while most of the trackers

Table 1: EAO for the six different trackers evaluated on VirtualPTB1 dataset.

Tracker	Crowdedness			Camera Altitude			Time of Day			Weather Condition				Occlusion		Scale Variation	
	1 Person	3 People	10 People	Low	Medium	High	Sunset/Sunrise	Midday	Night	Normal	Snow	Fog	Lightstorm	Low	High	No	Yes
ECO	55.85	65.86	50.33	63.49	58.55	50.00	60.00	58.29	53.76	56.31	55.12	55.23	62.72	62.64	42.23	64.61	48.93
BACF	48.94	55.85	39.95	43.45	54.03	47.25	52.87	47.00	44.87	45.62	49.00	48.16	50.20	54.86	29.34	54.22	41.32
$STAPLE_{CA}$	40.52	58.06	37.51	38.39	46.63	51.06	47.74	44.78	43.57	40.85	44.00	47.44	49.15	51.80	26.96	50.41	39.50
$SAMF_{CA}$	35.20	46.75	38.14	31.52	44.15	44.41	42.04	44.27	33.77	36.27	40.85	39.13	43.88	44.77	26.47	43.78	35.68
$MOSSE_{CA}$	22.01	24.24	20.11	8.98	12.28	45.11	25.01	20.69	20.66	23.47	18.65	21.17	25.21	23.67	17.7	21.59	22.75
DCF_{CA}	19.45	22.79	18.13	10.49	15.28	34.61	20.91	16.83	22.63	16.42	17.80	21.56	24.72	22.04	14.66	22.61	17.24

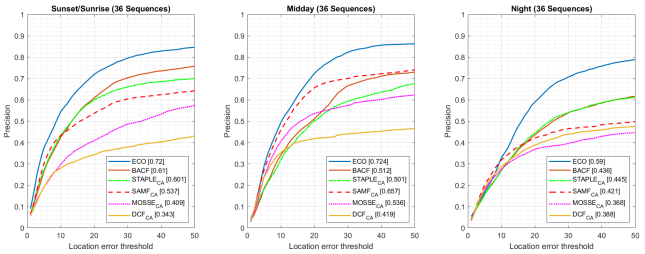

Figure 3: Precision plots for different times of the day. Generally, trackers perform poorly at night scenes as compared to theirs at sunrise, sunset and midday scenes.

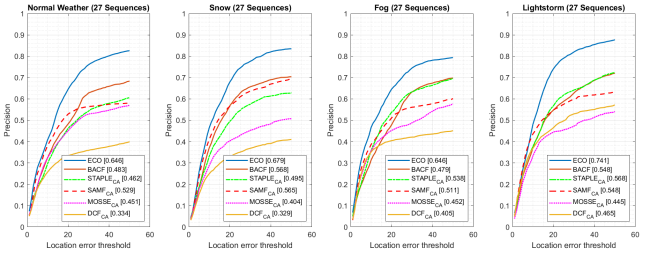

Figure 4: Precision plots for different weather conditions. Apparently, each weather condition effects each tracker differently. ECO's performance degrades noticeably for foggy weather conditions. On the other hand, MOSSE(CA) performs poorly in snowy weather conditions.

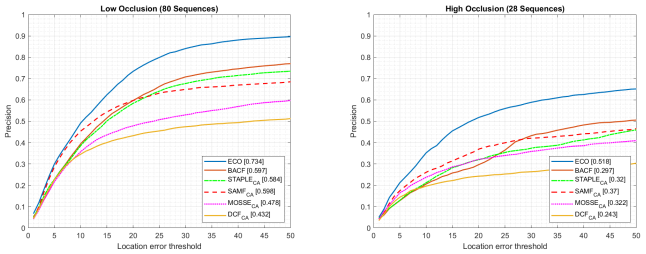

Figure 5: Precision plots for different occlusion levels. Clearly, all trackers perform poorly when the object of interest is highly and frequently occluded.

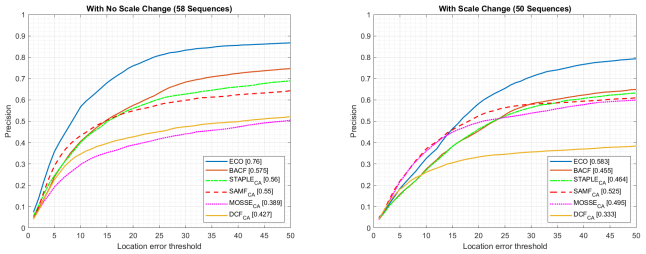

Figure 6: Precision plots for scale variation. The performance of the trackers degrades noticeably as the size the object of interest changes dramatically.

can deal with low levels of occlusion, the overall performance of the trackers deteriorates for the sequences where the persons being tracked are occluded either by other persons or other scene elements, such as trees or moving cars. Moreover, as shown in Fig. 6, due to the design of these trackers, most of them fail to keep track of the person of interest when large scale variations are observed throughout the sequences.

1.1. Discussion

As discussed before, some of these attribute classes such as weather conditions or different times of day have not been investigated in previous benchmark datasets. As can be seen from the results, some of the attributes from these newly evaluated classes are in fact the attributes that most the state-of-the-art trackers evaluated in this paper struggle to deal with. Our analysis reveals that the most challenging attributes seem to be high crowdedness, high camera altitude, night time, foggy weather, high occlusion and scale variation. While some attributes affect the overall performance of all trackers all together, some others have a direct effect only on a few trackers. For instance, it is clear from Fig. 2 that higher camera altitudes improve the performance of MOSSE(CA) and STAPLE(CA) trackers while the opposite is observed for ECO tracker. At the same time, from Fig. 6, we can conclude that MOSSE(CA) was the only tracker that exhibits robustness against variations in the scale of the people being tracked. Moreover, as shown in Fig. 4, snowy weather conditions cause the performance of MOSSE(CA) to decay noticeably while the fog was the main challenging attribute for ECO among all the other weather conditions. When the sequences where all six trackers performed the worst in terms of precision are examined, it is clear that the failures are not due to occlusions (in fact, the tracked person is fully visible in most of them), but due to

Figure 7: Sample frames from the most challenging synthetic sequences where the tested trackers mostly fail.

Table 2: Comparison of the synthetic, the real and the mixed sequences used in the training experiments.

Dataset	# of Sequences	Min Frames	Mean Frames	Max Frames	Total Frames
Synthetic	97	800	995	1187	96519
Real	97	46	497	4873	48171
Mix	194	46	746	4873	144690

the presence of extreme weather or low illumination conditions or low camera altitudes. Fig. 7 shows some of these challenging sequences. These results demonstrate that the correlation filter based trackers in our analysis and their online learning mechanisms cannot cope with the aforementioned challenges.

2. Additional Details About the Use of Synthetic Data for Training Deep Trackers

In Fig. 8, sample images from the training and testing sequences used in the second set of experiments are given. The ones in the blue and the red frames belong to the real and the synthetic training sequences, respectively. On the other hand, the green frame contains the test sequences that are used for evaluating the deep trackers. It is seen from the test sequences that different scene attributes such as scene crowdedness, camera altitude, camera type, and illumination condition are all present in these sequences. This is critical to ensure the validity of the test sequences for being a good proxy of the tracking scenarios in practice. Additionally, in Table 2, we show the statistics of the datasets used in the experiments. It is worth noting that the number of sequences is the same for both the synthetic and the real sequences used in the experiments E1 and E2, respectively. However, the average number of frames per sequence and the total number of frames in the synthetic dataset are almost twice those in the real dataset.

Figure 8: Sample images from the training and the testing sequences used in the second set of experiments. The blue and the red frames show samples from the real and synthetic sequences used in training, respectively. The green frame demonstrates samples from the whole 28 real sequences used in testing.