Supplementary Material for VmAP: A Fair Metric for Video Object Detection

Anupam Sobti Vaibhav Mavi M Balakrishnan Chetan Arora
Indian Institute of Technology Delhi
anupamsobti@cse.iitd.ac.in

ACM Reference Format:
Anupam Sobti Vaibhav Mavi M Balakrishnan Chetan Arora. 2021. Supplementary Material for VmAP: A Fair Metric for Video Object Detection. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3474085.3475383

1 ABLATION: UNIFYING CRITERIA

We experiment with the following unifying criteria:

- **Location:** The criterion described in the main paper and used in other experiments.
- **Appearance:** We use 3D histogram of RGB values of pixels inside the bounding box b and use the resulting 512-dimensional vector (${V^b}_{app}$) as the feature representing the appearance of b. For bounding boxes b_1 and b_2, the unifying criteria $\mathcal{U}_a$ is defined as:

$$\mathcal{U}_a(b_1, b_2) = ||{V^{b_1}}_{app} - {V^{b_2}}_{app}|| \quad (1)$$

  where $||.||$ represents the euclidean distance between vectors The bounding box b_2 is added to the set if $\mathcal{U}_a(b_1, b_2) < \theta_{app}$, where θ_{app} is the threshold.
- **Location + Appearance:** A bounding box b_2 is added to the set with b_1 as its first frame if $\mathcal{U}_a(b_1, b_2) < \theta_{app}$ and $\mathcal{U}_l(b_1, b_2) < \theta_{IoU}$
- **Time duration:** A bounding box is added to a set if the number of frames already present in the set is less than a certain threshold (θ_t).

The results are shown in Table 1. There is not much difference in the ranking of the algorithms while the absolute numbers change as per the criteria. Fig. 1 shows an example of sets formed in the same video with different unifying criteria. As one may observe, the appearance criteria may be more suited in counting like applications, while the location criteria is more suitable to perceive and control systems (for optimal path planning and control tasks, etc.).

Figure 2 shows the average set length for each class. The sets are smaller for the apperance and appearance-location criteria. As expected, the set size is fixed for the time criteria. We believe that all criteria are useful in some scenarios, e.g., location criterion in perceive-and-control systems, the time criterion in a latency-critical application and the apperance-location criterion for safety-critical applications where every small change needs to be tracked.

Detector	mAP	$\mathcal{U}_l$	$\mathcal{U}_a$	$\mathcal{U}_{al}$	$\mathcal{U}_t$
MEGA	83.4	56.6	71.6	66.1	76
RDN	81.3	53.65	67.3	60.6	72.5
FGFA	78.9	49.4	63.6	55.4	69.3
DFF	75.3	45.6	59.9	51.3	65.2

Table 1: The table shows that although the range of values are quite different, there isn't any difference in the ranking of algorithms by VmAP with different set formation strategies.

Detector	Video Recall		False Positives	
	P-R	VP-VR	P-R	VP-VR
horse				
MEGA-BASE	54.5	35.98	467.0	50.0
DFF	45.5	26.98	441.0	1.0
FGFA	46.56	32.28	656.0	93.0
RDN	54.5	33.86	452.0	38.0
MEGA	55.03	37.57	408.0	38.0
lion				
MEGA-BASE	73.33	60.0	5.0	0.0
DFF	66.67	66.67	1.0	0.0
FGFA	60.0	60.0	5.0	0.0
RDN	100.0	100.0	26.0	5.0
MEGA	100.0	100.0	22.0	0.0
car				
MEGA-BASE	82.85	45.65	5590.0	207.0
DFF	77.7	38.92	6565.0	208.0
FGFA	80.74	44.2	6028.0	260.0
RDN	84.43	52.64	5741.0	317.0
MEGA	81.53	51.19	5554.0	373.0
cattle				
MEGA-BASE	73.06	57.51	233.0	14.0
DFF	72.02	55.44	437.0	39.0
FGFA	62.69	53.37	345.0	23.0
RDN	78.76	63.73	515.0	31.0
MEGA	83.94	67.36	553.0	34.0

Table 2: Table showing false positive comparison of PR and VP-VR operating points

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '21, October 20–24, 2021, Virtual Event, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475383

Figure 1: The figure shows the first and last frame of two of the sets formed by using the location criteria (left) and appearance criteria (right). In the upper row, the appearance criteria declares a new set when the object has come too close. This might be dangerous for perceive and control systems like self driving cars. In the second row, the appearance criteria declares a new set even when the train hasn't moved much. In both the cases, the location criteria gives a better estimation.

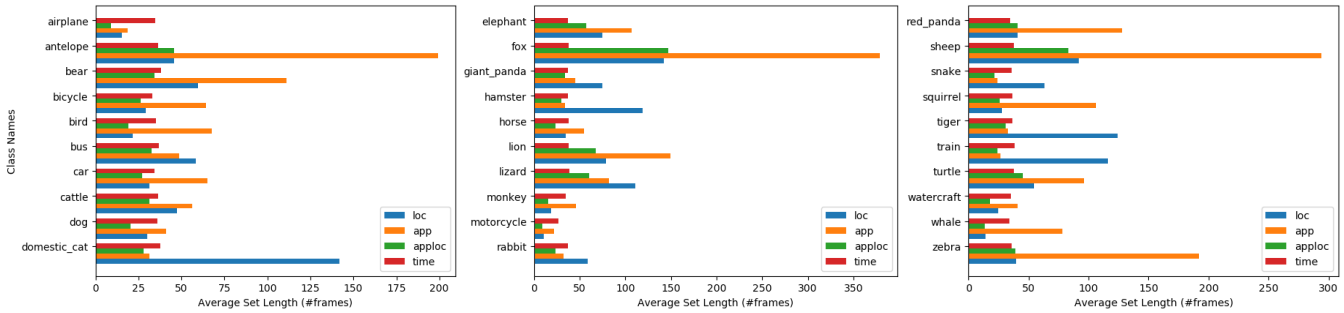


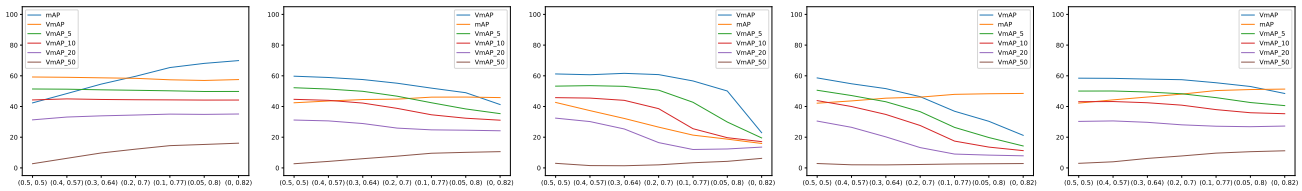
Figure 2: The sets are smaller for the appearance ($\mathcal{U}_a$) and even smaller for the ($\mathcal{U}_{al}$) criteria. As expected, the set size for the time criteria is the same ($\mathcal{U}_t$). We believe that all criteria could be useful in different scenarios as discussed in Sec. 1.

2 ABLATION: VMAP DEFINITION

In the main paper, we define a True Positive Set as a set in which at least one of the bounding boxes of the object in the set has an $IOU > IOU_{th}$. One may question the decision of taking at least one of the bounding boxes in the set. In this section, we verify whether the number of frames detected within a set is a critical parameter. We do this by performing the experiments with synthetic detectors as done in the main paper. The following definitions are compared:

(1) **VmAP** This is the baseline used in the paper. At least one bounding box in the set must have an $IOU > IOU_{th}$ as compared to the ground truth boxes.

(2) **VmAP_*N***: An alternative to using a single frame per set is to use a percentage of the set length as a threshold for counting a true positive. For example, in the VmAP_5 definition, a set is considered true positive if the detected bounding box matches the ground truth boxes in at least 5% of the frames within the set.



(a) Size (b) Illumination (c) Contrast (d) Speed (e) Color

Figure 3: The figure demonstrates how the different VmAP definitions respond to increase in different types of bias. We find that as the % value on number of frames in a set is increased, the behavior of the metric resembles that of mAP. In higher percentage values, a true positive set becomes very difficult to achieve resulting in extremely low scores. We, therefore, recommend the usage of VmAP as defined in the main paper.

Detector	Video Recall		False Positives	
	P-R	VP-VR	P-R	VP-VR
airplane				
MEGA-BASE	84.42	77.39	387.0	78.0
DFF	76.88	69.72	573.0	83.0
FGFA	80.53	68.59	615.0	57.0
RDN	86.06	78.89	445.0	74.0
MEGA	85.55	77.64	403.0	91.0
antelope				
MEGA-BASE	91.34	87.4	122.0	15.0
DFF	88.19	78.74	125.0	15.0
FGFA	88.98	81.89	107.0	30.0
RDN	92.91	90.55	96.0	40.0
MEGA	71.65	71.65	26.0	26.0
bear				
MEGA-BASE	91.6	57.14	633.0	31.0
DFF	87.39	67.23	561.0	22.0
FGFA	94.12	82.35	666.0	326.0
RDN	95.8	45.38	892.0	26.0
MEGA	94.12	56.3	775.0	36.0
bicycle				
MEGA-BASE	78.88	57.37	525.0	65.0
DFF	68.92	53.39	576.0	77.0
FGFA	70.52	53.78	399.0	42.0
RDN	77.29	58.96	542.0	56.0
MEGA	74.5	54.18	452.0	27.0
bird				
MEGA-BASE	41.46	28.85	383.0	28.0
DFF	35.29	35.01	132.0	96.0
FGFA	33.89	30.81	189.0	60.0
RDN	48.74	36.69	558.0	94.0
MEGA	41.46	36.41	210.0	100.0

Table 3: Table showing false positive comparison of PR and VP-VR operating points

3 OPERATING POINT COMPARISON

Tables 2, 3 and 4 show the reduction in false positives when using the confidence threshold corresponding to the optimal point on the P-R and VP-VR curves. As shown in the main paper, the number of false positives always decreases. In some cases, the decrease in set recall (VR) is also observed, which can be improved using traditional false-positive/false-negative trade-off using confidence threshold tuning.

4 POST-TRACKING ASSESSMENT

As discussed in the main paper, the Video Recall (VR) measure follows the FR + Tracker measure more closely as compared to the Frame Recall (FR) measure. Fig. 4 shows this trend in the decreasing order of ground truth measure (FR + Tracker). This indicates the suitability of the metric for measurement of the true ability of an object detector, when it would be used in conjunction with a tracker.

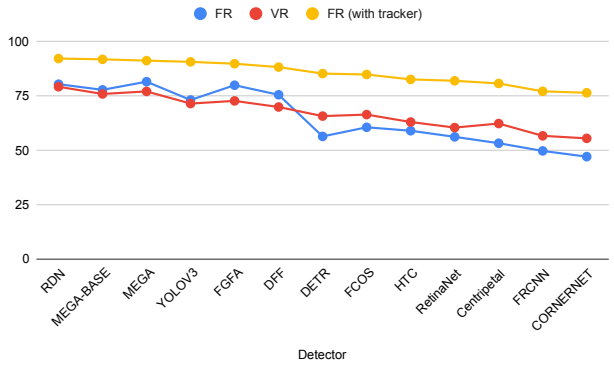


Figure 4: The detectors are presented in the order of ground truth rankings (Frame Recall using a Detector + Ideal Tracker). While the Frame Recall is low for some detectors (see FGFA, DETR), the post-tracking performance is quite high. On the other hand, VR follows the post-tracking performance closely.

Detector	Video Recall		False Positives	
	P-R	VP-VR	P-R	VP-VR
bus				
MEGA-BASE	67.06	56.47	398.0	44.0
DFF	64.71	43.53	321.0	7.0
FGFA	57.65	50.59	156.0	28.0
RDN	69.41	52.94	292.0	12.0
MEGA	72.94	55.29	508.0	26.0
dog				
MEGA-BASE	79.81	57.08	1220.0	122.0
DFF	71.46	52.9	1489.0	120.0
FGFA	79.35	57.31	929.0	71.0
RDN	83.06	62.41	1041.0	131.0
MEGA	85.15	65.66	1078.0	96.0
domestic cat				
MEGA-BASE	77.27	27.27	307.0	0.0
DFF	63.64	27.27	437.0	41.0
FGFA	81.82	22.73	242.0	5.0
RDN	72.73	36.36	188.0	0.0
MEGA	13.64	9.09	33.0	0.0
elephant				
MEGA-BASE	90.09	61.26	749.0	8.0
DFF	89.19	61.26	502.0	40.0
FGFA	85.59	56.76	556.0	21.0
RDN	92.79	73.87	772.0	30.0
MEGA	95.5	81.98	1019.0	27.0
fox				
MEGA-BASE	61.9	57.14	15.0	2.0
DFF	66.67	57.14	46.0	1.0
FGFA	66.67	61.9	30.0	3.0
RDN	71.43	66.67	28.0	3.0
MEGA	76.19	66.67	28.0	3.0
giant panda				
MEGA-BASE	70.97	59.68	164.0	9.0
DFF	70.97	53.23	373.0	2.0
FGFA	69.35	59.68	188.0	1.0
RDN	77.42	53.23	228.0	1.0
MEGA	30.65	30.65	3.0	3.0

Table 4: Table showing false positive comparison of PR and VP-VR operating points

5 VIDEO FOR TIMELINE PLOT

We also have a video for the timeline plot in the supplementary material. This is to further demonstrate how a timeline plot is able to convey the events in an entire video through a simple diagram.