Localization Recall Precision (LRP): A New Performance Metric for Object Detection

Kemal Oksuz, Baris Can Cam(✉), Emre Akbas, and Sinan Kalkan

Department of Computer Engineering,
Middle East Technical University, Ankara, Turkey
{kemal.oksuz,can.cam,eakbas,skalkan}@metu.edu.tr
http://image.ceng.metu.edu.tr

Abstract. Average precision (AP), the area under the recall-precision (RP) curve, is the standard performance measure for object detection. Despite its wide acceptance, it has a number of shortcomings, the most important of which are (i) the inability to distinguish very different RP curves, and (ii) the lack of directly measuring bounding box localization accuracy. In this paper, we propose "Localization Recall Precision (LRP) Error", a new metric specifically designed for object detection. LRP Error is composed of three components related to localization, false negative (FN) rate and false positive (FP) rate. Based on LRP, we introduce the "Optimal LRP" (oLRP), the minimum achievable LRP error representing the best achievable configuration of the detector in terms of recall-precision and the tightness of the boxes. In contrast to AP, which considers precisions over the entire recall domain, oLRP determines the "best" confidence score threshold for a class, which balances the trade-off between localization and recall-precision. In our experiments, we show that oLRP provides richer and more discriminative information than AP. We also demonstrate that the best confidence score thresholds vary significantly among classes and detectors. Moreover, we present LRP results of a simple online video object detector and show that the class-specific optimized thresholds increase the accuracy against the common approach of using a general threshold for all classes. Our experiments demonstrate that LRP is more competent than AP in capturing the performance of detectors. Our source code for PASCAL VOC AND MSCOCO datasets are provided at https://github.com/cancam/LRP.

Keywords: Average precision · Object detection
Performance metric · Optimal threshold · Recall-precision

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-01234-2_31) contains supplementary material, which is available to authorized users.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11211, pp. 521–537, 2018.
https://doi.org/10.1007/978-3-030-01234-2_31

1 Introduction

Today "average precision" (AP) is the de facto standard for performance evaluation in object detection competitions [8,14,28], and in the studies on still-image object detection [6,13,16,24], video object detection [9,12,36] and online video object detection [17,34]. AP not only enjoys such vast acceptance but it also appears to be unchallenged. Except for a small number of papers which do ablation studies [13,24], AP appears to be the sole criterion used to compare object detection methods.

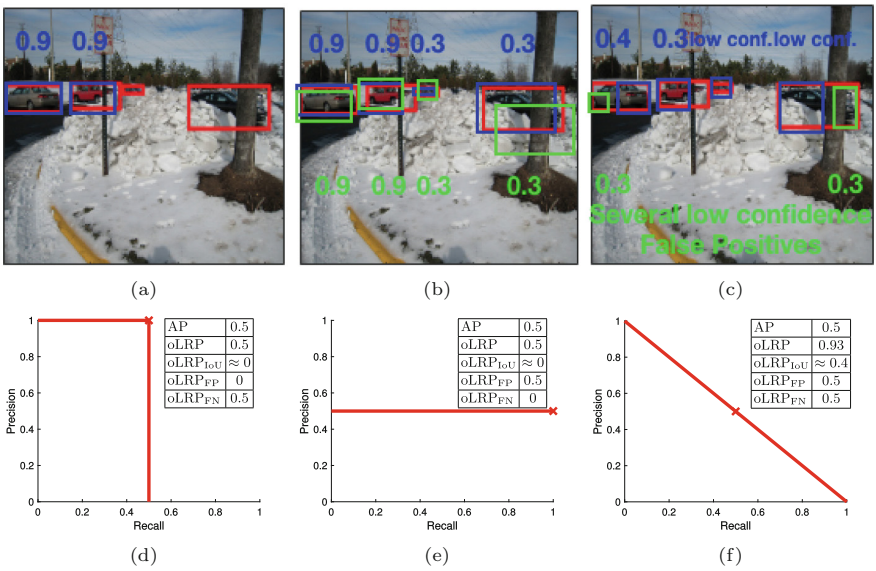


Fig. 1. Three different object detection results (for an image from ILSVRC [28]) with very different RP curves but the same AP. AP is unable to identify the difference between these curves. **(a, b, c)** Red, blue and green colors denote ground-truth, true positives; false positives respectively. Numbers are detection confidence scores. **(d, e, f)** RP curves, AP and LRP results for the corresponding detections in (a, b, c). Red crosses indicate Optimal LRP points. (Color figure online)

Despite its popularity, AP has certain deficiencies. First, **AP cannot distinguish between very different RP curves**: In Fig. 1, we present detection results of three hypothetical object detectors. The detector in (a) detects only half of the objects but with full precision; this is a low-recall-high-precision detector. In contrast, the detector in (b) detects all objects; however, for each correct detection it also produces a close-to-duplicate detection which escapes non-maxima suppression. Hence, detector (b) is a high-recall-low-precision detector. And the detector in (c) is in between; it represents a detector with higher precision at lower recall and vice versa. Despite their very different characteristics, the APs of these detectors are exactly the same (AP = 0.5). One needs to

inspect the RP curves in order to understand the differences in behavior, which can be time-consuming and impractical with large number of classes such as in the ImageNet object detection challenge [28] with 200 classes.

Another deficiency of AP is that **it does not explicitly include localization accuracy**: One cannot infer from AP the tightness level of the bounding box detections. Nevertheless, since extracting tighter bounding boxes is a desired property, nearly every paper on the topic discusses the issue mostly qualitatively [6,9,16,17,24] and some quantitatively by computing AP scores for different intersection-over-union (IoU) thresholds [13,16,24]. However, this quantitative approach does not directly measure the localization accuracy either and for the qualitative approach, it is very likely for the sample boxes to be very limited and biased. We discuss other less severe deficiencies of AP in Sect. 3.

A desirable performance metric is expected to include all of the factors related with performance. In object detection, the most important three factors are (i) the localization accuracy of the true positives (TP), (ii) the false positives (FP) rate and (iii) the false negative (FN) rate. Being able to evaluate a detector based on these factors is another desirable property for a performance measure since it can reveal improvement directions. Furthermore, a performance metric should reveal the RP characteristics of a detector (as LRP achieves in Fig. 1). This ability would benefit certain applications. For instance, using a high-precision detector is common in visual tracking methods [3,4,31,32,37], while initializing the tracker, known as *tracking by detection* as faster response times are required. Also, in online video object detection, the current approach is to use a still-image object detector with a general threshold (e.g., Association-LSTM [17] uses SSD [16] detections with confidence score above 0.8). A desirable performance measure should help in setting an optimal confidence score threshold per class.

In this paper, we propose a new metric called the "Localization-Recall-Precision Error" (LRP, for short). LRP involves appropriate components closely related to the precision, recall, and IoU and each parametrization of LRP corresponds to a point on the RP curve. We propose the "Optimal LRP", the minimum achievable LRP error, as the alternative performance metric to AP. Optimal LRP alleviates the drawbacks of AP, represents the tightness of the bounding-boxes and the shape of the RP curve via its components and is more suitable for ablation studies. Finally, based on Optimal LRP, a confidence score thresholding method is proposed to decrease the number of detections in an optimal manner. Our extensive experiments confirm that LRP is a highly capable metric for comparing object detectors thoroughly.

2 Related Work

Information Theoretic Performance Measures: Several performance measures have been derived on the confusion matrix. Among them, the most relevant one is the F-measure [25] defined as the harmonic mean of precision and recall. However, F-measure violates the triangle inequality, and therefore, it is not suitable as a metric [20] and it is not symmetric in the positive and negative classes.

These violations and its incapacity to measure bounding box tightness prevent its use for comparison among detectors in a consistent manner. Moreover, [5] points out that, except for accuracy, all information theoretic measures have undefined intervals. For example, F-measure is undefined when the number of TP is 0 even if there are detections. AP is an information theoretic measure, too, with deficiencies discussed in Sects. 1 and 3.

Point Multi-target Tracking Performance Metrics: Object detection is very similar to the multi-target tracking problem. In both problems, there are multiple instances to detect, and the localization, FN and FP rates are common criteria for success. Currently, component-based performance metrics are the accepted way of evaluating point multi-target tracking filters. The first metric to combine the localization and cardinality (including both FP and FN) errors is the Optimal Subpattern Assignment (OSPA) [29]. Following OSPA, several measures and metrics have been proposed as its variants [19,23,26,27,29,30,35]. Similarly, CLEAR multi-object tracking metrics [1] considers only FP and mismatch rate while ignoring the localization error. However, similar measures and metrics are lacking in the object detection literature, though similar performance evaluation problems are observed.

Setting the Thresholds of the Classifiers: The research on the optimization of a precision-recall balanced performance measure is mostly concentrated around the F-measure. [7] considers maximizing F-measure at the inference time using plug-in rules, while [18,33] offer maximization during training for support vector machines and conditional random fields. Similarly, [15] aims to find optimal thresholds for a probabilistic classifier based on maximizing the F-measure. Finally, [21] presents a theoretical analysis of optimization of the F-measure, which also confirms the threshold-F-measure relationship depicted in [15,22].

In summary, we see that existing methods mostly focus on the F-measure for optimizing the thresholds for classifiers, which, however, has the aforementioned drawbacks. Moreover, F-measure is shown to be concave with respect to its inputs, number of TPs and FPs [15], which makes the analytical optimization impossible. In addition, none of these studies have considered the object detection problem in particular, thus no localization error is directly included for these measures. Therefore, different from the previous work, we specifically are interested in performance evaluation and optimal thresholding of the deep object detectors. Moreover, we directly optimize a well-behaving function which has a smaller domain in practice in order to identify the class-specific thresholds.

3 Average Precision: An Analysis and Its Deficiencies

Due to space constraints, we omit the definition of AP and refer the reader to the accompanying supplementary material or [8]. There exist minor differences in AP's practical usage. For example, AP is computed by simply integrating over 11 points (that divide the entire recall domain in equal pieces) in the PASCAL VOC 2007 challenge [8] whereas in MSCOCO [14], 101 points are used. Precision values at intermediate points are simply interpolated to prevent wiggles in the curve,

by setting it to the maximum precision obtained in the interval of higher recall than the current point. While a single intersection-over-union (IoU) threshold, which is 0.5, is used in PASCAL VOC [8]; a range of IoU thresholds (from 0.5 to 0.95) are used in MSCOCO; the average AP over this range of IoU thresholds is also called mAP.

AP aims to evaluate the precision of the detector over the entire recall domain. Thus, it favors the methods that have precision over the entire recall domain, instead of the detectors whose RP curves are nearer to the top-right corner. In other words, AP does not compare the maximum but the overall capability/performance of the detectors. The most important two deficiencies of AP are discussed in Sect. 1. In the following, we list other, more minor deficiencies.

AP is not Confidence-Score Sensitive. Since the sorted list of the detections is required to calculate AP, a detector generating results in a limited interval will lead to the same AP. As an example, consider only 2 detections with same confidence score in Fig. 1 out of 4 ground truths. Note that setting the confidence scores to any value (i.e. 0.01) leads to the same AP as long as the order is preserved.

AP does not suggest a confidence score threshold for the best setting of the object detector. However, in a practical application, detections are usually required to be filtered owing to time limitations. For example, the state-of-the-art online object detector [17] applies a confidence score threshold of 0.8 on the SSD method [16] and obtains $12fps$ in this fashion.

AP uses interpolation between neighboring recall values, which is especially problematic for classes with very small size. For example, "toaster" class of [14] has 9 instances in the validation 2017 set.

4 Localization-Recall-Precision (LRP) Error

Let X be the set of ground truth boxes and Y be the set of boxes returned by an object detector. To compute $\mathrm{LRP}(X, Y_s)$, the LRP error of Y_s against X at a given score threshold s ($0 \leq s \leq 1$) and IoU threshold τ ($0 \leq \tau < 1$); first, Y_s, the set of detections with confidence score larger than s, is constructed and detections in Y_s are assigned to ground-truth boxes in X, as done for AP. Once the assignments are made, the following values are computed: (i) N_{TP}, the number of true positives; (ii) N_{FP}, the number of false positives; (iii) N_{FN}, the number of false negatives. Using these quantities, the LRP error is:

$$\mathrm{LRP}(X, Y_s) := \frac{1}{Z}\left(w_{IoU}\mathrm{LRP}_{IoU}(X, Y_s) + w_{FP}\mathrm{LRP}_{FP}(X, Y_s) + w_{FN}\mathrm{LRP}_{FN}(X, Y_s)\right), \tag{1}$$

where $Z = N_{TP} + N_{FP} + N_{FN}$ is the normalization constant; and the weights $w_{IoU} = \frac{N_{TP}}{1-\tau}$, $w_{FP} = |Y_s|$, and $w_{FP} = |X|$ control the contributions of the terms. The weights make each component easy to interpret, provide further information about the detector and prevent the total error from being undefined whenever

the denominator of a single component is 0. LRP_{IoU} represents the IoU tightness of valid detections as follows:

$$\text{LRP}_{IoU}(X, Y_s) := \frac{1}{N_{TP}} \sum_{i=1}^{N_{TP}} (1 - IoU(x_i, y_{x_i})), \tag{2}$$

which measures the mean bounding box localization error resulting from correct detections. Another interpretation is that $1 - \text{LRP}_{IoU}(X, Y_s)$ is the average IoU of the valid detections.

The second component, LRP_{FP}, in Eq. 1 measures the false-positives:

$$\text{LRP}_{FP}(X, Y_s) := 1 - Precision = 1 - \frac{N_{TP}}{|Y_s|} = \frac{N_{FP}}{|Y_s|}, \tag{3}$$

and false negatives are measured by LRP_{FN}:

$$\text{LRP}_{FN}(X, Y_s) := 1 - Recall = 1 - \frac{N_{TP}}{|X|} = \frac{N_{FN}}{|X|}. \tag{4}$$

FP and FN components together represent precision-recall of the corresponding Y_s by $1 - \text{LRP}_{FP}(X, Y_s)$ and $1 - \text{LRP}_{FN}(X, Y_s)$ respectively. Denoting the IoU between $x_i \in X$ and its assigned detection $y_{x_i} \in Y_s$ by $IoU(x_i, y_{x_i})$, the LRP error can be equally defined in a more compact form as:

$$\text{LRP}(X, Y_s) := \frac{1}{N_{TP} + N_{FP} + N_{FN}} \left(\sum_{i=1}^{N_{TP}} \frac{1 - IoU(x_i, y_{x_i})}{1 - \tau} + N_{FP} + N_{FN} \right). \tag{5}$$

LRP penalizes each TP by its erroneous localization normalized by $1 - \tau$ to the [0,1] interval, each FP and FN by 1 that is the penalty upper bound. This sum of error is averaged by the total number of its contributors, i.e., $N_{TP} + N_{FP} + N_{FN}$. So, with this normalization, LRP yields a value representing the average error per bounding box in the [0,1] interval, where each component equally contributes to the error. When necessary, the individual importance of IoU, FP, FN can be changed for different applications. To this end, the prominent component can be multiplied by a factor (say C) both in the numerator and the denominator [19]. This implies having C artificial errors for each error of the prominent type.

Overall, the ranges of total error and the components are $[0, 1]$ and lower value implies better performance. At the extreme cases; 0 for LRP means that each ground truth item is detected with perfect localization, and if LRP is 1, then no valid detection matches the ground truth (i.e., $|Y_s| = N_{FP}$). LRP is undefined only when the ground truth and detection sets are both empty (i.e., $N_{TP} + N_{FP} + N_{FN} = 0$), i.e., there is nothing to evaluate.

As for the parameters, s is the confidence score threshold, and τ is the IoU threshold. Since the RP pair is directly identified by the FP&FN components, each different detection set Y_s corresponds to a specific point of the RP curve. For this reason, decreasing s corresponds to moving along the RP curve in the

positive recall direction. τ defines minimum overlap for a detection to be validated as a TP. In other words, higher τ means we require tighter BBs. Overall, both parameters are related with the RP curve: A τ value sets the RP curve and an s value moves along the RP curve to evaluate the LRP error.

In the supplementary material, we prove that LRP is a metric.

5 Optimal LRP (oLRP) Error: The Performance Metric and Thresholder

Optimal LRP (oLRP) is defined as the minimum achievable LRP error with $\tau = 0.5$, which makes oLRP parameter independent:

$$\text{oLRP} := \min_{s} \text{LRP}(X, Y_s). \tag{6}$$

For ablation studies and practical requirements, different τ values can be adopted. In such cases, oLRP@τ can be used to denote the Optimal LRP error at τ.

oLRP searches among the confidence scores to find the best balance for competing precision-recall-IoU. The RP setting of the RP curve that oLRP has found corresponds to the top-right part of the curve, where the optimal balanced setting resides. We call a curve *sharper* than another RP curve, if its peak point at the top-right part is nearer to the $(1, 1)$ RP pair. To illustrate, the RP curves in Fig. 1(d) and (e) are sharper than that in Fig. 1(f).

The components of oLRP are coined as optimal box localization (oLRP_{IoU}), optimal FP (oLRP_{FP}), and optimal FN (oLRP_{FN}) components. In this case, oLRP_{IoU} describes the mean average tightness for a class, and oLRP_{FP} and oLRP_{FN} together pertain to the sharpness of the curve since the corresponding RP pair is the maximum achievable performance value of the detector for this class. One can directly pinpoint the sharpness point by $1 - \text{oLRP}_{FP}$ and $1 - \text{oLRP}_{FN}$. Overall, different from AP, oLRP aims to find out the best class specific setting of the detector and it favors sharper ones that also represent better BB tightness.

Denoting oLRP error of class $c \in C$ by oLRP_c, Mean Optimal LRP (moLRP) is defined as follows:

$$\text{moLRP} := \frac{1}{|C|} \sum_{c \in C} \text{oLRP}_c. \tag{7}$$

As in mAP, moLRP is the performance metric for the entire detector. Mean optimal box localization, FP and FN components, denoted by moLRP_{IoU}, moLRP_{FP}, moLRP_{FN} respectively, are similarly defined as the mean of the class specific components. Different from the components in oLRP, the mean optimal FP and FN components are not necessarily on the average of the RP curves of all classes due to averaging moLRP_{FP} (i.e., precision) with different moLRP_{FN} (i.e., recall) values but still provides information on the sharpness of the RP curves as shown in the experiments.

Owing to its filtering capability, oLRP can be used for thresholding purposes. If a problem needs an image object detector as the backbone and processing is to be completed within limited time, then only a small subset of the detections should be selected. For such methods, using an overall confidence score for the object detector is a common approach [17]. For such a task, oLRP identifies the class-specific best confidence score thresholds. One possible drawback of this method is that validated detections can still be too large to be processed in the desired limited time. However, by accepting larger LRP errors, higher confidence scores can be set, but again in a class-specific manner. Second practical usage of oLRP is about the deployment of the devised object detector into a platform in which confidence scores are to be discarded for user-friendliness. In such a case, one needs to set the τ threshold considering the application requirements while optimizing for the best confidence score.

In essence, calculating oLRP is an optimization problem. However, thanks to the smaller search space, we propose to discretize the s domain into 0.01 spaced intervals and search exhaustively in this limited space.

6 Experimental Evaluation

In this section, we analyze the parameters of LRP, represent its discriminative power on common object detectors and finally show that the class specific thresholds increase the performance of a simple online video object detector.

Evaluated Object Detectors: We evaluate commonly used deep object detectors; namely, Faster R-CNN, RetinaNet, and SSD. For Faster R-CNN and RetinaNet variants, we use the models by [11] and for SSD variants, the models of [10] are utilized. For the variants, we use R50, R101 and X101 while referring to the ResNet-50, ResNet-101 and RexNeXt-101 backbones respectively and FPN for feature pyramid network. All models are tested on "MS COCO validation 2017" including 80 classes and 5000 images.

6.1 Analyzing Parameters s and τ

Using Faster R-CNN (X101+FPN) results of the first 10 classes and mean-error for clarity, the effects of the s and τ are analyzed in Fig. 2 and 3. We observe

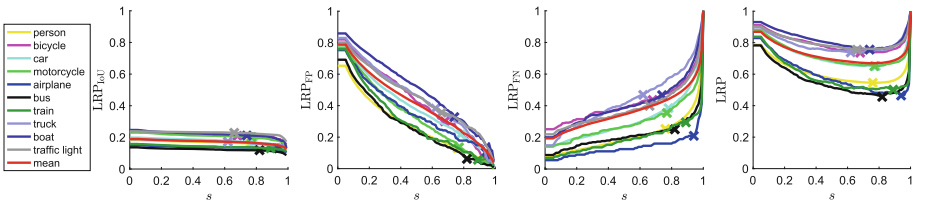


Fig. 2. For each class, LRP components & total error of Faster R-CNN (X101+FPN) are plotted against s. The optimal confidence scores are marked with crosses.

that box localization component is not significantly affected by increasing s, except for large s, where the error slightly decreases since the results tend to be more "confident". FP and FN components act in contrast to precision and recall respectively, as expected. Therefore, lower curves imply better performance for these components. Finally, the total error (oLRP) has a second-order shape. Since the localization error is not affected significantly by s, the behavior of the total error is mainly determined by FP and FN components, which result in the global minima of the total error to have a good precision and recall balance.

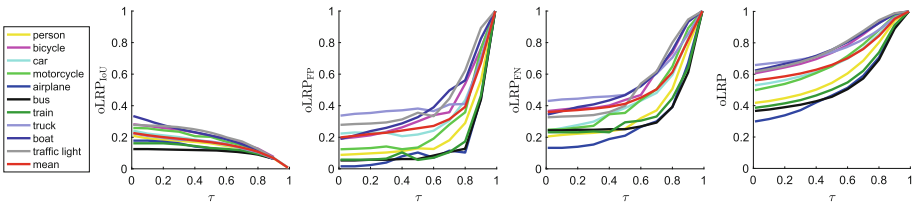


Fig. 3. For each class, oLRP and its components for Faster R-CNN (X101+FPN) are plotted against τ. The mean represents the mean of 80 classes.

In Fig. 3, oLRP and moLRP are plotted against different τ values. As expected, larger τ values imply lower the box localization component (oLRP_{IoU}). On the other hand, increase τ causes FP and FN components to increase rapidly, leading to higher total error (oLRP). This is intuitive since at the extreme case, i.e., when $\tau = 1$, there are hardly any valid detections and all the detections are false positives, which makes oLRP to be approximately 1. Therefore, oLRP allows measuring the performance of a detector designed for an application that requires a different τ by also providing additional information. In addition, investigating oLRP for different τ values represents a good extension for ablation studies.

Table 1. Performance comparison of common object detectors. R50, R101 and X101 represent the backbone networks used by ResNet-50, ResNet-101 and RexNeXt-101, respectively, and FPN refers to the feature pyramid network. s^*_{min} and s^*_{max} denote minimum and maximum class-specific thresholds respectively for oLRP. Note that unlike AP, lower scores are better for LRP.

	mAP	mAP@0.5	moLRP	$\text{moLRP}_{\text{IoU}}$	moLRP_{FP}	moLRP_{FN}	s^*_{min}	s^*_{max}
SSD-300	0.161	0.383	0.854	0.281	0.403	0.622	0.05	0.53
SSD-512	0.284	0.481	0.763	0.202	0.331	0.549	0.08	0.63
Faster R-CNN (R50)	0.348	0.557	0.714	0.183	0.292	0.484	0.18	0.93
RetinaNet (R50+FPN)	0.357	0.547	0.711	0.169	0.293	0.503	0.26	0.60
Faster R-CNN (R50+FPN)	0.379	0.593	0.689	0.175	0.259	0.454	0.41	0.94
RetinaNet (X101+FPN)	0.398	0.595	0.677	0.161	0.255	0.462	0.28	0.70
Faster R-CNN (R101+FPN)	0.398	0.613	0.673	0.168	0.255	0.436	0.37	0.94
Faster R-CNN (X101+FPN)	0.413	0.637	0.663	0.171	0.256	0.413	0.39	0.94

6.2 Evaluating Common Image Object Detectors

General Overview: Table 1 compares the detectors using mAP as the COCO's standard metric, mAP@0.50, moLRP and the class-specific threshold ranges. We observe that moLRP values are indicative of the known performances of the detectors. For any type of the detector, each new property (i.e., including FPN, increasing depth, using ResNext for Faster R-CNN and RetinaNet, increasing input size to 512 for SSD) decreases moLRP as expected. Moreover, the overall order is consistent with mAP except for RetinaNet (X101+FPN) and Faster R-CNN (R101+FPN), which are equal in terms of mAP; however, Faster R-CNN (R101+FPN) surpasses RetinaNet (X101+FPN) in terms of moLRP, which is discussed below. Note that moLRP_{FP} and moLRP_{FN} values in Table 1 are also consistent with the sharpness of the RP curves of the methods as presented in Fig. 4. To illustrate, Faster R-CNN (X101+FPN) has the best moLRP_{FP}, moLRP_{FN} combination, corresponding to the sharpest RP curve. Another interesting example pertains to the RetinaNet (X101+FPN) and Faster R-CNN (R50+FPN) curves. For these methods, moLRP_{FP} and moLRP_{FN} comparison slightly favors Faster R-CNN (R50+FPN), which is justified by their PR curves in Fig. 4.

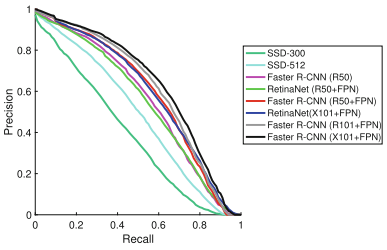
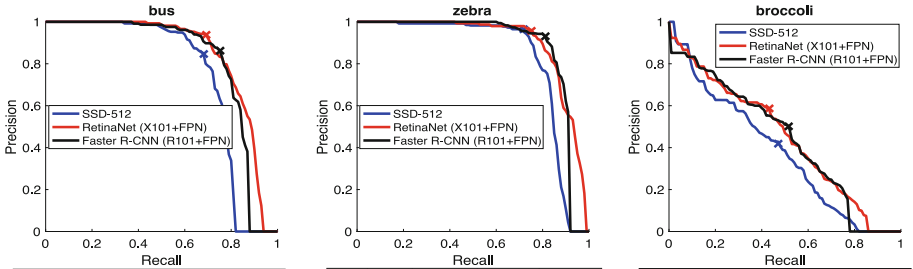


Fig. 4. Average RP curves of the common detectors.

Class-Based Comparison and Interpreting the Components: Now we analyze oLRP on a class-basis and look at the individual components to get a better feeling about the characteristics of methods – see Fig. 5. For all three classes, oLRP is determined at the RP pairs where there exists a sharp precision decrease on the top right part of the curve. Moreover, intuitively, these pairs provide a good balance between precision and recall. Considering the FP and FN components, one can infer the structure of the curve. For all methods, the "zebra" class has the sharpest RP curves which correspond to lower FP & FN error values. For example, Faster R-CNN has 0.069 and 0.188 FP and FN error values, respectively. Thus, without looking at the curve, one may consider that the peak of the curve resides at $1-0.069=0.931$ precision and $1-0.188=0.812$ recall. For the "broccoli" curve, a less sharp one, the optimal point is at $1-0.498=0.502$ and $1-0.484=0.516$ as precision and recall respectively. Similar to "zebra", these values suggest that the peak of the curve is around the center of the RP range. The localization component (oLRP_{IoU}) shows that the tightness of the boxes for the "bus" class is better than that of"zebra" for all detectors even though "zebra" has a sharper RP curve. For RetinaNet, average IoU is $1-0.106=0.894$ and $1-0.122=0.878$ for the "bus" and"zebra" classes respectively. With this analysis, we also see that it is easy to compare the tightness of the boxes among the methods and classes.



AP	oLRP	oLRP_{IoU}	oLRP_{FP}	oLRP_{FN}
0.742	0.550	0.129	0.154	0.318
0.840	0.478	0.106	0.062	0.307
0.810	0.488	0.119	0.138	0.247

AP	oLRP	oLRP_{IoU}	oLRP_{FP}	oLRP_{FN}
0.827	0.519	0.156	0.035	0.282
0.899	0.451	0.122	0.043	0.248
0.880	0.445	0.138	0.069	0.188

AP	oLRP	oLRP_{IoU}	oLRP_{FP}	oLRP_{FN}
0.375	0.864	0.261	0.581	0.529
0.444	0.813	0.218	0.413	0.567
0.436	0.814	0.227	0.498	0.484

Fig. 5. Example RP curves representing the optimal configurations marked with crosses. The curves are drawn for $\tau = 0.5$. The tables in the figures represent the performance of the methods with respect to AP and moLRP. The rows of the table correspond to SSD-512, RetinaNet (X101+FPN) and Faster R-CNN (R101+FPN) respectively. Note that unlike AP, lower scores are better for LRP.

Same mAP but Different Behaviors, Faster R-CNN vs. RetinaNet: Now we compare two detectors with equal AP in order to identify their characteristics using the components of moLRP; namely, RetinaNet (X101+FPN), a single shot detector and Faster R-CNN (R101+FPN), a two-step detector. Firstly, we use the box localization component ($\text{moLRP}_{\text{IoU}}$) in Table 1 to discriminate between these two detectors. The standard metric used in MS COCO aims to include the localization error by averaging over 10 mAP values. Since 1.8% difference for these two detectors is present in the mAP@0.5, one can infer that RetinaNet seems to produce more tight boxes. However, this inference is possible only by examining all 10 mAP results one by one and still it is not possible to quantize this tightness. In contrast, $\text{moLRP}_{\text{IoU}}$ directly suggests that, among all the detectors in Table 1, RetinaNet (X101+FPN) produces the tightest bounding boxes with an average tightness of $1 - 0.161 = 0.839$ in IoU terms.

Secondly, we compare the sharpness of the same two detectors, which are evidently different (Fig. 4). RetinaNet (X101+FPN) produces 486,108 bounding boxes for 36,781 annotations, whereas Faster R-CNN (R101+FPN) yields only 127,039 thanks to its RPN method. For RetinaNet, confidence scores of 57% of the detections are under 0.1, and 87% of them are under 0.25 (these values are 29% and 56% for Faster R-CNN), which generally causes RetinaNet to have lower or equal precision than Faster R-CNN throughout the recall domain except for the tail of the RP curve. In the tail of RetinaNet, owing to its large number of results, it has some precision even though that of Faster R-CNN drops to 0. Figure 5 illustrates this phenomenon, which is best observed in the "zebra" curve. Even though RetinaNet has higher AP than Faster R-CNN with 0.899 to 0.880, this AP difference originates from the large number of RetinaNet detections, which causes the better RP curve tail. This shallow curve-longer tail phenomenon

is observed to be more or less valid for more than 50 classes including the ones in Fig. 6. On the other hand, oLRP and thus moLRP do not favor these kind of detectors but the sharper ones as shown in Fig. 5, which causes Faster R-CNN (R101+FPN) to have lower Optimal LRP error for "zebra" class.

Overall, even though RetinaNet has the best bounding box localization, Faster R-CNN (R101+FPN) with the same AP has lower mean oLRP error. Moreover, considering the RP curve of these variants, Faster R-CNN is sharper than RetinaNet as shown in Fig. 4. This is also validated by the components with nearly equal moLRP_{FP} and difference in moLRP_{FN} in favor of Faster R-CNN. Similarly, both moLRP_{FP} and moLRP_{FN} for RetinaNet (R50+FPN) are greater than those of Faster R-CNN (R50) due to the same shallow curve-longer tail phenomenon, preventing its RP curves to be sharper. Again, what makes RetinaNet (R50+FPN) to have better performance regarding both mAP and moLRP is its strength to produce tight bounding boxes as shown in Table 1.

6.3 Better Threshold, Better Performance

In this experiment, we demonstrate a use-case where oLRP helps us to set class-specific optimal thresholds as an alternative to the naive approach of using a general threshold for all classes. To this end, we developed a simple, online video object detection framework where we use an off-the-shelf still-image object detector (RetinaNet-50 [13] trained on MS-COCO [14]) and built three different versions of the video object detector. The first version, denoted with B, uses the still-image object detector to process each frame of the video independently. The second and third versions, denoted with G and S, respectively, again use the still-image object detector to process each frame and in addition, they link bounding boxes across subsequent frames using the Hungarian matching algorithm [2] and update the scores of these linked boxes using a simple Bayesian rule (details of this simple online video object detector is given in the Supplementary Material). The only difference between G and S is that while G uses a validated threshold of 0.5 (see s^* of B in Table 2 and Fig. 1 in Supplementary Material for validation) as the confidence score threshold for all classes, S uses the optimal threshold per class which achieves the oLRP error. We test these three detectors on 346 videos of ImageNet VID validation set [28] for 15 object classes which also happen to be included in MS COCO.

AP vs. oLRP: We compare G with B in order to represent the evaluation perspectives of AP and oLRP – see Fig. 6 and Table 2. Since B is a conventional object detector, with conventional RP curves as illustrated in Fig. 6. On the other hand, in order to be faster, G ignores some of the detections causing its maximum recall to be less than that of B. Thus, these shorter ranges in the recall set a big problem in the AP evaluation. Quantitatively, B surpasses G by 7.5% AP. On the other hand, despite limited recall coverage, G obtains higher precision than B especially through the end of its RP curve. To illustrate, for the "boat" class in Fig. 6, G has significantly better precision after approximately between 0.5 and 0.9 recall even though its AP is lower by 6%. Since oLRP

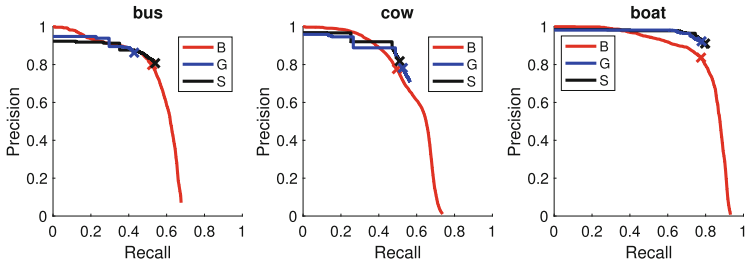


Fig. 6. Example RP curves of the methods. Optimal RP pairs are marked with crosses.

compares methods concerning their best configurations (i.e. the peak of their RP curves), this difference is clearly addressed comparing their oLRP error in which G surpasses S by 4.1%. Furthermore, the superiority of G is shown to be its higher precision since FN components of G and S are very close while FP component of G is 8.6% better, which is also the exact difference of precisions in their peaks of RP curves.

Therefore, while G seems to have very low performance in terms of AP, for 12 classes G reaches better peaks than B as illustrated by the oLRP values in Table 2. This suggests that oLRP is better than AP in capturing the performance details of the methods.

Table 2. Comparison among B, G, S with respect to AP & oLRP and their best class-specific configurations. The mean of class thresholds are assigned as N/A since the thresholds are set class-specific and the mean is not used. Note that unlike AP, lower scores are better for LRP.

	Method	airplane	bicycle	bird	bus	car	cow	dog	cat	elephant	horse	motorcycle	sheep	train	boat	zebra	mean
AP	B	**0.681**	**0.630**	**0.547**	**0.565**	**0.555**	**0.587**	**0.463**	**0.601**	**0.661**	**0.473**	**0.602**	**0.561**	**0.713**	**0.829**	**0.816**	**0.619**
	G	0.621	0.445	0.492	0.398	0.417	0.510	0.416	0.568	0.588	0.441	0.571	0.547	0.600	0.769	0.765	0.544
	S	0.645	0.535	0.500	0.485	0.419	0.492	0.434	0.569	0.589	0.444	0.573	0.545	0.609	0.792	0.782	0.561
oLRP	B	0.627	0.776	0.718	0.702	0.759	0.692	0.728	0.700	0.625	0.723	0.692	0.677	**0.583**	0.594	0.436	0.669
	G	0.606	0.783	0.691	0.727	**0.758**	0.679	0.714	**0.697**	0.614	**0.699**	**0.654**	**0.648**	0.586	0.553	0.432	0.656
	S	**0.603**	**0.762**	**0.687**	**0.688**	0.759	**0.678**	**0.712**	**0.697**	**0.613**	0.701	0.655	0.649	**0.583**	**0.551**	**0.425**	**0.651**
$oLRP_{IoU}$	B	0.182	0.271	0.169	0.177	0.207	0.145	0.166	0.203	0.170	0.155	0.192	0.154	0.159	0.199	0.128	0.179
	G	0.181	0.258	0.170	0.160	0.207	0.151	0.165	0.200	0.170	0.160	0.195	0.155	0.156	0.195	0.128	0.177
	S	0.186	0.270	0.170	0.173	0.207	0.148	0.170	0.200	0.170	0.160	0.194	0.155	0.159	0.197	0.131	0.179
$oLRP_{FP}$	B	0.080	0.228	0.300	0.203	0.303	0.224	0.242	0.248	0.095	0.246	0.158	0.141	0.099	0.163	0.034	0.184
	G	0.006	0.116	0.174	0.137	0.311	0.218	0.229	0.279	0.071	0.221	0.049	0.078	0.091	0.077	0.016	0.142
	S	0.087	0.226	0.184	0.193	0.320	0.182	0.269	0.283	0.075	0.231	0.084	0.078	0.110	0.089	0.030	0.163
$oLRP_{FN}$	B	0.383	0.427	0.478	0.477	0.499	0.504	0.533	0.394	0.395	0.540	0.448	0.494	0.344	0.224	0.220	0.424
	G	0.359	0.523	0.480	0.571	0.493	0.473	0.512	0.372	0.388	0.494	0.415	0.467	0.360	0.221	0.227	0.424
	S	0.326	0.389	0.489	0.461	0.488	0.490	0.480	0.369	0.385	0.493	0.406	0.468	0.339	0.203	0.202	0.398
s^*	B	0.38	0.31	0.44	0.27	0.49	0.61	0.42	0.49	0.49	0.52	0.45	0.51	0.41	0.45	0.31	N/A
	G	0.00	0.69	0.97	0.68	0.00	0.96	0.48	0.70	0.33	0.64	0.60	0.84	0.59	0.90	0.00	N/A
	S	0.00	0.54	0.98	0.45	0.00	0.91	0.49	0.64	0.39	0.58	0.63	0.85	0.55	0.89	0.54	N/A

Effect of the Class-Specific Thresholds: Compared to G, owing to the class-specific thresholds, S has 2.3% better mAP and 0.6% better moLRP as shown in Table 2. However, since the mean is dominated by s^* around 0.5, it is better to focus on classes with low or high s^* values in order to grasp the effect of the approach. The "bus" class has the lowest s^* with 0.27. For this class, S surpasses G by 8.7% in AP and 4.1% in oLRP. This performance increase is also observed for other classes with very low thresholds, such as "airplane","bicycle" and "zebra". For these classes with lower thresholds, the effect of class-specific threshold on the RP curve is to stretch the curve in the recall domain (maybe by accepting some loss in precision) as shown in the "bus" example in Fig. 6. Not surprisingly,"cow" is one of the two classes for which AP of S is lower since its threshold is the highest and thereby causing recall to be more limited. On the other hand, regarding oLRP, the result is not worse since this time the RP curve is stretched through the positive precision, as shown in Fig. 6, allowing better FP errors. Thus, in any case, lower or higher, the threshold setting method aims to discover the best RP curve. There are four classes in total for which G is better than S in terms of oLRP. However, note that the maximum difference is 0.2% in oLRP and these are the classes with thresholds around 0.5. These suggest that choosing class-specific thresholds rather than the common general thresholding approach increases the performance of the detector especially for classes with low or high class-specific thresholds.

7 Conclusion

We introduced a novel performance metric, LRP, as an alternative to the dominantly used AP. LRP has a number of advantages over AP, which we demonstrated in the paper: (i) AP cannot distinguish between very different RP curves whereas LRP, through its error components, provides a richer evaluation in terms of TP, FN and localization. (ii) AP not does have a localization component and one needs to calculate AP@τ with different τ values. However, LRP explicitly includes a localization error component ($1 - \text{oLRP}_{\text{IoU}}$ gives the mean localization accuracy of a detector). (iii) There are many practical use cases where one needs to set a detection threshold in order to obtain detections to be used in a subsequent stage. Optimal LRP provides a practical solution to this problem, which we demonstrated for online video object detection.
Supplementary Material. Supplementary material contains a detailed definition of AP, a result on the distribution of confidence thresholds, a description of the online detector and the proof that LRP is a metric.

Acknowledgements. This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) through project called "Object Detection in Videos with Deep Neural Networks" (project no 117E054). We also gratefully acknowledge (i) the support of NVIDIA Corporation with the donation of the Tesla K40 GPU and (ii) the computational resources kindly provided by Roketsan Missiles Inc. used for this research. Kemal Oksuz is supported by the TÜBİTAK 2211-A National Scholarship Programme for Ph.D. students.

References

1. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. EURASIP J. Image Video Process. **2008**(1), 246309 (2008). https://doi.org/10.1155/2008/246309
2. Bourgeois, F., Lassalle, J.C.: An extension of the munkres algorithm for the assignment problem to rectangular matrices. Commun. ACM **14**(12), 802–804 (1971)
3. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Gool, L.V.: Robust tracking-by-detection using a detector confidence particle filter. In: IEEE International Conference on Computer Vision ICCV (2009)
4. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Online multiperson tracking-by-detection from a single, uncalibrated camera. IEEE Trans. Pattern Anal. Mach. Intell. **33**(9), 1820–1833 (2011)
5. Brzezinski, D., Stefanowski, J., Susmaga, R., Szczech, I.: Visual-based analysis of classification measures with applications to imbalanced data. arXiv: 1704.07122 (2017)
6. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems, NIPS (2016)
7. Dembczynski, K.J., Waegeman, W., Cheng, W., Hüllermeier, E.: An exact algorithm for F-measure maximization. In: Advances in Neural Information Processing, NIPS, pp. 1404–1412 (2011)
8. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
9. Feichtenhofer, C., Pinz, A., Zisserman, A.: Detect to track and track to detect. In: IEEE International Conference on Computer Vision, ICCV (2017)
10. Ferrari, P.: A keras port of single shot multibox detector. https://github.com/pierluigiferrari/ssd_keras. Accessed 13 Mar 2018
11. Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., He, K.: Detectron. https://github.com/facebookresearch/detectron. Accessed 13 Mar 2018
12. Kang, K., et al.: T-CNN: tubelets with convolutional neural networks for object detection from videos. IEEE Trans. Circuits Syst. Video Technol. **PP**(99), 1 (2017)
13. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision, ICCV (2017)
14. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
15. Lipton, Z.C., Elkan, C., Naryanaswamy, B.: Optimal thresholding of classifiers to maximize F1 measure. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds.) ECML PKDD 2014. LNCS (LNAI), vol. 8725, pp. 225–239. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-44851-9_15
16. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
17. Lu, Y., Lu, C., Tang, C.: Online video object detection using association LSTM. In: IEEE International Conference on Computer Vision, ICCV (2017)
18. Musicant, D.R., Kumar, V., Ozgur, A.: Optimizing F-measure with support vector machines. In: The Florida Artifical Intelligence Research Society Conference, FLAIRS Conference (2003)

19. Oksuz, K., Cemgil, A.T.: Multitarget tracking performance metric: deficiency aware subpattern assignment. IET Radar Sonar Navig. **12**(3), 373–381 (2018)
20. Powers, D.M.W.: What the F-measure doesn't measure: features, flaws, fallacies and fixes. arXiv: 1503.06410 (2015)
21. Puthiya Parambath, S., Usunier, N., Grandvalet, Y.: Optimizing F-measures by cost-sensitive classification. In: Advances in Neural Information Processing, NIPS (2014)
22. Quevedo, J.R., Luaces, O., Bahamonde, A.: Multilabel classifiers with a probabilistic thresholding strategy. Pattern Recogn. **45**(2), 876–883 (2012)
23. Rahmathullah, A.S., Garcia-Fernandez, A.F., Svensson, L.: Generalized optimal sub-pattern assignment metric. In: IEEE International Conference on Information Fusion, FUSION (2017)
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, NIPS (2015)
25. Rijsbergen, C.J.V.: Information Retrieval. 2nd edn. Butterworth-Heinemann (1979)
26. Ristic, B., Vo, B.N., Clark, D.: Performance evaluation of multi-target tracking using the OSPA metric. In: IEEE International Conference on Information Fusion, FUSION (2010)
27. Ristic, B., Vo, B.N., Clark, D., Vo, B.T.: A metric for performance evaluation of multi-target tracking algorithms. IEEE Trans. Signal Process. **59**(7), 3452–3457 (2011)
28. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
29. Schuhmacher, D., Vo, B.T., Vo, B.N.: A consistent metric for performance evaluation of multi-object filters. IEEE Trans. Signal Process. **56**(8), 3447–3457 (2008)
30. Shi, X., Yang, F., Tong, F., Lian, H.: A comprehensive performance metric for evaluation of multi-target tracking algorithms. In: International Conference on Information Management, ICIM (2017)
31. Shu, G., Dehghan, A., Shah, M.: Improving an object detector and extracting regions using superpixels. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2013)
32. Stalder, S., Grabner, H., Van Gool, L.: Cascaded confidence filtering for improved tracking-by-detection. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6311, pp. 369–382. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15549-9_27
33. Suzuki, J., McDermott, E., Isozaki, H.: Training conditional random fields with multivariate evaluation measures. In: International Conference on Computational Linguistics and the Annual Meeting of the Association for Computational Linguistics, ACL-44 (2006)
34. Tripathi, S., Lipton, Z.C., Belongie, S.J., Nguyen, T.Q.: Context matters: refining object detection in video with recurrent neural networks. In: British Machine Vision Conference, BMVC (2016)

35. Vu, T., Evans, R.: A new performance metric for multiple target tracking based on optimal subpattern assignment. In: IEEE International Conference on Information Fusion, FUSION (2014)
36. Zhu, X., Dai, J., Yuan, L., Wei, Y.: Towards high performance video object detection. arXiv: 1711.11577 (2017)
37. Zou, X., Wen, J.: Detection of object security in crowed environment. In: IEEE International Conference on Communication Problem-Solving, ICCP (2015)