Quantifying Path Smoothness in Video Object Tracking by Detection

Mohammed Gasmallah[†,*], François Rivest[†,‡,*], Farhana Zulkernine[†,*], Mélanie Breton[◇]
[†] School of Computing, Queen's University, Kingston, Canada
[‡] Department of Mathematics and Computer Science, Royal Military College of Canada, Kingston, Canada
[◇] Defence Research and Development Canada, Valcartier, Canada

Abstract

Object detection and tracking are important areas of research in computer vision. Computer vision solutions to object detection are typically single-frame solutions. To perform tracking by detection, these solutions typically do object detection on a per-frame basis, thus losing any temporal information from previous frames. Many multi-object tracking solutions report the average precision performance on video datasets, but they do not evaluate the temporal qualities of these solutions. In video, not only the detection of objects is important but the temporal motion attributes of an object's path, such as its velocity, acceleration, and jerk, are important as well. Many implementations of Object Tracking by Detection systems have run into the problem of motion smoothing for bounding box paths. This paper focuses on quantifying the smoothness of detected object paths within some temporal window. We propose using two smoothness metrics from the field of biokinematics and adapt them for use with detections. Finally, using these metrics, we evaluate the ground truth and two popular object detectors, at the time of experimentation (YOLOv3 and Retinanet), on the entire MOT17 dataset. The results show that the metrics are useful in determining object smoothness, and provide us with an additional approach to evaluate an algorithm's performance in object tracking. The experiments also demonstrate that YOLOv3 produces smoother bounding boxes than Retinanet. All supplemental graphs and data are shown in our appendix[1].

Keywords: Object Detection, Motion, Tracking, Metrics, Evaluation

1. Introduction

As we progress in the era of real-time information processing, video object detection and tracking are becoming increasingly important areas of research in computer vision [1–4]. Object detection in machine learning has greatly increased in popularity with a variety of different approaches being taken to solve the problem such as those from You-Only-Look-Once (YOLO) and Faster R-CNN [3, 5]. The ability to detect objects and categorize them in a scene allows systems to make complex and important decisions regarding real-world objects.

Object detection performance using deep learning is often evaluated using the Mean Average Precision (mAP) after Intersection over Union (IOU) thresholding [6]. The mAP metric is useful for comparing the predicted boxes with the ground truth boxes. However, when applied to video, it does not provide any additional information about its temporal quality. In video, bounding box position and class are not the only information present and the quality of predictions can differ in a multitude of different ways. Many object tracking by detection systems have run into the problem of bounding box path smoothness and other forms of motion smoothness problems [7–9]. In this paper, we propose two smoothness metrics for use with object tracking by detection scenarios.

[1] Appendix hosted at:https://anonymous.4open.science/r/QPSMVOTD-062D/Object_Smoothness_Appendix.pdf

[*] {11mhg, farhana.zulkernine }@queensu.ca, francois.rivest@{rmc.ca, mail.mcgill.ca}

Object path smoothness is vital in qualitative assessments of object predictions in video and could be useful in determining which object detection methods should be used in production systems. Cognitive science research in object recognition has found that a smoothness constraint on the development of object recognition is necessary for enhanced colour/shape recognition and binding [10]. To quantify object smoothness, we go to the field of biokinematics for inspiration. Since smooth coordinated movements are often good characteristics for healthy human motor behaviour [11, 12], various smoothness metrics have been developed in the field of biokinematics to assess sensory-motor performance in patients [11]. We chose two metrics based on the findings of Balasubramanian et al. [12], namely Log Dimensionless Jerk (LDLJ) and Spectral Arc Length (SAL). We adapt these metrics for use in bounding box path scenarios and analyze the results of these metrics on the Multi-Object Tracking Dataset [13].

The main contributions are as follows:

- We propose the use of a temporal intersection over union (TIOU) to evaluate the speed of a bounding box path.
- We propose modifications to the Log Dimensionless Jerk metric for usage with object detection, and evaluate it using two object detection deep convolutional neural networks and the ground truth.
- We propose modifications to the Spectral Arc Length metric for usage with object detection, and evaluate it using two different object detection deep convolutional neural networks and the ground truth.
- We perform statistical tests on the results of applying both metrics to the entire Multi-Object Tracking (MOT) dataset to demonstrate that there are significant differences in the performance of each method.

The rest of this paper is organized as follows: Section 2 gives an overview of object detection networks and the two networks that we use in the experimentation. Section 3 discusses the object detection and tracking metrics that are most commonly used in the field. It also explains the adaptations made to the metrics. The methods, the experiments, and the testing of the metrics are reported in Section 4. We summarize the results and conclude in Section 5.

2. Related Work

2.1. Object Detection Networks

YOLO is a state-of-the-art, real-time object detection system for use on standard object detection tasks [3]. YOLO is a single-frame, one-stage object detection system that performs quick but accurate detections using a 53-layer feature extractor known as Darknet-53 [3, 14]. The YOLO version used in this paper is YOLOv3, which contains nine possible anchor boxes organized in groups of 3 [15]. Each group corresponds to a different scale, allowing a large variety of predictions in terms of bounding box scales [15]. This means that the final prediction tensor for YOLOv3 is of shape $[N, N, (3 * (4 + 1 + \text{num classes}))]$ where N denotes the number of grids used to divide the input image, 4 is the bounding box coordinates, and the additional value is the objectness score (a binary classification value indicating whether or not something is an object). Redmon and Farhadi [3] use $N = 13$. The predicted boxes are then extracted from this tensor using the equations proposed by Redmon et al. [15].

Retinanet is a state-of-the-art, one-stage object detector that shares many similarities with previous dense, two-stage object detectors such as Region-Proposal Network, and Fast-RCNN [16]. Retinanet focuses on using a feature pyramid network backbone and a novel focal loss to deal with the class imbalance in object detection datasets [16]. The feature pyramid backbone constructs efficient multi-scale features from a single image [16]. Retinanet uses 9 anchor boxes, grouped into 3 different scales of aspect ratios. This is comparable to the aspect ratio scales and anchor boxes that YOLOv3 uses. Since the time of starting experimentation, new object detectors have since been published, including newer versions of YOLO. In this paper, however, we focus on Retinanet with Resnet-50 (a 50-layer version of Resnet) and YOLOv3 as the baseline networks to evaluate our proposed metrics.

2.2. **Object Detection and Tracking Metrics**

One of the most commonly used evaluation metrics is the Intersection over Union (IOU) [6]. This metric is a useful way to determine true positives and false positives when comparing predictions against ground truths [2, 6]. The metric is often used to match predicted boxes with ground truth boxes based on IOU and then threshold boxes which are not close to any ground truth. A Generalized Intersection over Union (GIOU) metric has been developed by Rezatofighi et al. [6] to use an overlap metric as a regression loss. This is useful as it generalizes the overlap metric for non-overlapping bounding boxes.

For object tracking systems, the Multi-Object Tracking Accuracy (MOTA) is one of the most widely used metrics to evaluate performance [13]. MOTA is used in the Multi-Object Tracking (MOT) dataset challenge, although they indicate that it may not serve well as a single performance measure [13]. The MOTA was initially introduced by Stiefelhagen et al. [17] and is defined as:

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t GT_t} \tag{2.1}$$

where t is the frame index, GT is the number of ground truth objects, FN is the number of false negatives, FP is the number of false positives, and IDSW is the number of mismatched errors. The IDSW can be calculated by counting the number of times an object path switches identity based on ground truth.

Additionally, the Multiple Object Tracking Precision (MOTP) metric is another commonly used metric in tracking challenges. The MOTP denotes the average dissimilarity between true positives and the corresponding ground truth [13]. For bounding boxes, the MOTP is defined as:

$$MOTP = \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t} \tag{2.2}$$

where c_t is the number of matches in frame t, $d_{t,i}$ is the l_2 distance between all coordinates of the matched bounding box i with its assigned ground truth object.

Finally, in most object tracking scenarios, the only measure of trajectory quality is known as the track quality [13]. Track quality is classified into "mostly tracked", "partially tracked" or "mostly lost" categories based on a percentage measure of successful detections over an entire scene [13].

2.3. **Movement Smoothness**

Balasubramanian et al. defines movement smoothness as “a quality related to the continuality or non-intermittency of a movement, independent of its amplitude and duration” [12]. A smoothness measure is a metric that can be given a movement profile and should provide a valid, sensitive, reliable and practical measure [12]. We focus on the Log Dimensionless Jerk (LDLJ) and the Spectral Arc Length (SAL) because these are the only existing smoothness measures in kinematic motor control literature with these properties [12]. Balasubramanian et al. note that only SAL is reliable against measurement noise [12].

The Log Dimensionless Jerk (LDLJ) is one of the older, most frequently used smoothness measures [12] and is defined as:

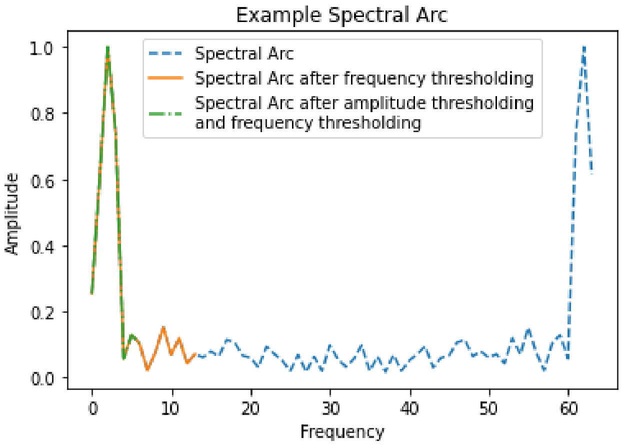


Figure 1. Example Spectral Arc after frequency thresholding and after amplitude thresholding.

$$DLJ = -\frac{(t_2 - t_1)^5}{v_{peak}^2} \int_{t_1}^{t_2} |\frac{d^2 v(t)}{dt^2}|^2 dt \tag{2.3}$$

$$LDLJ = -ln|DLJ| \tag{2.4}$$

where t_1 is the start time, t_2 is the end time, v_{peak} is the peak velocity, and $v(t)$ is the velocity at time t. The LDLJ is often used to quantify smoothness and coordination in kinematic tasks to analyze sensorimotor differences in stroke patients [11, 12]. However, Balasubramanian et al.[12] have found the LDLJ to be relatively sensitive to sensor noise.

The Spectral Arc Length (SAL) is a novel smoothness metric that is more reliable and robust than other previously used smoothness metrics [12]. The intuition behind this metric is that movements can be thought of as being composed of numerous low-frequency components and high-frequency interfering components [11]. Based on this concept, Balasubramanian et al. define the SAL as the negative arc length (length along a curve) of the amplitude and frequency-normalized Fourier Magnitude of the speed profile [11, 12]. The SAL has been used in kinematic research as well as in assessing surgical skills with regard to the surgeons' operational smoothness [18]. The SAL is defined as:

$$\eta_{sal} \triangleq -\int_0^{\omega_c} \sqrt{(\frac{1}{\omega_c})^2 + (\frac{d\hat{V}(\omega)}{d\omega})^2} d\omega \tag{2.5}$$

$$\hat{V}(\omega) \triangleq \frac{V(\omega)}{V(0)} \tag{2.6}$$

where ω_c is the frequency threshold, ω is the frequency, $\hat{V}(\omega)$ is the normalized amplitude of the speed profile at frequency ω, $V(\omega)$ is the amplitude of the speed profile at frequency ω, and thus $V(0)$ is the amplitude of the speed profile at frequency 0.

Note that SAL requires two hyper-parameters: a frequency threshold and an amplitude threshold. Balasubramanian et al. use a frequency threshold of $\omega_c = 40\pi rad/s$ and an amplitude threshold of 0.05. These values were tuned for patient trials in kinematic research and so may not work well for our purposes. An amplitude threshold of 0.05 only allowed for 1 frequency bin in our use case, and thus made the spectral analysis useless as we require a curve from which we could extract arc length (see subsection 4.4). An example spectral arc graph can be seen in Fig. 1.

3. Proposed Metrics

In this paper, we adapt the LDLJ and SAL to measure bounding box smoothness over time. We begin by constructing a speed profile for the proposed bounding boxes. We need a single metric that encompasses the smoothness of a box in terms of the bounding box position (x and y) and the change in scale (w and h). Since IOU encodes the shape properties of the objects compared to a region and gives a normalized measure of their area [6], we can use the formula $1 - IOU$ between bounding boxes at times t and $t+1$ to encode the speed profile of a box at time t. Therefore, we formulate the speed profile as a temporal IOU (TIOU) with the following equations:

$$\begin{aligned} v_{TIOU}(t) &= 1 - TIOU \\ v_{IOU}(t) &= 1 - \frac{|A_t \cap A_{t+1}|}{|A_t \cup A_{t+1}|} \end{aligned} \tag{3.1}$$

$$\begin{aligned} v_{TGIOU}(t) &= 1 - TGIOU \\ v_{TGIOU}(t) &= 1 - \left(TIOU - \frac{|C_t \setminus (A_t \cup A_{t+1})|}{|C_t|} \right) \end{aligned} \tag{3.2}$$

where $v_{TIOU}(t)$ is the TIOU at time t, $v_{TGIOU}(t)$ is the Temporal Generalized IOU (TGIOU) at time t, A_t is the bounding box at time t, and C_t is the smallest enclosing convex box for A_t and A_{t+1}.

If an object A is stationary and does not move between time t and $t+1$, the $v_{TIOU}(t) = 0$ and $v_{TGIOU}(t) = 0$. The maximal value for these metrics is based on the smallest possible overlap. These maximal values are $v_{TIOU}(t) = 1$ and $v_{TGIOU}(t) = 2$ because the IOU has a lower bound of 0 and the GIOU allows for a lower bound of -1 (where a value between 0 and -1 represents the distance of the bounding boxes from one another) [6]. We assume that a correct bounding box does not likely move beyond itself within one frame, considering 24 frames per second (fps) is the frame rate of most video cameras and of the MOT17 dataset videos. Therefore, we focus on the IOU formulation only. We demonstrate the effectiveness of using the TIOU as a measure of the bounding box movement profile in subsection 4.2 by plotting the TIOU of an object path and its smoothed variants.

3.1. Adapted Log Dimensionless Jerk

To use the TIOU as a speed profile in LDLJ, some modifications are required. First, a non-moving object would have a $v_{TIOU}(t) = 0$, which could lead the DLJ term in the ln of Eq.(2.4) to be zero. To prevent this we modify the LDLJ as follows:

$$LDLJ_{obj} = -ln|1 + DLJ| \tag{3.3}$$

This allows for $DLJ = 0$ and it does not greatly affect the LDLJ calculation. When comparing two objects, the object with the higher LDLJ is smoother (see Fig. 3a). As an additional modification, the v_{peak} is not set per trial and is instead set for all trials. Since we are using the TIOU as a measure of speed, there is a theoretical peak of 1 (the maximally overlapping bounding boxes results in a TIOU value of 1), and so we set $v_{peak} = 1.0$. Finally, we need to find an appropriate window length N, to perform the LDLJ calculation. Although the entire scene may be used, using a rolling window for the LDLJ calculation allows for an online measure of network performance in terms of bounding box prediction smoothness. This is illustrated in subsection 4.3.

3.2. Adapted Spectral Arc Length

To use the SAL with TIOU speed profiles, we take the discrete Fourier transform of $v_{TIOU}(t)$ profiles. We employ a sliding Discrete Fourier Transform (DFT) [19] to perform an online metric calculation. The sliding DFT requires a minimum of N samples (where N is the window length) before the DFT is valid, so we do not calculate the SAL for the first N samples. When comparing two objects, the one with the higher SAL is smoother (see Fig. 3b). Additionally, to adapt the SAL, we do not normalize per trial, as this leads to the inability to compare inter-trial results. To resolve this, we do not normalize by $V(0)$ in Eq.(2.6). Finally, as we are adopting this spectral arc length from another field, we perform some tests on the frequency and magnitude thresholds to find appropriate parameters by analyzing the effect of the parameters on the final SAL as explained in subsection 4.4.

4. Experiments

In this section, we go over our methodology and experimental results for evaluating and analyzing LDLJ and SAL. In subsection 4.2, we show that the TIOU ($v_{TIOU}(t)$) is a suitable speed profile for a bounding box. In subsection 4.3, we find the appropriate window length for LDLJ and SAL on bounding box smoothness calculations using a single object path as an evaluation of the hyperparameter. In subsection 4.4, we find the best amplitude and frequency thresholds that allow the most information to be collected for calculating the SAL. In subsection 4.5, we examine the intuition that the ground-truth path is the smoothest. Finally, in subsection 4.6, we analyze the performance of YOLOv3, Retinanet, and the Ground Truth object paths using the LDLJ and SAL.

4.1. Methods

To evaluate and analyze the feasibility of using LDLJ and SAL as smoothness metrics, we use the Multi-Object Tracking (MOT) Dataset as it contains a variety of scenes with a large variety of clearly labelled object paths [13]. Scenes in this dataset have a frame rate of 24fps. To match network detections to object paths, we use the IOU metric to find the closest match for each ground truth and assign the predictions accordingly. This method of assigning ground truths to predictions is how networks are trained in object detection internally [3, 14–16]. Finally, as all evaluations are on a sliding window, we use a stride of 4 frames as this is the minimum number of frames required for the jerk to be calculated.

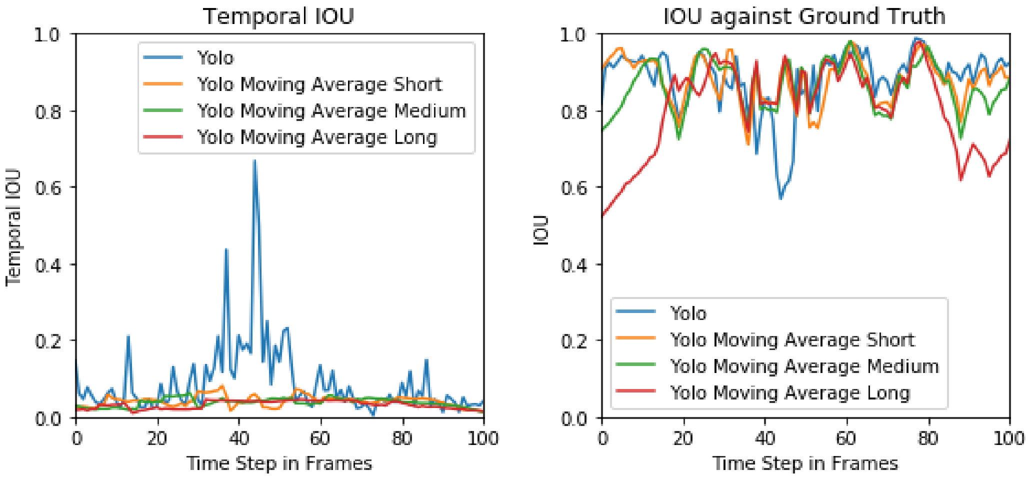
4.2. IOU Experimentation

Smoothness measures require a speed signal from which we can measure the smoothness of an object's path. Instead of using four signals (x, y, w, h), a smoothness measure that could encompass all four of these signals would be ideal. We did not directly use the norm of the change in these four signals as x, y operates on a different scale than w, h but instead, use the TIOU.

To demonstrate the effectiveness of the TIOU as the speed signal according to the definition given in section 3, we formulate an experiment using the path of object ID 1 in MOT17-09 predicted by YOLOv3. We compare these paths against paths generated from their moving average. If our smoothness measure is correct, larger moving averages will have smoother profiles. The moving average is defined as:

$$\mathrm{MA}_t = \frac{1}{w}\sum_{i=t-w}^{t} P_i \tag{4.1}$$

where t is the frame number, w is the window length, and P_i is the sample at frame i.



(a)

(b)

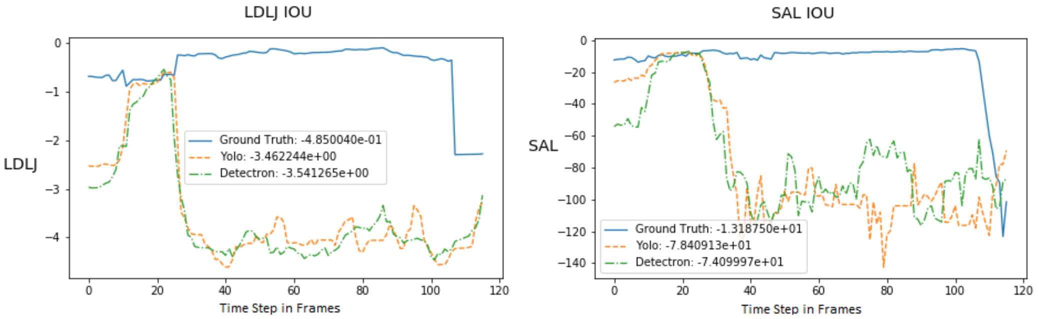
Figure 2. Plots of experiments with moving average on the effect of $v_{TIOU}(t)$ and IOU between predicted and ground truth.

The short moving average has a $w = 8$, the medium moving average has a $w = 16$ and the long moving average has a $w = 32$. All of these paths' center (x, y) values are plotted in Fig. 5 in Appendix A. In Fig. 2a, we plot the TIOU of these bounding box paths and in Fig. 2b, we plot the IOU of the predicted bounding box paths against their corresponding ground truth. From these figures, we see that the moving average reduces the spikes in movement in coordinate space, smoothing the TIOU plot correspondingly. This demonstrates the effectiveness of the TIOU as a speed signal for measuring the smoothness of an object's path.

4.3. **Determining Window Length**

To effectively use the LDLJ and SAL, we need to determine a suitable time window N that will encompass enough information about object paths. It should be noted that larger window lengths allow for more information about an object's motion characteristics. To determine a suitable time window, we begin by plotting the LDLJ and SAL at the following range of window lengths in terms of frames: {8, 16, 32, 64, 96, 128}.

The plots for these experiments can be seen in Figs. 6-7 in Appendix B. We note that the decrease in the smoothness of the last few collected points in the figures is due to ground truth moving in and out of frame. Based on these experiments we note that LDLJ is more sensitive to the window size than SAL. To choose a window size, we must balance local information with global information. A small window has much local information, but not enough global information about the movement profile of the bounding box. Similarly, a large window can often flatten out local information in favour of global information. Considering this balance, the window length we choose is 64 frames as this window allows for the ground truth to have changes in smoothness (allowing intratrial comparisons). We plot the graph of the moving LDLJ and SAL with a window of 64 frames on object path ID 1 in MOT17-09 in Figs. 3a-3b. The mean LDLJ and SAL values are shown in Tables 1-2.



(a) *(b)*

Figure 3. Plots for a window size of 64 on MOT17-02 object path 1 (higher is smoother). Results are shown for ground truth, YOLOv3 and Retinanet as a running plot.

Mean Log Dimensionless Jerk

Window Length	8 Frames	16 Frames	32 Frames	64 Frames	96 Frames	128 Frames
Ground Truth	$\mathbf{-1.63e^{-6}}$	$\mathbf{-2.20e^{-4}}$	$\mathbf{-1.83e^{-2}}$	$\mathbf{-4.85e^{-1}}$	**-1.82**	**-3.35**
YOLO	$-5.86e^{-5}$	$-6.88e^{-3}$	$-4.20e^{-1}$	-3.46	-6.01	-7.92
Retinanet	$-6.07e^{-5}$	$-7.16e^{-3}$	$-4.44e^{-1}$	-3.54	-6.10	-7.98

Table 1. Mean LDLJ values at a variety of window lengths (higher is smoother) for MOT17-02 object path 1.

Mean Spectral Arc Length

Window Length	8 Frames	16 Frames	32 Frames	64 Frames	96 Frames	128 Frames
Ground Truth	**-1.28**	**-2.20**	**-5.97**	**-13.19**	**-20.04**	**-27.28**
YOLO	-2.08	-8.88	-28.86	-78.41	-151.01	-246.27
Retinanet	-2.93	-8.95	-26.88	-74.10	-139.63	-215.67

Table 2. Mean SAL values at a variety of window lengths (higher is smoother) for MOT17-02 object path 1.

4.4. Determining Amplitude and Frequency Thresholds

SAL as defined in Eq.(2.5)-(2.6) requires two thresholding parameters [11]. Fig. 1 is an example of the SAL of a movement profile. These thresholds are useful in making SAL robust to noise [12]. We note, however, that the original parameters provided by [11] were not suitable for the object bounding box paths as they were initially found for patient sensorimotor trials.

We do a grid search on frequency threshold and amplitude threshold at a variety of ranges. In [11], they use 5 as their frequency threshold which corresponds to their sampling frequency of 100Hz, we began with this value and incremented it by 5 up until 35 as our corresponding sampling frequency is 24Hz. Any frequency bin beyond the frequency threshold tested is ignored to calculate SAL. For the amplitude threshold, we begin with an exceptionally small value of $1e^{-5}$ and in log scale, we increase this threshold until we reach $1e^{-1}$. We begin with this small value to allow more information into the SAL calculation and try to find the effect that increasing this threshold may have. Once the spectral arc reaches either the frequency threshold or the amplitude threshold, all other frequency bins are ignored.

The results of our experimentation are presented in Table 3 and a few insights are readily apparent. Firstly, we note that changes in the amplitude threshold are minimal. We choose an amplitude threshold of $1E^{-5}$ as a lower amplitude threshold is preferable in allowing more information to be included in the SAL calculation. Finally, the frequency threshold has a scaling effect on the Spectral Arc Length up until a threshold of 25. Frequency thresholds

	Amplitude Threshold					
		$1e^{-5}$	$1e^{-4}$	$1e^{-3}$	$1e^{-2}$	$1e^{-1}$
Frequency Threshold	5	-7.72	-7.72	-7.72	-7.72	-7.21
	10	-8.67	-8.67	-8.67	-8.65	-8.59
	15	-9.59	-9.59	-9.59	-9.56	-8.59
	20	-10.60	-10.60	-10.60	-10.56	-9.22
	25	-13.19	-13.19	-13.19	-13.13	-11.55
	30	-13.19	-13.19	-13.19	-13.13	-11.55
	35	-13.19	-13.19	-13.19	-13.13	-11.55

Table 3. Spectral Arc Length Threshold Experimentation on MOT17-02 object path 1 (higher is smoother) of the Ground Truth path. This is used to view the effect of parameters on the Spectral Arc Length to choose suitable parameters.

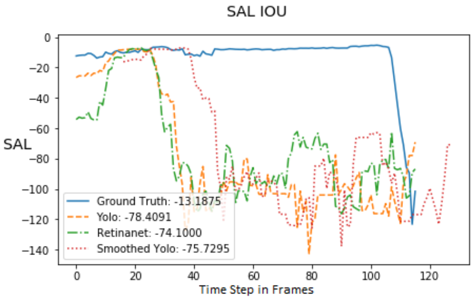


Figure 4. SAL plot for frequency threshold 25 and amplitude threshold $1e^{-5}$ of MOT17-02 object path 1.(Note higher is smoother)

beyond 25 frequency bins do not affect the SAL calculation. Due to these findings, we choose to use a frequency threshold of 25 and an amplitude threshold of $1e^{-5}$ for object path analysis. The plot for this set of hyperparameters is presented in Fig. 4 and the remaining set of hyperparameters can be found in Appendix C, Figs. 8-11. These experiments show that a frequency threshold of 25 and an amplitude threshold of $1e^{-5}$ are useful for bounding box path smoothness evaluation.

4.5. **Ground Truth Smoothness Analysis**

It may seem intuitive to believe that ground truth is the smoothest path however, we develop an experiment using the moving averages from subsection 4.2 and the ground truth of that very same object path to demonstrate otherwise. We plot the LDLJ and the SAL of those object paths using the hyperparameters chosen after experiments from subsections 4.3-4.4 in Figs. 6-11. This analysis of the moving average paths and the ground truth shows that the ground truth is not always the smoothest path and simply matching the ground truth does not necessarily lead to the smoothest bounding box motion characteristic.

4.6. **Complete Analysis of LDLJ and SAL in YOLO, and Retinanet**

Previous subsections present experiments on the path of object ID 1 in the MOT17-09 scene. In this subsection, we report the LDLJ and SAL averages on all object IDs in all scenes in the MOT Dataset. We plot the box plot of the LDLJ and SAL values on all object IDs in all scenes in MOT in Figs. 12-13 of Appendix D. The histogram plots of the LDLJ and SAL means are all shown in Appendix D Figs. 15-16.

If these metrics are indicative of object path smoothness, we would expect that the ground truth would be the smoothest. In our experimentation in subsection 4.6 with these

Metrics	LDLJ	SAL
Ground Truth	$\mathbf{-0.21} \pm 0.18$	$\mathbf{-50.82} \pm 15.33$
YOLO	-0.65 ± 0.64	-57.85 ± 12.04
Retinanet	-0.74 ± 0.74	-65.69 ± 10.96

Table 4. Mean LDLJ and mean SAL values for all networks as well as Ground Truth on all objects that are present in the ground truth, and predicted by YOLOv3 and Retinanet in all scenes of the Multi-Object Tracking Dataset. Note that for both LDLJ and SAL, higher is smoother.

Metric	F value	$P <$
SAL IOU	910.69	*0.0001*
LDLJ IOU	641.14	*0.0001*

Table 5. Results of one-way ANOVA test on the mean SAL and mean LDLJ of all objects in the ground truth, and predicted by YOLOv3 and Retinanet in all scenes of the MOT Dataset. This shows that both the SAL and the LDLJ generate distinct distributions based on the smoothness of the objects detected.

metrics, this expectation holds (see Table 4). Although LDLJ is known to be unreliable when affected by sensor noise [12], the object detection scenario has very little such noise. In Table 4 we see that both, the LDLJ and the SAL, claim that Retinanet is worse than YOLO for motion smoothness. Retinanet has many more object proposals than YOLO [15, 16] and thus may have more high-frequency noise in object paths for the long term. This may explain the reason for its lower smoothness metrics.

To show that the LDLJ and the SAL properly differentiate between the ground truth, YOLOv3, and Retinanet, we perform two one-way ANOVA tests on the mean SAL and the mean LDLJ of all object IDs that are present in the ground truth and predicted by YOLOv3 and Retinanet from the MOT Dataset. To maintain comparability, if an object is not predicted by either YOLOv3 or Retinanet, it is not included in the one-way ANOVA tests. The results of these ANOVA tests can be seen in Table 5. With $p < 0.0001$ for both SAL and LDLJ, we decided to conduct a multi-comparison post hoc test to determine which population means are significantly different from the others. As the population means were found to be normally distributed with D'Agostino's K-squared Test, we conducted a Tukey's Honest Significant Difference (HSD) Test with $\alpha = 0.05$. The results of this test on the LDLJ means can be seen in Table 6 and the results on the SAL means can be seen in Table 7. As the null hypothesis for Tukey's HSD test is that all population means are the same, we note that both LDLJ and SAL can differentiate all population means. This supports that both the LDLJ and SAL can be used as reliable ways to determine the smoothness of object paths generated.

Group 1	Group 2	Mean Diff	p adjusted	lower	upper	Reject H_0
Ground Truth	Retinanet	-0.53	0.001	-0.57	-0.50	**True**
Ground Truth	YOLOv3	-0.44	0.001	-0.48	-0.41	**True**
Retinanet	YOLOv3	-0.09	0.001	-0.05	0.13	**True**

Table 6. Multiple Comparison of LDLJ Means using Tukey HSD with α of 0.05. A reminder that H_0 is that all population means are the same.

Group 1	Group 2	Mean Diff	p adjusted	lower	upper	Reject H_0
Ground Truth	Retinanet	-14.87	0.001	-15.68	-14.05	**True**
Ground Truth	YOLOv3	-7.03	0.001	-7.84	-6.21	**True**
Retinanet	YOLOv3	7.84	0.001	7.02	8.66	**True**

Table 7. Multiple Comparison of SAL Means using Tukey HSD with α of 0.05. A reminder that H_0 is that all population means are the same.

5. Conclusion

In this paper, to quantify bounding box path smoothness, we adapted two smoothness metrics from the field of kinematics for use in object bounding box path analysis in object tracking by detection challenges. We show the process by which we adapt the smoothness metrics for bounding box path analysis, and show that these metrics can quantify object path smoothness. We provided implementation details for LDLJ and SAL on object paths in video, and we analyzed the window size for both metrics and found the best hyperparameters for SAL (subsections 4.3-4.4) for this purpose. Finally, we compare and analyze the results of using these metrics on all objects, in all scenes of the MOT17 dataset after detection by both YOLOv3 and Retinanet. These results showed that the proposed metrics, LDLJ and SAL adapted for temporal IOU, can differentiate multi-object tracking systems by their smoothness and that YOLOv3 tends to produce smoother bounding boxes than Retinanet.

In the future, we plan to investigate the differentiability of these metrics to use them for regularization in object tracking by detection systems such as recurrent video object detectors. This would enable a system that not only detects objects but attempts to predict smoother object paths. Experimental results have found that animals can better recognize objects with smoother input [10]. This suggests that learning from temporal input, instead of static frames, can improve object recognition in machine vision systems. Similarly, biasing the production of smooth predictions through smoothness regularization may improve learning in object detection systems.

Acknowledgements

This project was supported by Defence Research and Development Canada (DRDC).

Appendix A.

Our appendix is currently being hosted at this link: https://anonymous.4open.science/r/QPSMVOTD-062D/Object_Smoothness_Appendix.pdf.

References

[1] L. Leal-Taixé, A. Milan, K. Schindler, D. Cremers, I. Reid, and S. Roth. "Tracking the Trackers: An Analysis of the State of the Art in Multiple Object Tracking". PhD thesis. Technical University Munich, 2017. URL: http://arxiv.org/abs/1704.02781.

[2] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. "Microsoft COCO: Common objects in context". In: *Lecture Notes in Computer Science* 8693 LNCS (2014), pp. 740–755. ISSN: 16113349. DOI: 10.1007/978-3-319-10602-1{_}48.

[3] J. Redmon and A. Farhadi. "YOLO9000: Better, faster, stronger". In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* (2017), pp. 6517–6525. ISSN: 0146-4833. DOI: 10.1109/CVPR.2017.690. URL: http://arxiv.org/abs/1612.08242.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[5] S. Ren, K. He, R. Girshick, and J. Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), pp. 1137–1149. ISSN: 01628828. DOI: 10.1109/TPAMI.2016.2577031.

[6] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese. "Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression". In: (2019). URL: http://arxiv.org/abs/1902.09630.

[7] A. Mhalla, T. Chateau, and N. E. B. Amara. "Spatio-temporal object detection by deep learning: Video-interlacing to improve multi-object tracking". In: *Image and Vision Computing* 88 (2019), pp. 120–131. ISSN: 02628856. DOI: 10.1016/j.imavis.2019.03.002. URL: https://doi.org/10.1016/j.imavis.2019.03.002.

[8] Z. Pan and C. W. Ngo. "Moving-object detection, association, and selection in home videos". In: *IEEE Transactions on Multimedia* 9.2 (2007), pp. 268–279. ISSN: 15209210. DOI: 10.1109/TMM.2006.887992.

[9] D. Park, C. L. Zitnick, D. Ramanan, and P. Dollar. "Exploring weak stabilization for motion feature extraction". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1.c (2013), pp. 2882–2889. ISSN: 10636919. DOI: 10.1109/CVPR.2013.371.

[10] J. N. Wood. "A smoothness constraint on the development of object recognition". In: *Cognition* 153 (2016), pp. 140–145. ISSN: 00100277. DOI: 10.1016/j.cognition.2016.04.013. URL: http://dx.doi.org/10.1016/j.cognition.2016.04.013.

[11] S. Balasubramanian, A. Melendez-Calderon, and E. Burdet. "A robust and sensitive metric for quantifying movement smoothness". In: *IEEE Transactions on Biomedical Engineering* 59.8 (2012), pp. 2126–2136. ISSN: 00189294. DOI: 10.1109/TBME.2011.2179545.

[12] S. Balasubramanian, A. Melendez-Calderon, A. Roby-Brami, and E. Burdet. "On the analysis of movement smoothness". In: *Journal of NeuroEngineering and Rehabilitation* 12.1 (2015), pp. 1–11. ISSN: 17430003. DOI: 10.1186/s12984-015-0090-9. URL: http://dx.doi.org/10.1186/s12984-015-0090-9.

[13] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, K. Schindler, A. Milan, L. Leal-taix, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler. "MOT16: A Benchmark for Multi-Object Tracking". In: (2016), pp. 1–12. URL: http://arxiv.org/abs/1603.00831.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. "You Only Look Once: Unified, Real-Time Object Detection". In: (2015). ISSN: 01689002. DOI: 10.1109/CVPR.2016.91. URL: http://arxiv.org/abs/1506.02640.

[15] J. Redmon and A. Farhadi. "YOLOv3: An Incremental Improvement". In: (2018). URL: http://arxiv.org/abs/1804.02767.

[16] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. "Focal Loss for Dense Object Detection". In: *Proceedings of the IEEE International Conference on Computer Vision*. Vol. 2017-Octob. 2017, pp. 2999–3007. ISBN: 9781538610329. DOI: 10.1109/ICCV.2017.324.

[17] R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo. *The CLEAR 2007 evaluation*. Vol. 4625 LNCS. April. 2008, pp. 3–34. ISBN: 9783540695684. DOI: 10.1007/978-3-540-68585-2{_}1.

[18] W. H. Jantscher. "Using Real-Time Smoothness Metrics to Deliver Haptic Performance Cues for a Dexterous Task". PhD thesis. RICE University, 2018. DOI: 10.1192/bjp.111.479.1009-a.

[19] R. Lyons and E. Jacobsen. "The sliding DFT". In: *IEEE Signal Processing Magazine* 20.2 (2003), pp. 74–80. ISSN: 1053-5888. DOI: 10.1109/MSP.2003.1184347.