ASSESSING AI POLYP DETECTION DELAY USING TRACKLETS

Samuel Gussman-Toh[1], Wei-Chung Cheng[2], Nicholas Petrick[2], Feng Yang[2]

[1]Georgia Institute of Technology, sgussman3@gatech.edu
[2]Centers for Devices and Radiological Health, U.S. Food and Drug Administration, U.S.A.

ABSTRACT

This study introduces a tracklet-based approach for estimating the polyp detection delay for a colon CADe system as part of overall standalone performance testing. We propose using *detection delay* as a benchmarking metric that is agnostic to processing time and dataset modifications. Our study addresses how polyps frequently come in and out of frame, creating distinct tracklets, and examines the impact this behavior can have on detection delay calculations. Lastly, we introduce detection delay curves to compare model performance over a range of sensitivities and use detection window curves to measure how sensitivity improves the longer a model observes a polyp.

***Index Terms*—** assessment, detection delay, computer aided detection, CADe, regulatory science, colonoscopy

1. INTRODUCTION

Real-time computer aided detection (CADe) algorithms for colonoscopy (colon CADe) are AI-enabled medical devices that help endoscopists reduce missed polyps by identifying potential polyps in real-time during endoscopy procedures. Numerous colon CADe devices have been developed for clinical use and studies demonstrate their ability to improve adenoma detection rate (ADR) [1], [2]. Arguably, the most clinically important time to detect a polyp is the first time it comes into view to alert the clinician. In procedure videos, subsequent appearances of the polyp could be due to repositioning the endoscope as part of diagnosis. This ties the clinical utility of colon CADe devices not just to if a polyp is detected, but to how quickly it is identified.

Standalone studies benchmarking colon CADe devices use clinically relevant performance metrics, enable objective comparisons, and mirror relevant clinical conditions by testing on full procedure videos (not snippets).

Detection delay is not consistently assessed and is often omitted in clinical and standalone studies in literature [3]. Studies that report temporal performance use inconsistent terms and metrics [4], [5], [6]. Frequently, they only report real-time performance metrics (e.g. frames per second or latency/lag). While these metrics have important clinical and usability implications, they often focus only on hardware and compute dependent delays. This makes it difficult to compare results across studies and generally does not address how long a polyp must be visible before the AI identifies it. In standalone studies that reported detection delay, it was referred to by a variety of terms (e.g., detection latency) and was calculated as the total elapsed time between a polyp's first appearance and AI detection [5]. However, polyps frequently leave the frame and may not be visible for the entire period before detection [6]. Using first appearance to detection may penalize models when the polyp is not visible and may not accurately reflect a model's performance. Current studies generally assess colon CADe performance at either the frame-level, the polyp-level (how well the model does at detecting each polyp), or the video level (per procedure) [7]. However, each polyp appearance could also be considered a separate instance for detection. We introduce *tracklet*-level assessment here to offer another way to evaluate colon CADe device performance.

Our study augments the assessment of real-time colon CADe performance by proposing (1) a standard *detection delay* metric and testing methods, and (2) an appropriate dataset for benchmarking performance. Our metric and methods measure how quickly models detect each polyp overall, every time it reappears in frame, and the first time it appears in frame. We also compare the differences in polyp-level detection delay when calculated using *all frames* vs. only using *visible frames*. Lastly, we consider how quickly models' sensitivity improves the longer they view a polyp.

2. METHODS

We define two terms for our study: detection delay and tracklets. *Detection delay* is the elapsed time between a polyp appearing and being detected by the CADe model, irrespective of processing time. We calculated detection delay using the number of frames between appearance and detection, but we convert from frames to seconds (sec) using a video's frame rate. This allows direct comparisons between videos with different frame rates.

A *Tracklet* is a consecutive series of frames where the same object is visible [8]. Polyps frequently come in and out of frame as the clinician repositions the endoscope, dividing each polyp into multiple tracklets. This is illustrated in **Fig. 1**, which shows an example from our dataset. Our study

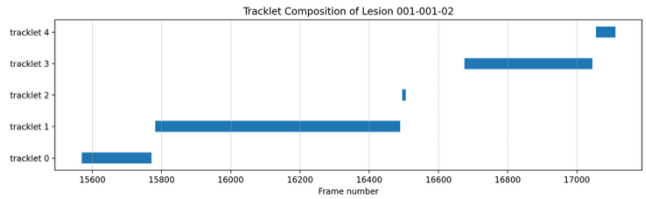

Fig. 2. Illustration of the five tracklets with varying gaps for polyp #2 in video 001-001.

assessed the temporal performance at both the polyp-level and the tracklet-level. We also compare the average detection delay of CADe models on the *first tracklet* vs. *all tracklets*

We assess the temporal performance of two AI CADe models to detect polyps in colonoscopy videos by fine tuning pre-trained YOLOv5 and YOLOv10 models using the combined SUN, PolypGen, KUMC2021, and annotated LDPolypVideo datasets.

A testing dataset was created from the REAL-Colon dataset, which contains 60 annotated full-procedure colonoscopy videos from four clinical sites [6]. Each video was trimmed to only the withdrawal phase of the procedure to focus on the clinically relevant portion of the procedure [9]. In our study, detection delay is assessed only on videos that contained histologically verified polyps. This resulted in 43 videos with 100 polyps visible in 268,037 of 1,409,455 frames. We calculated various metrics describing our test dataset, including the total duration of each polyp using all frames between the first and last visible frame, and the duration of each polyp in only frames where the polyp was visible. We created ground-truth tracklets for each polyp, measuring their length (in frames), the gaps between them, and the number of tracklets per polyp. We analyzed all tracklets and first tracklets. For all metrics, we calculated the mean (μ), standard deviation (σ), and quartiles (Q_{25}, Q_{50}, Q_{75}). To visualize the dataset, we created a histogram showing the lengths of the tracklets and the number of tracklets per polyp.

The models were evaluated on all videos and predictions were saved without applying a confidence threshold. An intersection-over-union (IoU) threshold of 0.1 was used to define true-positive detections. Different confidence thresholds were then applied to calculate the average detection delay at various levels of frame sensitivity.

The average polyp-level and tracklet-level detection delays are reported for the dataset. The polyp-level detection delay is calculated (1) using all frames between the polyp's first appearance and detection by the model including gaps (*all frames*), and (2) using only the frames where the polyp was visible excluding gaps (*visible frames*). Tracklet-level detection delay is defined as the elapsed time between the start of a tracklet and detection. Delay curves were created by plotting average detection delays over a range of sensitivities.

Lastly, we introduce a *detection window* to measure how a model's sensitivity changes as a function of polyp observation time. A detection window is a time window (*t*), where a true positive detection is only counted if it occurs within the first *t* seconds. Using a series of *t* values, we plot a detection window curve illustrating how sensitivity changes with observation time.

3. RESULTS

The average polyp duration was calculated for *all frames* and *visible frames* using every polyp in the test dataset. For *all frames*, the average polyp ground truth (GT) duration was μ = 151.2 sec (σ = 142.3 sec), with quartiles 67.7 sec (Q_{25}), 113.4 sec (Q_{50}), and 163.1 sec (Q_{75}). The shortest duration was 2.8 sec and the longest was 1020.4 sec. Using only visible frames, the average polyp duration was μ = 100.4 sec (σ = 90.7 sec), with quartiles 52.3 sec (Q_{25}), 78.6 sec (Q_{50}), and 113.8 sec (Q_{75}). The shortest duration was still 2.8 sec, while the longest duration decreased 24.1% to 774.1 sec.

The 100 polyps in the test dataset were composed of 2438 individual tracklets. The average number of tracklets per polyp was μ = 24.38 (σ = 25.13), with quartiles 10.75 (Q_{25}), 17 (Q_{50}), and 28.25 (Q_{75}). The minimum number of tracklets per polyp was 1 and the maximum was 169. **Fig. 2.** shows the tracklet distribution illustrating the range and varying number of tracklets for the polyps in our dataset.

Figure 3 shows the GT tracklet duration distribution for all tracklets (blue) and first tracklets (orange). The average tracklet duration was μ = 4.11 sec (σ = 7.58 sec), with quartiles 0.37 sec (Q_{25}), 1.47 sec (Q_{50}), and 4.28 sec (Q_{75}). The shortest duration was 1 frame (0.03 sec) and the longest was 75.71 sec. When only considering the first tracklets, the average tracklet duration is μ = 4.38 sec (σ = 6.94 sec), with quartiles 0.77 sec (Q25), 1.72 sec (Q50), and 3.73 sec (Q75). The shortest tracklet was also only 1 frame (0.04 sec) while the longest was 42.12 sec.

The average gap between GT tracklets is μ = 2.17 sec (σ = 10.01), with quartiles 0.08 sec (2 frames, Q_{25}), 0.36 sec (9 frames, Q_{50}), and 1.16s (31 frames, Q_{75}). The shortest gap was 0.03 sec (1 frame), and the longest gap was 228.5 sec (6849 frames). The gaps for the first tracklet were generally longer. The average gap between the first and second tracklet was μ

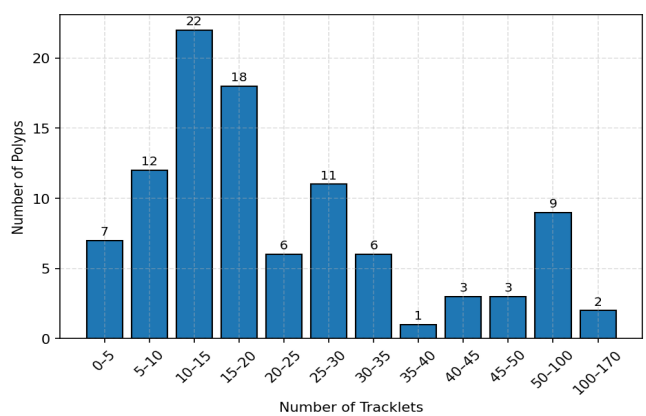

Fig. 1. Distribution of per-polyp GT tracklets in our dataset. It included a wide range in number of tracklets.

= 2.29 sec (σ = 9.85 sec), with quartiles 0.23 sec (6.5 frames, Q_{25}), 0.60 sec (15 frames, Q_{50}), and 1.35 sec (38.5 frames, Q_{75}). The shortest gap is the same (0.03 sec, 1 frame), while the longest gap is 96.72 sec (2418 frames).

The detection delay results were highly similar for both YOLOv5 and YOLOv10, so only the results from YOLOv5 are included here. **Fig. 4.** shows YOLOv5's polyp-level detection delay curves calculated using *all frames* and *visible frames*. At a frame sensitivity of 0.9, the model detects all 100 polyps with an average detection delay of 0.54 sec (*all frames*) and 0.32 sec (*visible frames*) – making the detection delay 69% larger when calculated using *all frames*. Over the entire range of sensitivities, the difference between the detection based on *all frames* and *visible frames* decreased as sensitivity increased. As a percentage, the difference varied from 48.8% (sen 0.7) up to 111.8% (sen 0.1).

YOLOv5 achieved average detection delay for tracklets (not polyps) of 0.07 sec at a frame sensitivity of 0.9. The average tracklet-level detection delay increased to 0.12 sec for the first tracklets. The detection delay curve for all tracklets and first tracklets is in **Fig. 5.**, showing that the average detection delay was longer for the first tracklet at well-performing levels of frame sensitivity. This broke down below a sensitivity of 0.3 (0.4 for YOLOv10), but there is a substantially smaller (24:1) sample size for the first tracklets.

The detection window curve of YOLOv5 is shown in **Fig. 6.** for both polyps and tracklets. Tracklet sensitivity plateaus at 0.93 after 1.2 sec of a tracklet appearing, while polyp sensitivity plateaus at 0.96 after 2.4 sec and slowly increases to 0.99 at a 9 sec window.

4. DISCUSSION

Because polyps are made of distinct tracklets, measuring the time a polyp is visible varies greatly if using all frames (as frequently done to calculate detection delay) or just visible frames, with a 34.8 sec difference between median times. We assessed the impact this has on polyp-level detection delay by creating two detection delay curves for comparison (**Fig. 4.**). The figure shows that the difference between the curves decreases as frame sensitivity improves. At a frame sensitivity of 0.9, the resulting delay is 69% larger using all frames compared with using only visible frames, even though this equals only a 0.22 sec difference (roughly 6 frames). The convergence of the curves is intuitive, because the differences disappear when the polyp is detected in the first tracklet (which is more likely with higher frame sensitivities). Still, it may be preferable to use visible frames when calculating average detection delay because it may represent the underlying AI performance better.

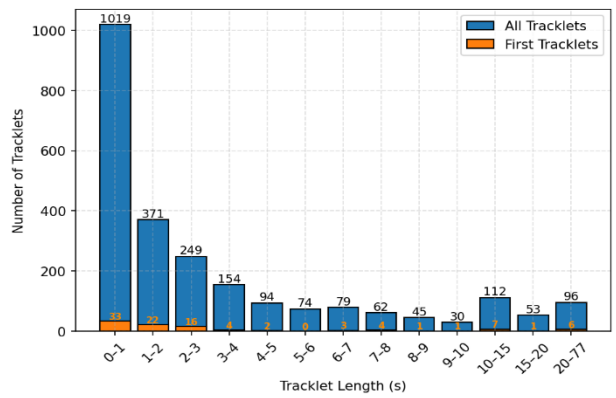

Fig. 3. Distribution of GT tracklet lengths. The numbers above the bars give the count of tracklets in that range. The blue bars show the overall number of tracklets, while the orange boxes illustrate only the first tracklets.

Polyps are composed of numerous tracklets – with 40% of polyps in our testing dataset containing between 10 and 20 tracklets. Tracklets tend to be short (41.8% are ≤1 sec), and over 50% of tracklets ≤2 sec. The first tracklets are similar to all tracklets in length. Because first tracklets are the most important opportunity for detection, we created tracklet-level detection delay curves for first tracklets and compared it to the curves for all tracklets. The curves show similar performance and converge as frame sensitivity improves. There is some instability in the first tracklet curve at lower sensitivities, but this is likely due to the reduced sample size.

Our detection window results may help to benchmark the clinically relevant sensitivity of a model – while our YOLOv5 achieves a polyp sensitivity of 0.99, this requires

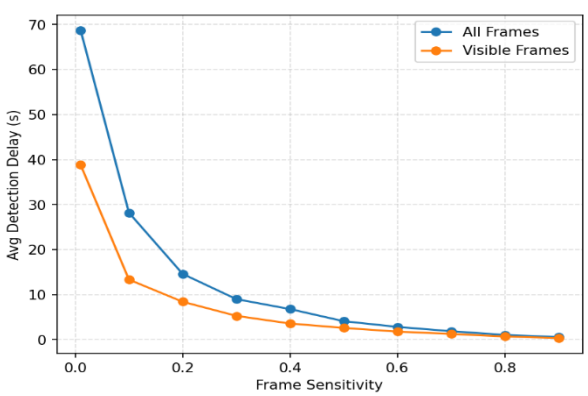

Fig. 4 The YOLOv5 polyp-level detection delay curve calculated for frame sensitivities ranging from 0.01 to 0.9.

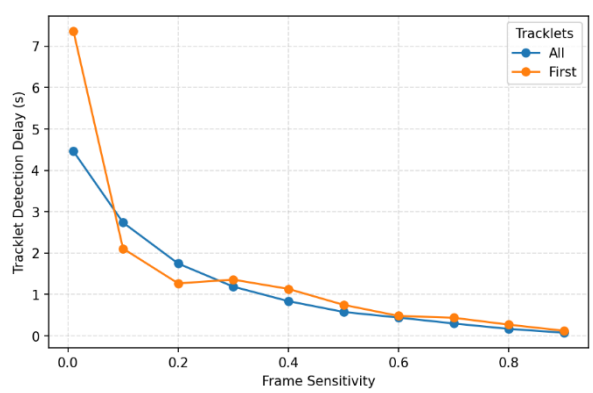

Fig. 5 The YOLOv5 tracklet-level detection delay curves for all tracklets (orange) and first tracklets (blue).

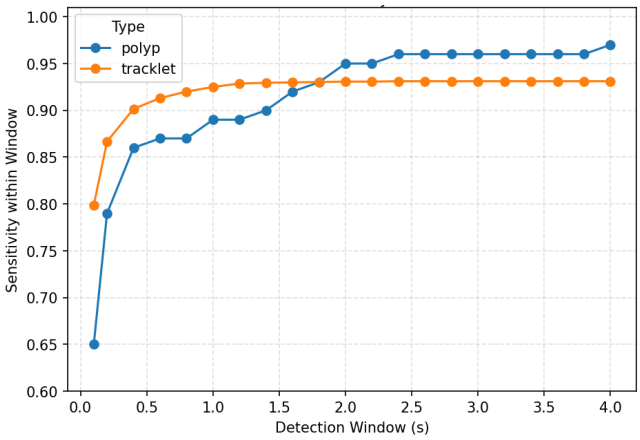

Fig. 6 The detection window curves for YOLOv5 calculated at a frame sensitivity of 0.9. The blue line shows polyp detection sensitivity and the orange curve shows tracklet sensitivity.

the polyp to be visible for ~9 sec (likely well past the point of clinical relevance). Instead, the plots show our YOLOv5 model can achieve a polyp sensitivity of 0.95 within a potentially more meaningful 2 sec. Similarly, the model achieves a good tracklet sensitivity of 0.93 after 1.2 sec, which is shorter than the median length of both all tracklets and first tracklets.

One limitation of this work is that it focused only on the connection between AI sensitivity and the detection delay without considering false positives in the analyses, despite the likely tradeoff between detection delay and the false positive rates in practice. Another limitation of our experiments is that we did not assess AI algorithms that aggregate predictions over multiple frames; instead, we treated each frame individually. Aggregating across frames may also impact the detection delay for an AI algorithm.

5. CONCLUSIONS

Detection delay is an important performance consideration in real-time colon CADe devices. In this study, we investigate *detection delay* as a potential metric for assessing an AI model's performance agnostic to both hardware latency and processing time. In addition, we introduce the concept of assessing delay based on a polyp's visible time (rather than using the total time between first appearance and detection) because we observed that polyps are often only visible within multiple short intervals (tracklets) during a colonoscopy exam. Calculating average detection delay for both polyps and tracklets may improve our understanding of a model's ability to both detect polyp early and across different instances in which it appears.

6. COMPLIANCE WITH ETHICAL STANDARDS

The work used only publicly available data sources and followed appropriate ethical standards in conducting this research and in writing the manuscript and followed all applicable laws and regulations regarding treatment of animals and human subjects.

7. ACKNOWLEDGMENTS

This article reflects the views of the authors and does not represent the views or policy of the U.S. Food and Drug Administration, the Department of Health and Human Services, or the U.S. Government. The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services.

8. REFERENCES

[1] A. Hann, J. Troya, and D. Fitting, "Current status and limitations of artificial intelligence in colonoscopy," *United Eur. Gastroenterol. J.*, vol. 9, no. 5, pp. 527–533, Jun. 2021, doi: 10.1002/ueg2.12108.

[2] I. Barua *et al.*, "Artificial intelligence for polyp detection during colonoscopy: a systematic review and meta-analysis," *Endoscopy*, vol. 53, no. 3, pp. 277–284, Mar. 2021, doi: 10.1055/a-1201-7165.

[3] P. Wang *et al.*, "Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study," *Gut*, vol. 68, no. 10, pp. 1813–1819, Oct. 2019, doi: 10.1136/gutjnl-2018-317500.

[4] A. Krenzer *et al.*, "A Real-Time Polyp-Detection System with Clinical Application in Colonoscopy Using Deep Convolutional Neural Networks," *J. Imaging*, vol. 9, no. 2, p. 26, Jan. 2023, doi: 10.3390/jimaging9020026.

[5] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information," *IEEE Trans. Med. Imaging*, vol. 35, no. 2, pp. 630–644, Feb. 2016, doi: 10.1109/TMI.2015.2487997.

[6] J. González-Bueno Puyal *et al.*, "Polyp detection on video colonoscopy using a hybrid 2D/3D CNN," *Med. Image Anal.*, vol. 82, p. 102625, Nov. 2022, doi: 10.1016/j.media.2022.102625.

[7] C. Biffi *et al.*, "REAL-Colon: A dataset for developing real-world AI applications in colonoscopy," *Sci. Data*, vol. 11, no. 1, p. 539, May 2024, doi: 10.1038/s41597-024-03359-0.

[8] G. Marchese Aizenman, P. Salvagnini, A. Cherubini, and C. Biffi, "Assessing clinical efficacy of polyp detection models using open-access datasets," *Front. Oncol.*, vol. 14, p. 1422942, Aug. 2024, doi: 10.3389/fonc.2024.1422942.

[9] L. Parolari, A. Cherubini, L. Ballan, and C. Biffi, "Towards Polyp Counting In Full-Procedure Colonoscopy Videos," Feb. 14, 2025, *arXiv*: arXiv:2502.10054. doi: 10.48550/arXiv.2502.10054.

[10] C. Hassan *et al.*, "Computer-aided detection-assisted colonoscopy: classification and relevance of false positives," *Gastrointest. Endosc.*, vol. 92, no. 4, pp. 900-904.e4, Oct. 2020, doi: 10.1016/j.gie.2020.06.021.