Automated Depth Sensing-Based Computer Vision for Dog Tail Wagging Interpretation

Devon Martin
Dept. Electrical Engineering
North Carolina State University
Raleigh, NC, USA
dfmartin43@gmail.com

Jeremy Park
Dept. Computer Science
North Carolina State University
Raleigh, NC, USA
jipark@ncsu.edu

Megan Carson
North Carolina State University
Raleigh, NC, USA
megash91@gmail.com

Margaret Gruen
College of Vetenary Medicine
North Carolina State University
Raleigh, NC, USA
megruen@ncsu.edu

Alper Bozkurt
Dept. Electrical Engineering
North Carolina State University
Raleigh, NC, USA
aybozkur@ncsu.edu

David L. Roberts
Dept. Computer Science
North Carolina State University
Raleigh, NC, USA
robertsd@csc.ncsu.edu

Abstract

Visual cues are commonly used by many animals, both intrinsically and explicitly, to communicate. Understanding and deciphering animal behavior and communication has been an area of active research, especially to assess emotions and mood. Dogs have been one of the most studied animals thanks to their integration into every aspect of human life. Objective measurement of dogs' behavioral communications is increasingly of interest. Remote cameras, computer vision, and signal processing techniques offer a non-contact system for objectively characterizing tail wag—a behavior commonly believed to be important for dog communication. Cameras do not compromise subject comfort or the mechanics of behavioral signals of interest. This study focuses on the tail as an indicator of emotional state and expands an existing Mask R-CNN computer vision methodology to derive detailed tail wag metrics across a population of 30 dogs in the presence (or absence) of certain stimuli. We updated the existing work with several thousand additional training images to make it more robust, at the cost of increased false positives. We have shown that this method works efficiently enough on most of videos in our data set to capture the tail wag signal in spite of streaming and detection difficulties. A good correspondence between these metrics and the video footage was observed. Our approach enables extracting tail position in three dimensions and deriving temporal metrics like speed and momentum. These novel capabilities allowed for performance of broad population statistical tests which revealed certain tail wag metrics to be different in the presence of certain visual stimuli. The findings of this study validate the potential for computer vision to provide higher resolution monitoring and continuous interpretation of dog tail movements and positions.

ACI 2024, December 02–05, 2024, Glasgow, United Kingdom

ACM ISBN 979-8-4007-1175-6/24/12
https://doi.org/10.1145/3702336.3702340

CCS Concepts

• **Applied computing** → **Bioinformatics**; • **Computing methodologies** → *Neural networks.*

Keywords

Computer Vision, Animal-Computer Interaction, Interpretability

ACM Reference Format:
Devon Martin, Jeremy Park, Megan Carson, Margaret Gruen, Alper Bozkurt, and David L. Roberts. 2024. Automated Depth Sensing-Based Computer Vision for Dog Tail Wagging Interpretation. In *The International Conference on Animal-Computer Interaction (ACI 2024), December 02–05, 2024, Glasgow, United Kingdom.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3702336.3702340

1 Introduction

Scientific advancements in animal cognition have been fueling the long-term aspiration of interspecies communication with particular focus on apes, parrots, dolphins, horses, rabbits, cats, and dogs [13]. Dogs are one of the most common animals in human environments, and they serve critical roles including companionship, emotional support, drug detection, guiding people with visual disabilities, medical alerts, herding, *etc.* Despite progress with this popular domesticated species, deciphering the nuances of communication, especially to assess their emotions and mood, remains a challenge due to intricacies and complexity of their social cues and expressions.

It has long been discussed that tail wag is one of the behavioral signals used by dogs for visual and tactile communication and as a display of emotional states [12]. Dogs regularly use tail orientation, in conjunction with body posture, movement, and facial expressions, to display emotion, intention, and motivation. The movement dynamics and many degrees of freedom of a wagging tail enables dogs' tails to be a predominant display of such information.

Computer vision, when combined with deep learning, provides a powerful technique to record, analyze, and interpret animal behavior. Semantic segmentation is a computer vision technique for extracting meaningful sections of complex image data. It can be

used for behavior tracking by identifying certain anatomical features corresponding to behavior. This has only recently been possible thanks in large part to Convolutional Neural Networks (CNN). CNNs have proven great at extracting features and developing feature maps from image data. Stacked CNNs are capable of more abstract feature detection, improving performance. One such CNN architecture, Mask Region-based Convolutional Neural Network (Mask R-CNN), has been developed to generate pixel masks corresponding to objects in the frame [9]. These masks enable the necessary infrastructure to provide high accuracy and timely frame-by-frame behavior monitoring to assess stereotypical behavior and temperament. Such techniques have been used for pose estimation using depth cameras [11] and more generally for markerless detection of anatomical features [28]. For example, in previous attempts to gauge the emotional state of dogs, pictures of facial expressions were categorized as positive anticipation or negative frustration [3].

Mancini [18] argues that in attempts to create communication with animals and human-animal interactions broadly, the animal's comfort and psychological well-being are paramount. With this in mind, a good design solution is to use camera-based systems. These systems are strongly motivated by [21] and implemented in [19, 23]. Camera systems can provide continuous, high-resolution tail wag identification while only requiring that the dog remain within the observation area, thereby minimizing animal discomfort. Being mounted from above, the cameras permit easy mobility and, importantly, do not impede natural tail movement in any way. Additionally, camera-based systems could allow for detecting multiple dog and multiple human subjects simultaneously for socialization experiments, and could be combined with other approaches like attention and posture monitoring for more stimulating, pet-aware toys.

In this paper, we present our latest efforts for an enhanced computer vision system for measuring and interpreting dogs' tail wagging behavior. The system uses a depth camera for 3D pose estimation of the tail, and the tail response is measured in the presence of specific visual stimuli to better quantify positive or negative responses. This work extends the Mask R-CNN pipeline described by Roberts et al. [23] to a full end-to-end pipeline from video recordings to tail position and wag behavior analytics. Roberts et al. reported on preliminary efforts using a neural network model trained predominately on video from two dogs only and showed relatively poor ability to generalize to unseen dogs. The work presented here updates this pipeline in two important ways: (1) completing the pipeline with metrics to judge tail position and motion, (2) adding significantly more training data from about 50 dog subjects to improve R-CNN detection rates. We present results indicating that the newly trained model is capable of much broader recognition, and that the performance of recognition of the tail by this model is sufficient to derive meaningful tail position and wag analytics that can support future investigation of behavioral communication in domestic dogs.

We chose the problem of monitoring tail wag. As far as we know, there has never been objective monitoring of tail wag besides Roberts et al. [23], and while this earlier system did detect dog tails, it was not robust enough for full-length video data. Our updated system is the first successful application of tail motion detection that works well enough on video recordings. The main contribution of this paper is the development of a computer vision system capable of objective tail wag measuring on video data. We also validated our system to confirm its usefulness to the behavior community and performed a final preliminary comparison using the extracted data from thirty dogs. This paves the way for a stimulus-based behavior classification using tail motion.

2 Background

We discuss the two prominent tail wag theories, prior studies pertaining to dog tail wag, and common approaches formerly used with computer vision systems on animal farming and sciences.

2.1 Tail Wag Theories

Earlier work on interpreting tail wagging behavior has focused on two prominent theories for why dogs wag their tails and relates to their development alongside the domestication process: 1) the domestication syndrome hypothesis, and 2) the domesticated rhythmic wagging hypothesis.

As described in the Belyaev experiments, the domestication syndrome hypothesis posits that tail wag arose unexpectedly as a byproduct of selection for tameness and other human-useful traits. Interestingly, Russian geneticist Dmitri Belyaev was able to produce human-compatible traits in wild foxes in as little as six generations of selective breeding for tameness [6]. These new traits included the ability to be petted, whining when humans leave, and tail wagging when experimenters approached. The last of these already indicates a movement towards human-dog communication. These tamed foxes are also capable of following human gaze, suggesting intention or interest recognition. Alternatively, the rhythmic wagging hypothesis posits that tail wag arose during the domestication process as a direct response to humans' proclivity for rhythmic stimuli [26].

2.2 Prior Studies in Dog Tail Wag

Observational studies of dogs in the 1960s and 70s regularly discussed tail movement in a limited context, as more an addendum to larger displays of dog emotion rather than a display in and of itself. Fox [8] and Tembrock [29] both reference tail movement and position as displays of motivational states. In particular, higher tail positions are associated with confidence and/or aggression while lowered tail position can be a neutral signal or may reflect fear and/or submission.

More recent experiments aimed not only to properly interpret the signal from tail motion, but also to use these motions to display proper emotional states. Some of these experiments used robotic tails to display information with attempts to use simulated tail wag signals to provoke responses from other dogs [2, 15] as well as humans [27]. Leaver and Reimchen [15] looked at dogs' decision to approach made in response to a short or long tail that was either sedentary or wagging. They found that longer tails were better at displaying signals and that wagging tails were more approachable. Quaranta et al. [22] presented stimuli to dogs and monitored the tail wag left vs. right bias. Positive stimuli invoked right-biased wags while neural or unfamiliar stimuli showed either no bias or left-biased wags. They relate these to the left or right brain hemisphere dominance. Ruge et al. [24] proposed a much higher-fidelity

ethogram model to improve user experience. They showed a large dependency on dog temperament and noted that wag duration and intensity were also important for interpretation, not just instantaneous orientation. The group acknowledges a need for further nuances for the model, which has been one of our motivations behind this presented study.

2.3 Computer Vision in Animal Applications

Within farming and animal sciences, animal measurements can be time demanding, costly, and stressful for the animals [4, 7, 16]. The recent popularity in digital cameras and 3D cameras has made computer vision systems (CVS) an acceptable method as a non-invasive and lower-cost alternative. RGB and IR cameras are commonly used, though depth cameras using time-of-flight or LiDAR and hyperspectral cameras have also been applied.

Deep learning approaches using these cameras have now been useful for a variety of tasks, including object sensing, mapping, recognition, motion tracking, semantic/image segmentation, scene interpretation, monitoring, phenotyping, image classification, and pose estimation. CVSs have been applied to primates, dogs, horses, cats, cattle, pigs, ovine, and poultry [4, 7, 16]. Many convolutional neural network architectures have been created for these tasks:

- Animal tracking: DeepLabCut, EZtrack, DeepPoseKit
- Image classification: AlexNet, Inception, MobileNet, DenseNet, ResNet, VGG, YOLO
- Object Detection: YOLO, R-CNN and Mask R-CNN, DenseNet, ResNet
- Semantic/Instance Segmentation: DeepLab, Mask R-CNN
- Pose Estimation: CPHR, DeepLabCut, DeepPose
- Custom training library: Microsoft's COCO

Important for this paper, a survey focusing on CVS for animal emotions was published recently [4]. Many of these studies focused on pain specifically, but other standard emotions were included, depending on the species. Depending on the study, either a discrete emotional state model or a dimensional approach model (classifying along different dimensions, often with valence) was used. These studies commonly employed facial motion tracking, such as custom Facial Action Coding System approaches that have been adapted for primates, dogs, horses, and cats. There have also been many studies on behavioral encoding in horses, dogs, and pigs, focusing on body expressions observed in posture [4]. Sometimes stimuli were controlled and administered to the test animals and in other studies, the animal was monitored over time and natural behaviors were grouped.

For our study, we performed instance segmentation using the Mask R-CNN approach, retraining it specifically for tail detection using a COCO library approach. The Mask R-CNN is optimal for this type of task because it performs object classification, detection and segmentation in parallel [16]. In addition to using traditional 2D image tools, we were able to acquire 3D data using a depth camera, making our approach unique.

3 Methodology

We detail the experiment with dogs and how the video data was acquired, then discuss ethics, followed by major updates to the prior work to make the system more robust, how image processing was conducted, and lastly discuss how final metrics were calculated.

3.1 Experimental Setup

This work augmented a previously presented data pipeline [23], following guidance from [21]. To briefly summarize the process, a flat space of about four square meters was monitored overhead with a RealSense D415 depth camera [5] (Intel, Portland, OR), which provided color, infrared, and depth images at 60 frames per second (FPS) (see Figure 2). This setup was also easily ignored by the dogs. Owners sat with their dogs leashed and visual stimuli were presented for periods of 15 seconds each (see Figures 1 and 3). Each session lasted about 6 minutes total and there were 42 participants.

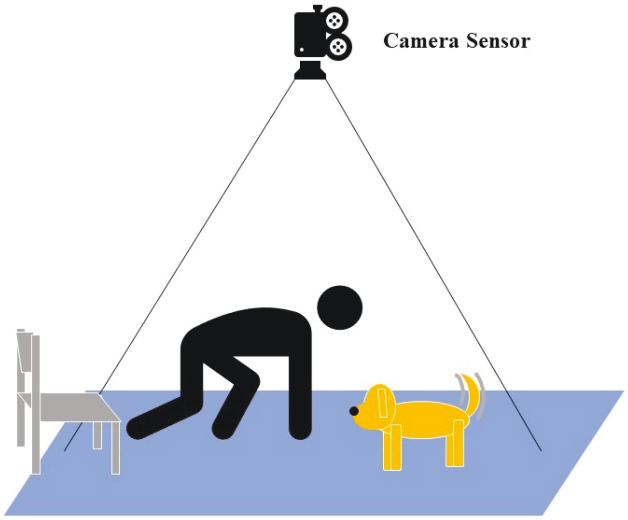

Figure 1: Computer vision setup for dog tail monitoring

The RGB and depth video streams were recorded in standard VGA resolution (640 × 480) at 60 FPS. The higher frame rate enabled more accurate measurement of tail position since the spatial displacement of the tail in 16ms is far lower than the standard 33ms encountered with a more common 30 FPS recording rate. The RealSense Viewer software recorded video streams in a format called ROS bag[1] that allowed the pixel-synced depth and RGB images to be replayed after recording as if they were being streamed from the camera itself. The primary benefit was the ability to replay the recordings with accurate timing for post-processing, and more importantly, to ensure the depth and RGB images remained in sync for the tail position analysis. The depth video provided pixel-linked RGB images and depth maps, meaning a pixel in the RGB image was correlated directly with a pixel in the depth map, allowing ascertainment of 3D information. A second RGB camera was used to monitor the visual stimulus being presented to the dog. The experimenter performed hand gestures within the field of view of both cameras for calibration.

A Maria database stored the data-heavy images in each frame from the RealSense camera (RGB, depth, and IR). Additionally, the

[1] http://wiki.ros.org/Bags

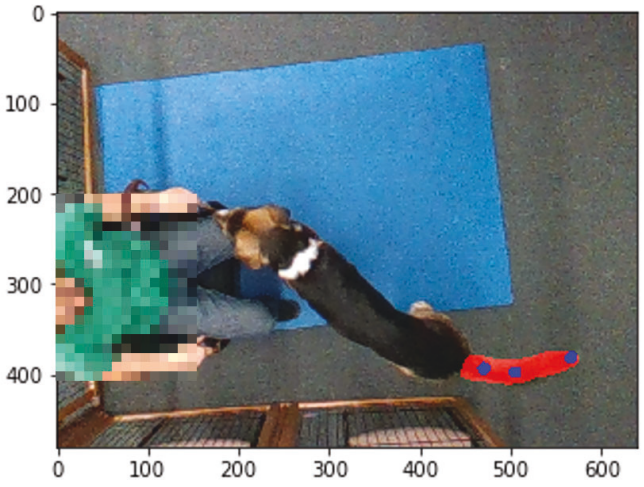

Figure 2: the top-view scene that is used by the computer vision algorithm for automated detection of the tail wag

database stored the final results- centroids and medial axis points with corresponding depths. This way, storage was optimized and future access to the results were easily obtainable for all later steps.

The outline below describes the image processing pipeline. The **boldfaced** items highlight the particular contribution of this paper. These changes are different from previous implementations [19, 23], and provide additional robustness.

(1) Acquire color and depth frames
   - Queue image frames from stored .bag file
   - Store image frames into pre-built Maria DataBase for easy access
(2) Perform image segmentation to identify the dog tail within the image frame
   - Use R-CNN to acquire all masks for base, end, entire tail, and foot
   - **Determine primary masks for main tail sections**
   - Store masks in Maria DataBase
(3) Derive tail the medial axis and main centroids
   - Identify dog tail skeletonization
   - **Use centroids to determine tail start and end points**
   - **Logically order skeleton points**
(4) **Derive standardized points for tail position**
   - **Establish orientation of tail position reference for each frame**
   - **Preprocess with interpolation and standardization**
   - **Calculate tail position, direction, angle, and tip score**
   - **Store scores in Maria DataBase for quick future reference**
(5) **Calculate time-specific metrics**
   - **Limit per Nyquist rate**
   - **Finite difference for speed and acceleration**
   - **Estimate angular momentum with speed and tail**

Figure 3: Presented visual stimuli during the computer vision tests included a) a large ball, b) a yellow box, c) a trash bag, d) nothing, e) an English Sheepdog stuffed dog, f) a Jack Russell stuffed dog, g) a Sheltie stuffed dog, and h) a real dog

3.2 Ethics

All procedures involving dog participants were conducted under an IACUC approved protocol (Protocol #18-053-O) and under the direction of a board certified veterinary behaviorist. The 30 dogs included in this work were recruited as volunteers via flyers and email listservs within the NC State College of Veterinary Medicine (CVM) students, staff, and faculty. Dogs visited the CVM for a single visit, and owners signed an informed consent prior to the start of the study [23]. Experiment volunteers could withdraw from the study at any time, and staff had the option of prematurely ending the experiment in the event of observed stress or discomfort.

3.3 Major Updates

We performed an early run-through of the dog video data. This exercise elucidated two major difficulties with a naive approach. 1- The original R-CNN was insufficient to robustly identify dog tail segments in images, and 2- the streaming method used to extract

individual images was not sufficient to attain each frame. Each of these challenges and their solutions are presented next.

3.3.1 R-CNN Retraining. Previous training for the R-CNN was discussed in [23]. To summarize the original model, a generic Mask R-CNN was implemented based on the open-source repository by Matterport [1]. To specialize the base model towards tail recognition, approximately 300 images were annotated from the original bag files from two dogs specifically. Performance on the two dogs used for training was decent but could not be generalized to the other dog images well. While this was a good starting model, we found inefficiencies with video detection of various dogs throughout our videos. We define this inefficiency in terms of useful frames, where a useful frame is one in which enough tail segments were successfully detected in order to detect the full dog tail. The base R-CNN run-through produced a very low percentage of useful frames per video (see Figure 4). Commonly, dog video files produced less than 10% useful frames, suggesting either long stretches of false negatives or very choppy tail identification, not good enough to detect the wag signal. The average percentage of useful frames was 20.97%.

This performance was not acceptable for our analysis, so we recruited student volunteers to annotate additional images using LabelBox (https://labelbox.com/). We had about 50 images for each dog in the study, producing an additional dataset size of approximately 2500 annotated images. This was over eight times the original dataset and was expected to greatly improve semantic segmentation performance. However, we noticed issues even when using the gold standard of human annotations. These included highly fluffy tails, very small dogs which are naturally harder to detect, and darker colored dogs whose tails could not be distinguished amongst regular fur. Being difficult for human detection, we cannot expect the computer vision model to perform any better, so this helps set an upper bound on expected performance. Using these additional annotations, we retrained the Mask R-CNN.

While the same dogs were used for both training and validation, the images from the dogs were split randomly. This means that despite of the many breeds and sizes that our model was trained on, our model may still suffer low generalizability. However, for purposes of tail detection in our video data, this model was sufficient.

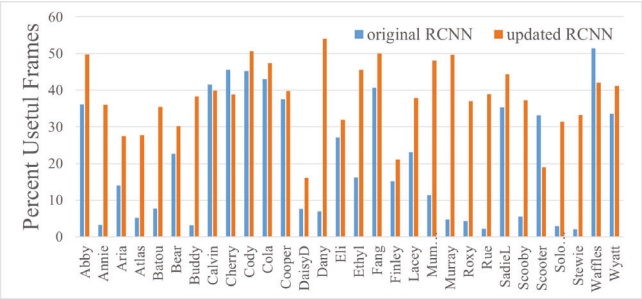

Figure 4: Semantic segmentation rate of tail detection

Training curves are shown in Figure 5. We observed that generally, losses decreased with epoch as expected. We selected the 157th epoch as our model as this best balanced the various validation losses. After updating the R-CNN with additional images, the average percentage of useful frames almost doubled to 38.03%. While still not great, most of dog videos had at least 30% useful frames. This score accounts for null time before and after the experiments, so the experimental detections would be higher. As we will see in subsequent analysis, even at 30% detections, we can detect the wag signal consistently.

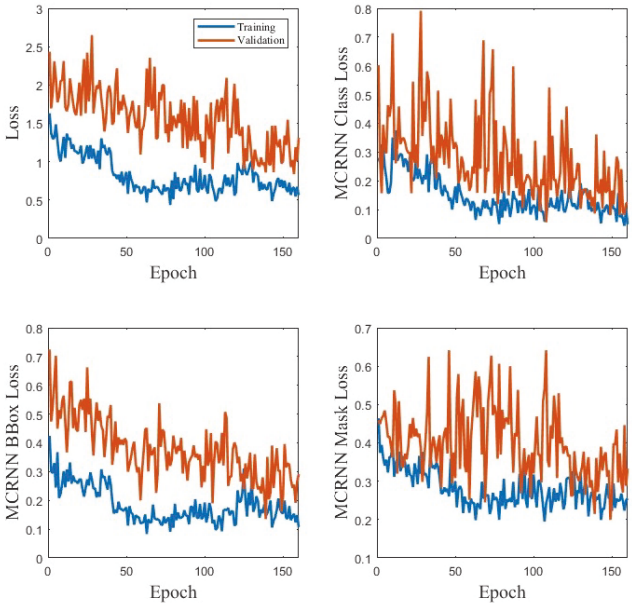

Figure 5: R-CNN retraining curves

Table 1: Updated R-CNN mAP Scores

Metric	IoU	area	maxDetections	Score
Average Precision	0.50:0.95	all	100	0.222
Average Precision	0.50	all	100	0.472
Average Precision	0.75	all	100	0.184
Average Precision	0.50:0.95	small	100	0.248
Average Precision	0.50:0.95	medium	100	0.235
Average Recall	0.50:0.95	all	1	0.251
Average Recall	0.50:0.95	all	10	0.305
Average Recall	0.50:0.95	all	100	0.305
Average Recall	0.50:0.95	small	100	0.270
Average Recall	0.50:0.95	medium	100	0.335

We used the COCO library's evaluation function with 500 validation images to report on mean average precision (mAP) and report these scores in Table 1. This showed expected scores around 0.2-0.3. The additional training for the R-CNN improved its tail detection rate (reducing type 2 error), but now detects tails where none are present (increased type 1 error). Figure 6 shows the number of detections of each type for each frame. Most frames have 0 or 1 detection as expected, but it is not uncommon for 2 or more detections, with fewer instances of higher numbers of detections.

Multiple detections are primarily caused by experimenters walking into the frame during setup and post-experiment, though we have also observed instances of the dog's leg being identified as a tail. To resolve this ambiguity, we selected a primary detection for each class by examining nearby tail sections and/or past detections. For example, if two end-of-tails are detected, and one is closer to a prior known location of an end-of-tail, then this one is primary. Alternatively, if two end-of-tails are detected and one is much closer to an entire-tail detection, then this one is primary. In cases of multiple detections, we utilized the following equation, with t being number of frames before the current frame, and r being pixel-wise distance between different tail subsection centroids:

$$s = \max\left(0,\ 1 - \frac{20 \cdot t + r}{600}\right)$$

The primary detection is the one with the higher score. We engineered this equation specifically for the following important properties:

(1) balancing a dependence between time and distance
(2) higher scores when closer in distance and time
(3) scores 0 when...
  (a) near-full frame-length away
  (b) frames more than 30 frames (1/2 sec) before
(4) simple function
  (a) piecewise linear
  (b) on the unit interval $s \in [0, 1]$

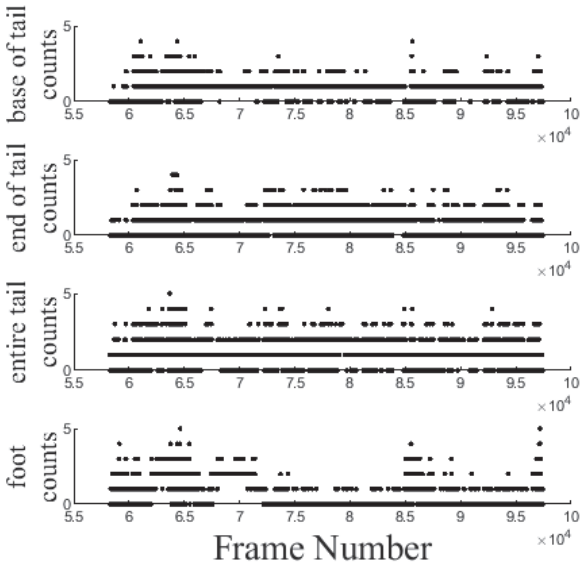

Figure 6: Counts of tail sections through the frames for one of the dogs.

3.3.2 Streaming. Because of the nature of RealSense bag files, data are streamed into a queue. However, the streaming process is not consistent enough to extract each sequential frame, skipping some and double counting others. The frame skip rates are shown in Figure 7, where about 47% of frames are sequential while 99% are within 6 frames from each other. Because of inconsistent stream speeds when multiple calls to a bag file are made, we implemented a 20% overlap between multiprocess calls to prevent large gaps in time.

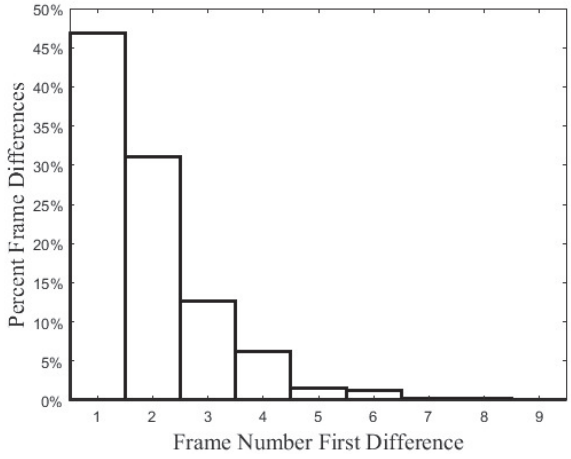

Figure 7: Frame number differences show how often frames are skipped from the streaming file. About 46.9% of frames are consecutive, while over 99.6% of frames are within the Nyquist rate. The Nyquist rate is a standard tool used in signal processing to ensure a signal capture a periodic phenomenon [20].

During experimentation, in addition to the specialized depth camera, we had a standard camera behind the experiment space, where the dog and stimulus being presented could be observed. Ideally, the bag file video and rear video should be temporally aligned, differing by no more than a constant, so that stimuli times can be aligned with tail wag times. Yet, the stream inconsistencies introduced non-constant time distortion. Fortunately, the bag files contain frame numbers as metadata and the video made from the bag frames was corrected by filling in empty times with the most recent last frame.

3.4 Processing

Once the two major problems were corrected, we proceeded with a more standard workflow highlighted in Figure 8. We used the R-CNN to extract the masks, use skeletonization to collect the medial axis points, and standardized the points before feature extraction. The following procedure was originally presented and is summarized from our earlier work [19]. However, it is greatly expanded upon for the final camera-based system presented in this paper.

3.4.1 Masks. The R-CNN returned a list of masks that corresponded to the sections of the image with either the base of the tail, the end of the tail, the entire tail, or a human foot.

We began with the prior R-CNN model [23] and analyzed the number of successfully processed frames. The main obstruction towards broader applicability was the accuracy of the Mask R-CNN [9]. Table 2 shows that the pipeline could successfully identify

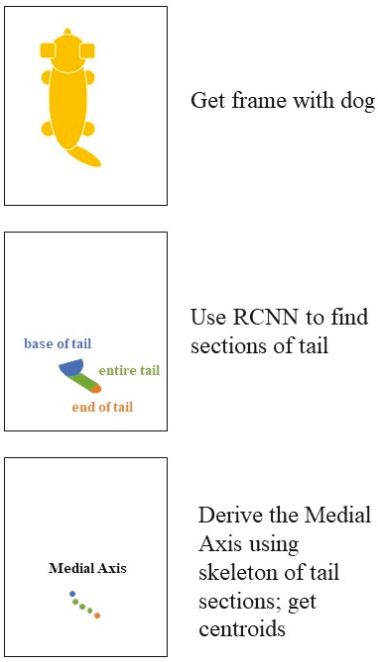

Figure 8: Data pipeline flowchart

the dog tail approximately 2/3 of the time. In particular, the base of the tail was easily identifiable with a detection rate of about 64% of the frames.

Table 2: Table of R-CNN results

Metric	Value
Percent tail detection	67.9%
Percent base of tail detection	64.1%
Percent end of tail Detection	38.5%

3.4.2 Medial Axis. Starting with a set from the medial axis, we order the skeleton points to orient the virtual tail. Our Mask R-CNN identifies the base, the body, and the end of the tail [23]. Ordering is accomplished using a piecewise optimal pathing approach. From one point, we connect it to the nearest point, and continue this process until all points are connected. For initialization, the centroid of either the base or the end of the tail is used as either the start or the end point, respectively [19].

After R-CNN detection and tail skeletonization, only the medial axis keypoints and centroids of the base-of-tail and end-of-tail were stored for further processing. This saved processing and storage time for all further steps without loss of needed information.

Based on which sections of the tail were detected in each frame, either a Base-Tail approach or End-Tail approach was used. The Base-Tail places the centroid of the base as the first point and sequentially finds the closest points along the tail. The End-Tail starts at the centroid of the end of the tail, then sets the furthest tail point from this centroid as the first point, then sequentially finds the closest points moving along the tail. Which method was used in which case depended on what subsets of the tail were detected as outlined in Table 3.

Table 3: Method used for determining orientation based on segments of the tail detected by the Mask R-CNN model. For *, if base detected within 30 frames ago: Base-Tail, else if end detected within 30 frames ago: End-Tail, else skip.

Base	End	Tail	Method
1	1	1	Base-Tail
1	1	0	skip
1	0	1	Base-Tail
0	1	1	End-Tail
1	0	0	skip
0	1	0	skip
0	0	1	*
0	0	0	skip

3.4.3 Standardization. After all medial axis keypoints were sequenced and before valid metrics were extracted, further preprocessing steps were necessary to make a proper comparison. These additional steps were:

- interpolation
- invert depth
- rotation

We wish to be able to compare medial axis points pertaining to the same small intervals along the tail. For this reason, we convert these points to percentiles along the tail. Another problem arises as the depth (z-axis) is recorded in different units than the pixels (x and y axes). For this reason, we kept overhead positions separate from depth positions where possible. For interpolation, points were selected along the length of the tail determined by pixel distance alone. Interpolation to a standard twenty points (shown in Figure 9c) allowed for alignment of points between frames.

Depth inversion was a simple operation to convert depth data to height data. Since scores are relative to the distance along the tail, we maintained heights as negative numbers.

Lastly, we translated and rotated the points such that the first point was at the origin and the first direction for the tail was facing directly $-\vec{y}$. The dot product equation was used to estimate a θ. A rotation was then completed with quaternions (see Figure 9d).

3.5 Deriving Features

We calculated four directional scores based strongly on the ethogram by [24]. These were position (up vs down), direction (left vs right), angle (degree from midline), and tip angle (greatest angle along tail end). The reference coordinate frame was designed to have the tail pointing directly backward, seen as a continued extension of the spine [19]. This aligned the y-axis along the dog's spine and the z-axis opposite gravity (see Figure 10). Scores were then calculated and normalized by the length of the tail, thereby returning values $x \in [-1, 1]$. For demonstration, Figure 11 shows several instances of tail positions with their corresponding scores.

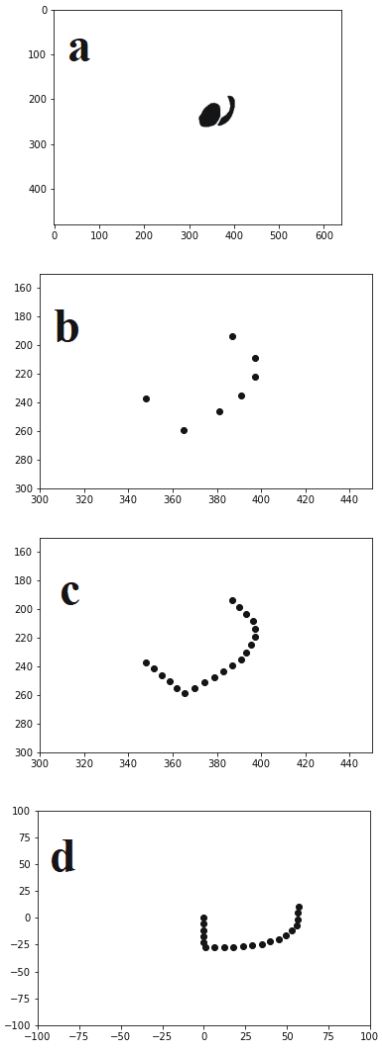

Figure 9: Medial Axis Preprocessing Steps (a) R-CNN identifies a base of tail and entire tail subsections (b) centroids and skeletonization provide medial axis keypoints (c) interpolation is used to standardize the points relative to position along the tail (d) translation and quaternions are used to rotate the medial axis to a standard format

As an example, the equation for direction (which varies along the x direction) is:

$$\text{direction score} = \sum_{i=1}^{20} \frac{||s_{x,i} - s_{x,i-1}||}{||s_i - s_{i-1}||}$$

Here, i indexes the 20 interpolated subsections along the length of the tail. Numerator terms, $s_{x,i}$, indicates the x-value (direction score) of tail segment i, while denominator terms, s_i, are the full 3-vector values of tail segment i. Tail tip angle was calculated by calculating the maximum angle between sequential tail subsections for the last few identified medial axis points.

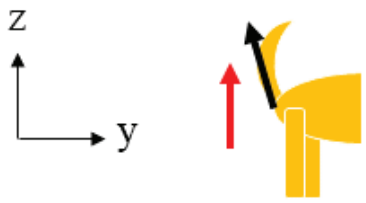

Figure 10: Tail Orientation

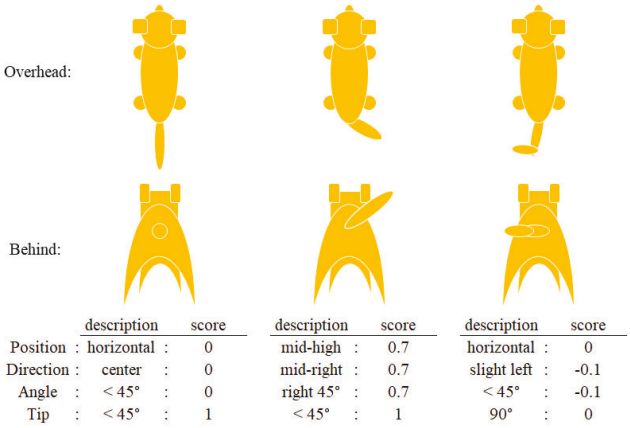

Figure 11: Dog tail positions and accompanying scores

Tail speed, angular momentum, and acceleration are all time-dependent metrics and are heavily affected by 1) frame skips and 2) proper interpolation for correct medial axis point comparison between frames. From prior experience, we estimated a maximum reasonable tail wag frequency of 5 Hz, which required a sampling frequency of 10 Hz as imposed by the Nyquist rate. With a camera rate of 60 fps, this means that at most five consecutive frames can be lost/skipped without loss of signal. Generally, speed and acceleration are estimated using finite difference methods. However, because of inconsistent frame detection rates, these equations were slightly altered from the traditional backward difference method. With s_1 and s_2 being a set of medial axis points in frames 1 and 2, i indexing tail axis points, and n being frame number, average speed is calculated as:

$$\text{avg speed} = \frac{\sum_{i=1}^{20} ||s_{2,i} - s_{1,i-1}||}{(n_2 - n_1) \cdot 20}, \quad n_2 - n_1 \leq 6$$

Acceleration is calculated similarly.

We are also interested in the tail's angular momentum (L) for motion analysis. Though we are interested in general motion of the tail for display purposes, angular momentum has been applied previously to understanding climbing motion in lizards and geckos [10, 25]. Angular momentum is the product of inertia and angular velocity. Inertia itself is the product of squared distance from a pivot and mass. With the base of the tail as the pivot (see Figure 12) and assuming that the tail has constant mass distribution, ρ, inertia is:

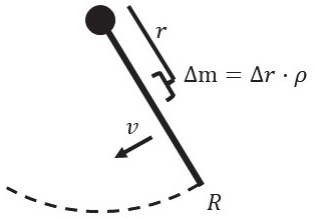

Figure 12: Tail momentum derivation

$$I = \int_0^R \rho \cdot r^2 dr = \frac{\rho}{3} \cdot R^3$$

and angular momentum simplified to the expression:

$$L = \frac{\rho}{3} \cdot R^3 \cdot v$$

Since ρ varies from dog to dog, we report L/ρ as a metric.

We report tail position, direction, angle, and tip angle at each frame as well as tail velocity, acceleration, and angular momentum.

4 Results

An example of metrics for one dog are shown in Figure 13. The top graph shows the visual stimuli times. These were presented for periods of 15 seconds, which corresponds to 1000 frames, which is about the length of these sections. We observed that near the beginning of the experiment (during setup), there are few metrics since the dog was offscreen during most of this time. We also observed a lull in position, direction, and speed a bit before the 70,000th frame, corresponding to a time when the dog was resting its tail. Lastly, we observed times of sporadic higher-variance position, angle, and speed scores, suggesting the tail was more active during these times. There are some momentary changes such as around frame 75,000, where we see a short-left bias in direction and tail angle turning slightly positive (forward-facing). There are also noteworthy periods of rapid direction change (strong wag) from frame 80,000 to 84,000. Times of high angular momentum, such as at frame 85,000, correspond to when the tail was commencing or terminating a tail wag.

We simplified the stimulus response by averaging metric data (either position, direction, angle, tip angle, speed, angular momentum, or acceleration) to times roughly aligned with the time the visual stimulus was presented. We elected to offset the response window by 30% to allow a delay for the dog to see and properly respond to the stimulus and residual effects after the curtain was closed. For example, if a stimulus was presented from second 100 to 115, the score would be averaged from 104.5 to 119.5.

4.1 Computer Vision Quality

In addition to collecting video data, we had previously annotated the number of wags per dog in response to each stimulus. With this ground truth data, we examined the quality of the R-CNN signals. Figure 14 shows direction score sequence examples from our method at times when visual stimuli were shown. Sections with high wag counts show obvious cyclical patterns as expected of tail

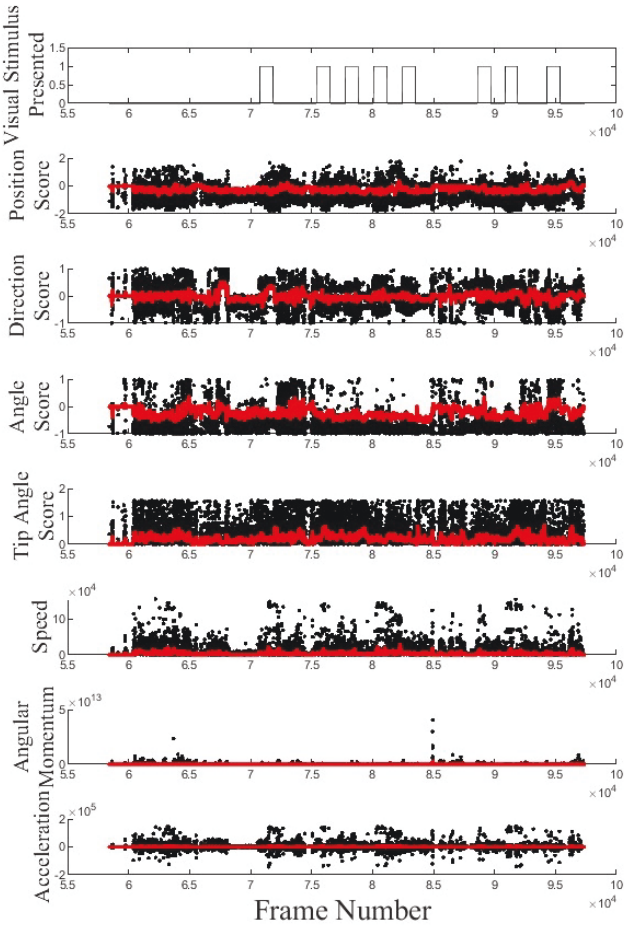

Figure 13: All metrics calculated throughout the experiment for one representative dog in our data set. Black is raw data; red is mean averaged trend. Visual stimulus is 1 when stimulus is being shown; order of stimuli is black ball, white gray fake dog, yellow box, small brown white fake dog, trash bag, white brown fluffy fake dog, nothing, real dog.

wagging. This suggests we are detecting the wag behavior that we claim.

Notably, only certain subsections of stimulus times showed expected tail wag periodicity. Because of Heisenberg's uncertainty principle, we cannot simultaneously perfectly know the time and frequency of a signal, but wavelet transforms are good at balancing temporal and frequency information to roughly estimated periods of signal [17]. To filter out the effects from non-wagging times, we used the wavelet transform with a standard Ricker wavelet, and observed good amplitudes in the pseudocolor plot at ordinate position 10 (Figure 15b). We then took the amplitude of this section and smoothed it with a moving average filter (15c). Signals consistently above the threshold of 0.25 were considered to be wag times. Once these times were detected, the following sinusoidal metrics were calculated:

- Offset: average of the 5th to 95th percentile of the sequence

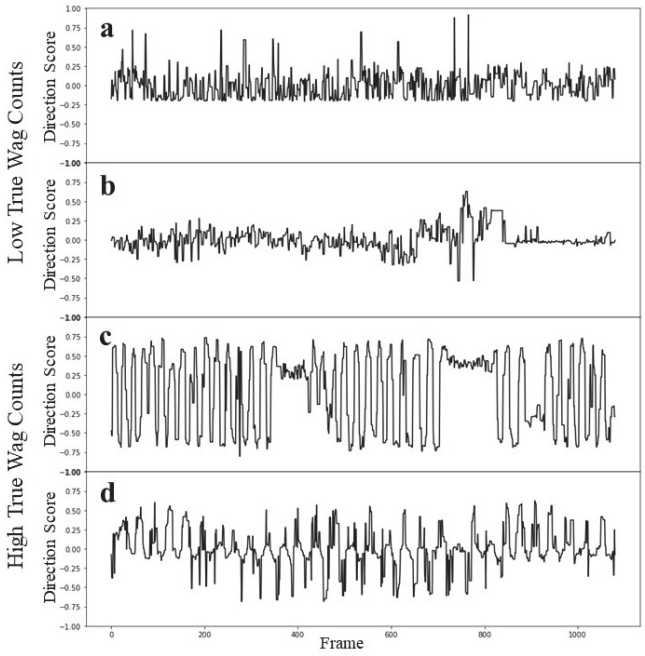

Figure 14: Subplots a and b show R-CNN-derived patterns with very low true wag values. Plots c and d show R-CNN-derived patterns with high true wag values. We can observe very clear cyclical patterns in c and d that are not observable in a and b.

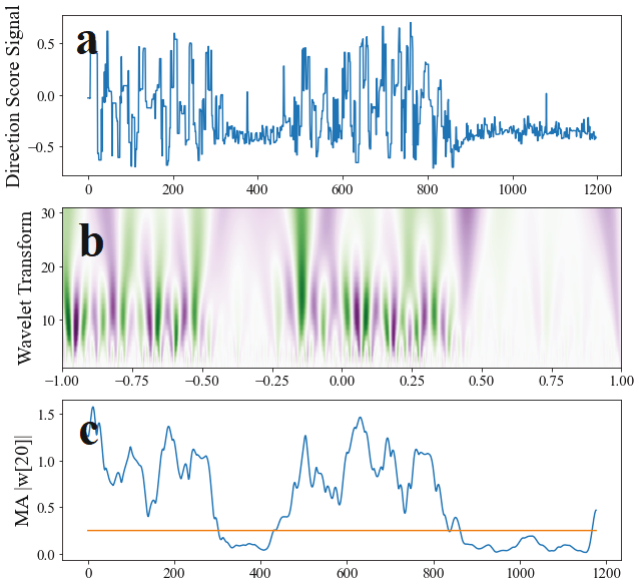

Figure 15: We identified more detailed wag times using a wavelet transform with the Ricker wavelet. a) the original signal, b) wavelet transform of (a), c) moving average of the absolute value of the wavelet transform where ordinate of (b) is 10. In this example, the wag times returned correspond to frames $[0, 303] \cup [431, 863]$.

- Amplitude: bisection of the 5th to 95th percentile of the sequence
- Peak Frequency: max frequency from the FFT between 1 and 4 Hz

Lastly, subsections were combined using a weighted average in proportion to the length of the subsection. This then returned an overall offset, amplitude, and frequency score for each stimulus time.

5 Discussion

To demonstrate that our system does work and captures the tail wag signal, we show a short section of one of the derived metrics, direction score, in Figure 16. It should be noted that direction corresponds to left and right movement relative to the spine from an overhead camera. We observed a high degree of correspondence between the direction score and the tail motion in the video. The dog was wagging its tail in sections A and C, stopped wagging in section B, and sat in section D. We observed that the direction score showed very low variance in sections B and D as can be expected. Although sections A and C appear to have high variance due to noise, they actually align with tail motion during a tail wag pulse, with a frequency around 2 Hz and amplitude from -0.75 to 0.75 times the length of the tail.

By zooming into a section of the tail wag, we can observe more clearly the sinusoidal pattern in the data. Figure 17 shows a cyclic pattern with a frequency of 1.8 ± 0.1 Hz. This proves that the video system is capable of capturing the tail wag signal. As a future direction, since wag signals are frequency signals that are determined temporally, wavelet analysis would be a good analytical direction. Wavelet spectrograms could immediately reveal wag behavior localized in time and frequency.

Figures 18 and 19 show each of our seven metrics' distributions of the dogs to each stimulus presented. We can observe that most stimuli show similar population distributions, except for "real dog" response, which has a slightly lower angle and tip angle score, suggesting that the tail was more backward during these times. The standard positions were around -0.7 for position, which is highly down relative to the base of the tail, 0 for direction, which indicates no left/right bias, as expected, and angle was -0.4, which is moderately behind the dog, also as expected. These plots show similar values for most stimuli. This is largely because these are raw data aggregated across all the dogs tested. We intend on performing an in-depth analysis on each subgroup and stimuli as a future work. Regardless, these plots do show the applicability of a high-level initial analysis using the data acquired from our system.

When we focused on the smaller time subsets corresponding to wagging, these signals became obvious to a human observer. We performed a verification test by counting wags from our system and comparing it to the true counts. Sequences were mean-centered and stepwise interpolated to account for NANs. Then the following three features were extracted:

- Frequency: derived from the FFT algorithm, we recorded the peak frequency from the power spectrum. This is helpful for estimating the number of wags.

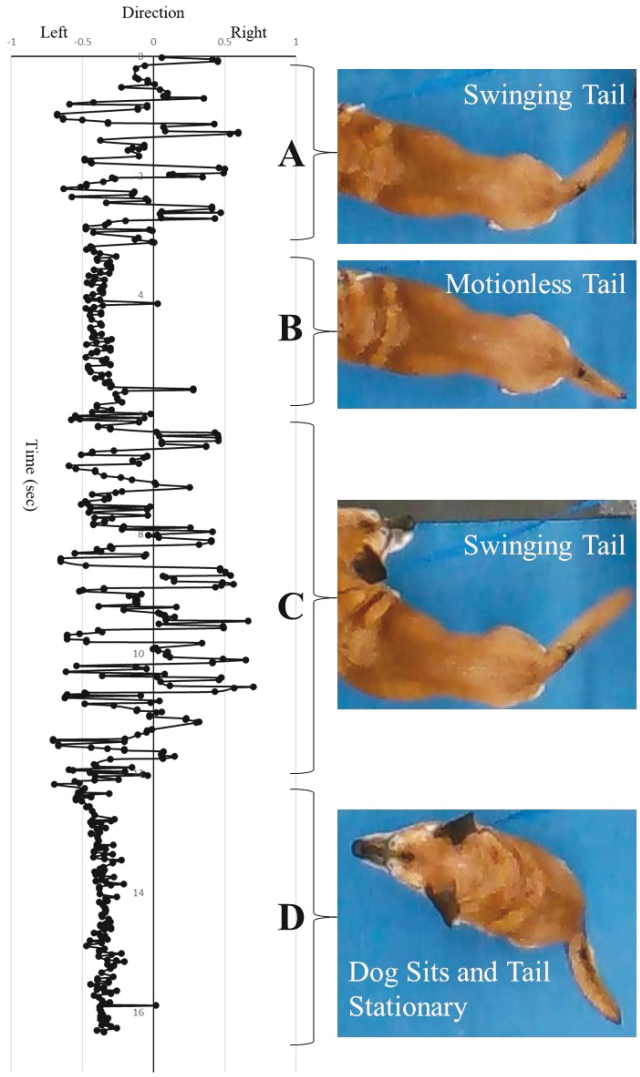

Figure 16: Detailed comparison of derived direction score and corresponding video frames. This 16-sec clip shows segments A and C to be when tail was wagging prominently, segment B during which the dog stopped wagging its tail, and segment D during which the dog sat down with the tail leaning to the left as expected of a negative direction score. Time is displayed top to bottom.

- Magnitude: this is the average of the absolute values of the first difference of the bandpassed signal.
- Kurtosis: calculated from the histogram of the bandpass filtered signal. This helps determine how common wagging is.

Using these features, we then optimized a linear combination of these few features using a 70/30 train-test split. We minimized the sum of squared error and found a coefficient of determination (R^2) with the true wag count of 0.45 ± 0.08, a moderate fit.

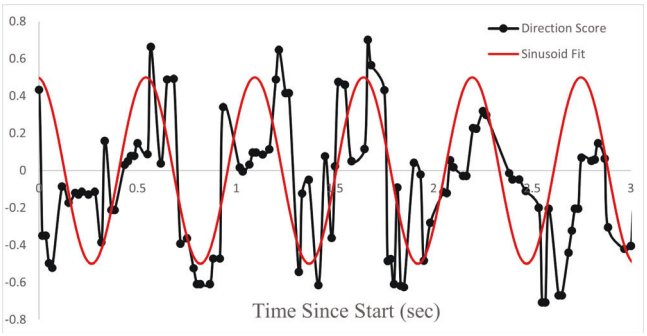

Figure 17: Fitting a simple sinusoid to a short segment of direction score elucidates a tail wag frequency of 1.8 ± 0.1 Hz.

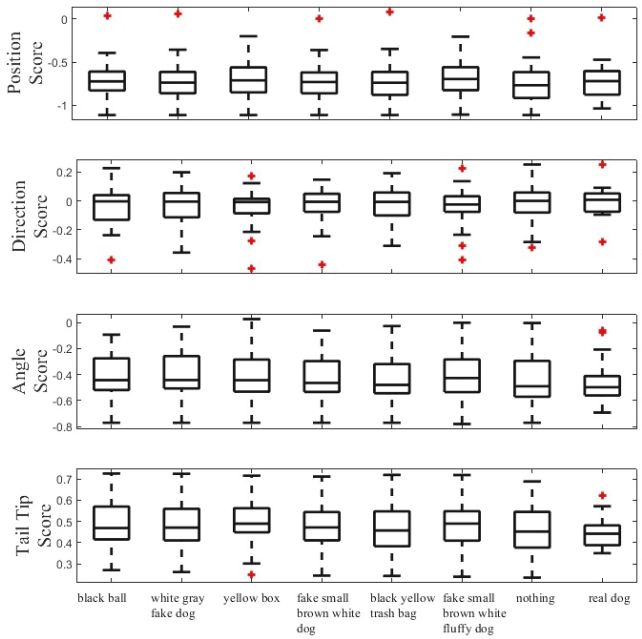

Figure 18: Distributions of the four position scores of all 30 dogs against each visual stimuli presented.

We next detail sources of uncertainty. The efficacy of computer vision systems can be evaluated along the lines of accuracy, ease of use, observer effects, and usability for paired studies.

As with all large system workflows, our camera-based system is susceptible to errors accrued from assumptions or randomness. Causes of deviations are similar to those already reported in our earlier work [19]. We summarize the causes of deviations that are still valid. 1- Being a neural network-based model, our RCNN naturally has object detection uncertainty (see Figure 4) [9]. If a tail section is not detected, the frame becomes less useful and robustness deteriorates. 2- Depth cameras specifically have random noise that is not shared with standard color images. This can alter values along the z-axis. 3- For our reference frame, the first few skeleton points of the medial axis presumably point directly along the y-axis. If

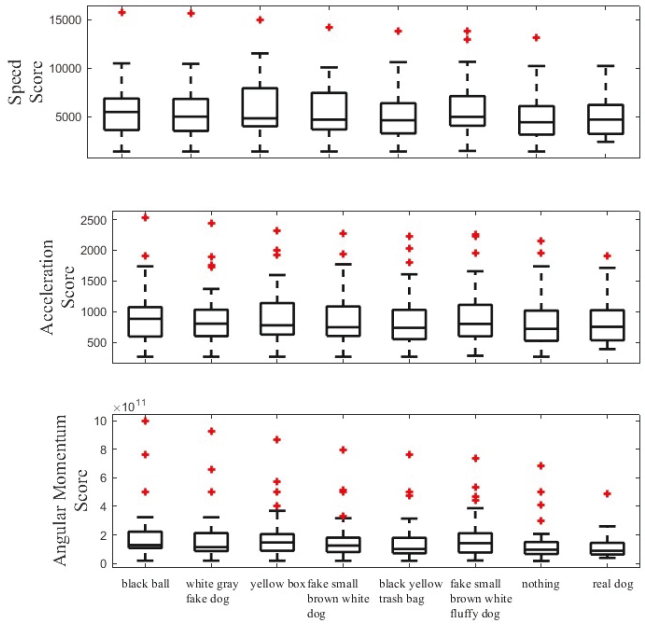

Figure 19: Distributions of the motion scores of all 30 dogs against each visual stimuli presented.

invalid, this assumption would create a systematic error affecting all other medial axis point estimates [19]. 4- Similarly, on the opposite end of the tail, the tail tip angle should be only at the very end of the tail, but may differ based on the dog species or tail conditions. 5- Dog-specific tail conditions can greatly affect performance, such as tail fluffiness or small tails, which confuses medial axis detection.

6 Future Work

The following future directions include continuing efforts originally presented in our earlier publication [19]. Here, we expand upon these earlier plans with more details.

Despite the desire for improved accuracy and the current low detection rates for individual frames, our computer detection system has still been sufficient for tail wag monitoring. Robustness was generally improved and the system functioned sufficiently in spite of many missed detections. However, we could consider implementing further robust procedures such as interpolation when within the dog tail wag Nyquist rate as only one frame detection is required in six consecutive frames presently. This is usually within reason using the presented methodology since the detection rates were improved using more training data. Alternatively, we can consider detrended fluctuation analysis (DFA) since this method is robust to the signal gaps commonly seen in the tail wag signals. Using the final derived metrics, future clustering analyses can be used to quantify the number of tail patterns or corresponding emotions [19]. Linear mixed models would serve as a means of comparison by accounting for the effects of different dog breeds, personalities, and effects of the different stimuli. With i for individual dog, j the stimulus, and k the feature, a model could be:

$$F_{k,ij} = \beta_0 + \beta_1 \times dog_i + \beta_2 \times stimulus_j$$

Using a computer vision approach, our tail detection methodology could be improved to a larger general ethology method using camera-based systems for posture and attention monitoring [21]. It could also be seen as a future open tool similar to DeepLabCut, which currently supports pose estimation [14]. As a more immediate improvement, a logical next step would be to identify owners' interactions with their dogs to monitor the human-dog dyad and more in-depth communication cues. Attached to the front-end of a feedback system, we can foresee this system being used for automated training. Vision-based tail wag tracking is a very useful research tool for understanding dog behavior better and with future application areas such as dog shelters, where dogs are already confined in spaces.

Overlapping tails, *e.g.* a curly-tailed or sickle-tailed dog, was originally a concern but was very uncommon, so we chose to ignore it at this time. This could be rectified in the future using the depth filter values along the cross section. If the depth of the center of the cross is more similar to one path than the other, it would suggest that it belongs to the former.

Even with all these improvements presented here, we anticipate variability and covariates amongst many well-established characteristics of the population of dogs, such as effects of dog breed, personality, left vs right preference, as well as tail traits, like fluffiness, curl, or simply the absence of a tail. Our post-analysis models specifically focused on the need to account for these population dynamics to best set the tail wag model baseline. It should be noted that the primary objective of this work is the completion of a tail wag monitoring toolkit. We did not attempt to create an interpretability framework because we anticipate that interpretations will change with future studies. All interpretations made in this paper were meant as examples of what could be claimed given the data from this system. In addition to controlling population dynamics, we spent extensive time analyzing the validity of the signals to understand the noise versus signal relationship and the limitations of the system.

Lastly, our updated system was designed to permit future studies on human-dog interaction. As such, the system, with further modifications, could be fully capable of identifying multiple dog tails in frame as well as a human in frame, whether the person be a trainer presenting commands, owner providing encouragement, or stranger causing confusion. While multiple tails are expected to be a minor update, human detection will be more difficult since we have not provided such training data yet.

7 Conclusion

This paper significantly expands the R-CNN methodology presented before [23] to work on continuous video data for dog tail wag tracking. We updated the original R-CNN with several thousand additional training images to make it more robust, at the cost of increased false positives. We then modified the data pipeline by taking advantage of both the temporal nature of the video as well as spatial locations of various detected class sections. We have shown that the method does work well enough on most of our videos to capture the tail wag signal in spite of streaming and detection difficulties. We observed a good correspondence between these metrics and the video footage. We then applied our system to 30

dog videos and extracted tail position in all three dimensions and derived temporal metrics like speed and momentum. The successful attempts to perform broad population statistical tests revealed some visual stimuli to be identifiable from tail wag alone.

Our system successfully detects tail position with a high resolution and without interfering with dog movement or natural tail motion. This study supports the literature and sets a standard for high-resolution monitoring extracted from tail motion. We expect that the higher resolution proposed with the presented methodology will inspire and help elucidate future studies of dog ethology, such as, in more dog-centric training or play environments.

Acknowledgments

The authors acknowledge the support from NSF through CCSS-1554367, EF-2319389 and ECC-1160483 and IBM Faculty Awards.

References

[1] Waleed Abdulla. 2017. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. (2017). https://doi.org/matterport/Mask_RCNN

[2] K. Artelle, L. Dumoulin, and T. Reimchen. 2010. Behavioural responses of dogs to asymmetrical tail wagging of a robotic dog replica. *Laterality* 0, 0 (2010), 1–7. https://doi.org/10.1080/13576500903386700

[3] T. Boneh-Shitrit, M. Feighelstein, A. Bremhorst, S. Amir, T. Distelfeld, Y. Dassa, S. Yaroshetsky, S. Riemer, I. Shimshoni, D. S. Mills, and A. Zamansky. [n. d.]. Explainable automated recognition of emotional states from canine facial expressions: the case of positive anticipation and frustration. *Scientific reports* 12, 22611 ([n. d.]). https://doi.org/10.1038/s41598-022-27079-w

[4] Sofia Broomé, Marcelo Feighelstein, Anna Zamansky, Gabriel Carreira Lencioni, Pia Haubro Andersen, Francisca Pessanha, Marwa Mahmoud, Hedvig Kjellström, and Albert Ali Salah. 2023. Going Deeper than Tracking: A Survey of Computer-Vision Based Recognition of Animal Pain and Emotions. *International Journal of Computer Vision* 131, 2 (2023), 572–590. https://doi.org/10.1007/s11263-022-01716-3

[5] Intel Corporation. 2018. Intel Realsense D400 Series (DS5) Product Family. https://software.intel.com/sites/default/files/Intel_RealSense_Depth_Cam_D400_Series_Datasheet.pdf.

[6] Lee Dugatkin. 2018. The silver fox domestication experiment. *Evo Edu Outreach* 11, 1 (2018), 1–5. https://doi.org/10.1186/s12052-018-0090-x

[7] Arthur Francisco Araújo Fernandes, João Ricardo Rebouças Dórea, and Guilherme Jordão de Magalhães Rosa. 2020. Image Analysis and Computer Vision Applications in Animal Sciences: An Overview. *Frontiers in Veterinary Science* 7, 551269 (2020), 1–18. https://doi.org/10.3389/fvets.2020.551269

[8] M. Fox. 1969. The anatomy of aggression and its ritualization in Canidae: developmental and comparative study. *Behaviour* 35 (1969), 243–258.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2980–2988. https://doi.org/10.1109/ICCV.2017.322

[10] Ardian Jusufi, Daniel Goldman, Shai Revzen, and Robert Full. 2008. Active tails enhance arboreal acrobatics in geckos. *PNAS* 105, 11 (2008), 4215–4219. https://doi.org/10.1073/pnas.0711944105

[11] Sinead Kearney, Wenbin Li, Martin Parsons, Kwang I. Kim, and Darren Cosker. 2020. RGBD-Dog: Predicting Canine Pose from RGBD Sensors. https://proxying.lib.ncsu.edu/index.php/login?url=https://www.proquest.com/working-papers/rgbd-dog-predicting-canine-pose-sensors/docview/2391023019/se-2 Copyright - © 2020. This work is published under http://arxiv.org/licenses/nonexclusive-distrib/1.0/ (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2022-08-17.

[12] M. Kiley-Worthington. 1976. The Tail Movements of Ungulates, Canids and Felids with Particular Reference to Their Causation and Function as Displays. *Brill* 56, 1/2 (1976), 69–115. https://www.jstor.org/stable/4533714

[13] Don Kulick. 2017. Human-Animal Communication. *Annu. Rev. Anthropol.* 46 (2017), 357–78. https://www.jstor.org/stable/44866095

[14] Jessy Lauer, Mu Zhou, Shaokai Ye, William Menegas, Steffen Schneider, Tanmay Nath, Mohammed Mostafizur Rahman, Valentina Di Santo, Daniel Soberanes, Guoping Feng, Venkatesh N. Murthy, George Lauder, Catherine Dulac, Mackenzie Weygandt Mathis, and Alexander Mathis. 2022. Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nature Methods* 19, 4 (April 2022), 496–504. https://doi.org/10.1038/s41592-022-01443-0 Number: 4 Publisher: Nature Publishing Group.

[15] S. Leaver and T. Reimchen. 2008. Behavioural Responses of Canis familiaris to Different Tail Lengths of a Remotely-Controlled Life-Size Dog Replica. *Behaviour* 145, 3 (2008), 377–390.

[16] Guoming Li, Yanbo Huang, Zhiqian Chen, Jr. Gary D. Chesser, Joseph L. Purswell, John Linhoss, and Yang Zhao. 2021. Practices and Applications of Convolutional Neural Network-Based Computer Vision Systems in Animal Farming: A Review. *Sensors (Basel, Switzerland)* 21, 1492 (2021), 1–44. https://doi.org/10.3390/s21041492

[17] S G. Mallat. 1998. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego.

[18] Clara Mancini. 2011. Animal-computer interaction: a manifesto. *interactions* 18, 4 (2011), 69–73.

[19] Devon Martin, Timothy Holder, Colt Nichols, Jeremy Park, David Roberts, and Alper Bozkurt. 2022. Comparing Accelerometry and Depth Sensing-Based Computer Vision for Canine Tail Wagging Interpretation. *Proceedings of the Ninth International Conference on Animal-Computer Interaction* 20 (2022), 1–8. https://doi.org/10.1145/3565995.3566025

[20] Harry Nyquist. 1928. Certain topics in telegraph transmission theory. *Transactions of AIEE* 47 (1928), 617–644.

[21] Patricia Pons, Javier Jaen, and Alejandro Catala. 2015. Developing a depth-based tracking system for interactive playful environments with animals. *ACE '15: Proceedings of the 12th International Conference on Advances in Computer Entertainment Technology* 59 (2015), 1–8. https://doi.org/10.1145/2832932.2837007

[22] A. Quaranta, M. Siniscalchi, and G. Vallortigara. 2007. Asymmetric tail-wagging responses by dogs to different emotive stimuli. *Current Biology* 17, 6 (Mar. 2007), 199–201. https://doi.org/10.1016/j.cub.2007.02.008

[23] David L. Roberts, Jeremy Park, Anthony Pappas, Margaret Gruen, and Megan Carson. 2019. Automated Tail Position Tracking with Millimeter Accuracy using Depth Sensing and Mask R-CNN. Print. *ACI'19, Sixth International Conference on Animal-Computer Interaction* (Nov. 2019). https://doi.org/10.1145/3371049.3371050

[24] Luisa Ruge, Elizabeth Cox, Clara Mancini, and Rachael Luck. 2018. User Centered Design Approaches to Measuring Canine Behavior: Tail Wagging as a Measure of User Experience. *Proceedings Of The Fifth International Conference On Animal Computer Interaction* (Dec. 2018). https://doi.org/10.1145/3295598.3295599

[25] Johanna Schultz, Robert Cieri, Tasmin Proost, Rishab Pilai, Mitchell Hodgson, Fabian Plum, and Christofer Clemente. 2021. Tail Base Deflection but not Tail Curvature Varies with Speed in Lizards: Results from an Automated Tracking Analysis Pipeline. *Integrative and Comparative Biology* 61, 5 (2021), 1769–1782. https://doi.org/10.1093/icb/icab037

[26] Leonetti Silvia, Cimarelli Giulia, Hersh A. Taylor, and Ravignani Andrea. 2024. Why do dogs wag their tails? *Biol. Lett.* 20, 1 (2024). http://doi.org/10.1098/rsbl.2023.0407

[27] Ashish Singh and James Young. 2013. A Dog Tail for Utility Robots Exploring Affective Properties of Tail Movement. *Human Computer Interaction Interact* 8118 (2013), 403–419.

[28] Raman Srinivasan, Rytis Maskeliūnas, and Robertas Damaševičius. 2022. Markerless Dog Pose Recognition in the Wild Using ResNet Deep Learning Model. *Computers* 11, 1 (2022), 2. https://proxying.lib.ncsu.edu/index.php/login?url=https://www.proquest.com/scholarly-journals/markerless-dog-pose-recognition-wild-using-resnet/docview/2621273692/se-2

[29] G. Tembrock. 1968. Land mammals. *In: Animal communication* (1968), 359–373.