Estimating Withdrawal Time in Colonoscopies

Liran Katzir[1], Danny Veikherman[1], Valentin Dashinsky[2], Roman Goldenberg[3], Ilan Shimshoni[4], Nadav Rabani[1], Regev Cohen[3], Ori Kelner[3], Ehud Rivlin[3], and Daniel Freedman[3]

[1] Google
[2] DRW; work performed while at Alphabet
[3] Verily Life Sciences
[4] University of Haifa; work performed while at Alphabet

Abstract. The Colonoscopic Withdrawal Time (CWT) is the time required to withdraw the endoscope during a colonoscopy procedure. Estimating the CWT has several applications, including as a performance metric for gastroenterologists, and as an augmentation to polyp detection systems. We present a method for estimating the CWT directly from colonoscopy video based on three separate modules: egomotion computation; depth estimation; and anatomical landmark classification. Features are computed based on the modules' outputs, which are then used to classify each frame as representing forward, stagnant, or backward motion. This allows for the optimal detection of the change points between these phases based on efficient maximization of the likelihood; from which the CWT follows directly. We collect a dataset consisting of 788 videos of colonoscopy procedures, with the CWT for each annotated by gastroenterologists. Our algorithm achieves a mean error of 1.20 minutes, which nearly matches the inter-rater disagreement of 1.17 minutes.

Keywords: colonoscopy, detection, visual odometry, withdrawal time

1 Introduction

Colorectal Cancer (CRC) claims many lives per year [2,1]; however, as is well known, CRC may be prevented via early screening. In particular, the colonoscopy procedure is able to both detect polyps in the colon while they are still precancerous, and to resect them. There is, however, some variation in the quality of colonoscopy procedures, as performed by different gastroenterologists. This paper is concerned with one aspect which underlies this variation, namely the time spent by the endoscopist during the withdrawal phase of the colonoscopy.

Background and Motivation By way of background, we briefly describe the structure of a colonoscopy. When the procedure commences, the goal of the physician is to insert the colonoscope all the way to the end the colon, known as the cecum; see Fig. 1. This is known as the colonoscopic insertion. The time it takes to reach the cecum is known as the Cecal Intubation Time (CIT). During this process the physician moves the endoscope both forwards and backwards. From time to time only the wall of the colon is seen by the camera.

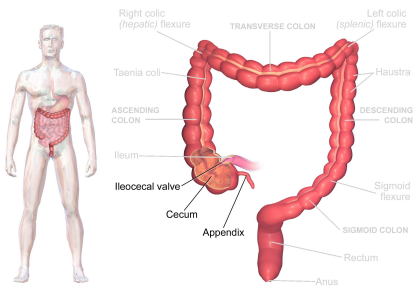

Fig. 1. The map of the colon, adapted with minor modifications from `wikipedia`. Our goal is to estimate the Colonoscopic Withdrawal Time (CWT), the duration elapsed from the time the colonoscope reaches the cecum until it has been entirely withdrawn.

Once the cecum has been reached the physician starts to slowly extract the colonoscope; the goal in this stage is to detect the polyps.[5] This phase is easier from the navigation point of view and the motion is usually backwards. Forward motion generally only occurs when a polyp is detected, examined and extracted. Thus, there is a clear distinction between colonoscopic insertion and colonoscopic withdrawal phases. The duration of the colonoscopic withdrawal is referred to as the Colonoscopic Withdrawal Time (CWT).

In this paper, we are particularly interested in measuring the CWT, as the CWT can impinge directly upon the successful detection and removal of polyps. Specifically, success in polyp detection is often measured by the Adenoma Detection Rate (ADR), defined as the fraction of procedures in which a physician discovers at least one adenomatous polyp. Several studies have found a positive correlation between the CWT and rates of neoplasia detection [7,42]. As a result, current guidelines recommend that the CWT be at least 6-7 minutes in order to achieve the desired higher ADR [8,33,26]. Higher ADR is directly linked to lower rates of interval CRC (a CRC which develops within 60 months of a negative colonoscopy screening) [24]; thus, ensuring a sufficiently high CWT is of paramount importance.

Overview of the Proposed Method Our goal is to estimate the CWT. In order to do so, we seek to find the first time point in the procedure where the operator stops inserting the endoscope deeper into the colon and starts the colonoscopic withdrawal phase. Typically, this happens at the cecum, see Fig. 1.

Our technique relies on the extraction of three key quantities: the egomotion of the camera, depth estimates of the colon, and detection of anatomical landmarks. The use of the egomotion is clear: it allows us to assess in which direction we are moving, which is an obvious differentiator between the colonoscopic insertion phase and the colonoscopic withdrawal phase. The use of the depth maps is more subtle: they help to distinguish between frames in which the camera is adjacent to the colon (which occur more often in colonoscopic insertion) and frames which see an unobstructed view of the colon (which occur more often

[5] Cheng *et al.* [11] studied the effect of detecting polyps during colonoscopic insertion and found it did not improve ADR.

in colonoscopic withdrawal). The use of landmark detection is straightforward: the presence of the relevant landmarks near the cecum – namely the appendiceal orifice, ileocecal valve, and triradiate fold – constitutes strong evidence the withdrawal phase has begun.

Our method then learns the optimal way of combining features based on these three key quantities in order to best estimate the probability that any given frame is in one of the following three phases: {`forward`, `stagnant`, `backward`}. Precise definitions of these three phases are given in Section 3; in brief, `forward` corresponds to the time moving forward from the rectum to the cecum; `stagnant` to the time spent inspecting the cecum; and `backward` to the time moving backward from the cecum to the rectum. Based on the per frame probabilities, we propose an optimization problem for determining the optimal temporal segmentation of the entire video into phases, which in turn gives the CWT. The main reason that this problem is challenging is that the accuracy of the estimated egomotion and depth images is not always high. We perform ablation studies to carefully show the role each of these features plays in the estimate.

Applications There are several uses of this segmentation procedure. The first use is as a performance metric for GIs. As we have already mentioned, the standard guidelines require that the physician spend at least six minutes after the ileocecal valve has been detected. In such scenarios the withdrawal time should not be measured manually nor by chronometric instruments. However, the amount of time the physician has actually spent can be easily estimated from the segmentation results, and used as a performance metric. Second, the system could be used in combination with existing automatic polyp detection systems, such as those described in [36,38]. In particular, some physicians prefer to detect and remove polyps only in the colonoscopic withdrawal, thus the polyp detection could be optionally turned off during the colonoscopic insertion phase. Indeed, given that our method is based on visual odometry computations, it could in principle store the approximate locations of any polyps detected in the colonoscopic insertion; then in the colonoscopic withdrawal when the physician returns to these locations she can be alerted to the existence of the polyp and take more care in finding it. Third, systems for detecting deficient colon coverage such as [17] could similarly be turned off during the colonoscopic insertion.

A final use case for such a system is as a tool for training novice endoscopists. In [25], it was shown empirically that as training proceeds, the CIT becomes shorter. This shows that the trainee becomes more proficient in navigation; however, this is not necessarily correlated to an improvement in ADR. Thus, navigation and polyp detection are two different capabilities which have to be mastered and the segmentation of the procedure can help in analyzing their proficiency separately.

Contributions and Paper Outline The main contributions of the paper are:

1. We propose a novel approach to withdrawal time estimation based on the combination of egomotion, depth, and landmark information.
2. We collect a gastroenterologist-annotated dataset of 1,447 colonoscopy videos for the purposes of training and validation.

3. We validate our algorithm on this dataset, showing that the algorithm leads to high quality segmentations. Specifically, our error is smaller in magnitude than inter-physician disagreement.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes our proposed method, focusing on egomotion and depth computation; detection of anatomical landmarks; derivation of features from the foregoing; a technique for combining these features into a per-frame phase classifier; and a method for video phase segmentation based on this classifier. Section 4 describes our dataset and presents experimental results, including ablation studies. Section 5 concludes the paper and discusses future work.

2 Related Work

Withdrawal Time Estimation We begin by discussing [12], which has addressed a very similar problem. The goal of this work is to detect the Cecal Intubation Time (CIT) based on the motion of the endoscope. The assumption made is that when the cecum is reached the magnitude of the motion is small. The relative motion is estimated between consecutive frames using a network which estimates the optical flow between them based on the Horn–Schunk algorithm [23]. Each motion is then classified as +1 for insertion, -1 for withdrawal, and 0 for stop. A section of the video where the sum of these values is lowest is assumed to be the turning point. One of the problems that must be dealt with is that in such videos, many of the frames are of low quality. Thus, at first frames are classified as informative or non-informative [5,13] and the optical flow algorithm is run only on the informative frames. This is due to the fact that on non-informative frames the estimated motion is usually classified as stop. In [28], the authors present a two-stage method for detecting the withdrawal point, i.e., the moment when the endoscope begins to withdraw. First, a deep network is trained to classify each frame whether it is an image of the ileocecal valve, the opening of the appendix or it contains background. Second, the trained classifier is used to generate a time series consisting of the per-frame ileocecal valve class probabilities. This temporal signal is then processed with sliding windows to identify the first window with a sufficient number of frames recognized as the ileocecal valve. Given this window, the withdrawal point is estimated as the last frame of the window.

Egomotion and Depth Estimation The first step of our algorithm runs a visual odometry and depth estimation procedure on a single colonoscopy video. In practice it is also possible to run these procedures separately. Visual odometry algorithms only recover the relative motion between consecutive frames. Classical methods extract and match feature points using descriptors such as SIFT or ORB and estimate from them the relative motion. A review of these methods can be found in [31,32]. Deep learning visual odometry also exists, for example [39]; and specialized visual odometry for endoscopy videos have also been developed [30,37,3].

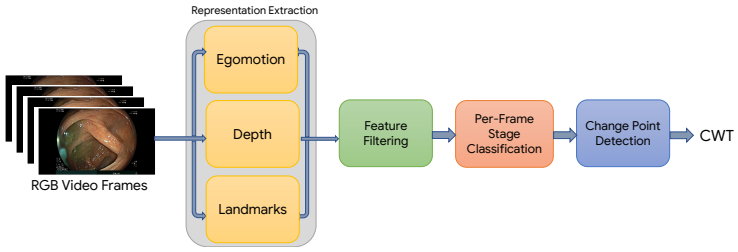

Fig. 2. The main blocks of the estimation pipeline.

Since relative motion and depth images are closely related as will be described below, deep learning methods which recover them together have been developed yielding superior results. In initial methods the training was performed in a supervised manner, where the ground truth depth and relative motion were available [4,40,44]. Following this unsupervised methods were developed [45,27,41,43,21]. The possibility of learning in an unsupervised manner is extremely useful in our case since for colonoscopy videos, ground truth depth and motion are not available. We obtain our depth images and relative motion using a state of the art method of that type [21], which will be reviewed below.
Phase Detection Our work may be thought as a type of phase detection on colonoscopy procedures, where there are three phases. There has been some related work on phase detection in medical procedures. In [34], the video frames are analyzed using a network based on AlexNet. The output of the network are detected tools and the phase to which the frame belongs. An HMM is then used to classify the phase of the frame taking into account temporal constraints. In a more recent work [16], the backbone is replaced with more modern networks and the HMM is replaced with a multi-stage convolutional network. This algorithm was used for analyzing surgery stages in [14]. A very recent validation of surgical phase detection on a much larger dataset is presented in [6].

3 Methods

3.1 The Estimation Pipeline

Our pipeline is illustrated in Fig. 2. It consists of three deep neural network modules which take as inputs RGB frames, the outputs of which are combined to generate a low dimensional representation. Following this stage, the pipeline has three sequential blocks which take the processed low dimensional representation to generate an estimate of the CWT. More specifically, the stages are as follows:
Representation Extraction: The RGB frames are first passed (in parallel) through deep neural network modules to transform them into a succinct representation. The models are egomotion and depth (Section 3.2), and a landmark classifier (Section 3.3). The exact representations extracted from each of these modules is described in Section 3.4.

Feature Filtering: The low dimensional representation derived above are noisy per-frame features. We therefore filter these features to make them more informative, by computing exponentially moving averages with different spans; absolute values; and running maxima (Section 3.5).
Per-Frame Phase Classifier: Given the robust features thus computed, this block combines the features into per-frame algorithmic phase probability estimates (Section 3.6).
Change-Point Detector: Given the per-frame probability estimates, we compute the change point which induces maximum likelihood (Section 3.7).

We now expand upon each of these stages of the pipeline.

3.2 Unsupervised Visual Odometry

From an RGB video of a colonoscopy, our goal is to estimate the relative motion between consecutive frames; and additionally, for each frame its corresponding depth image. Given that the vast majority of current endoscopes are monocular we consider the monocular setting. We adopt the *struct2depth* method [10,9,21], which is unsupervised. This is beneficial in the colonoscopy setting, as neither the ground truth position of the colonoscope nor the depth image are available for training. Furthermore, the method does not assume that the camera is calibrated, which can be useful in cases where the camera parameters are unknown.

Our method, like many algorithms for unsupervised depth and motion estimation [45,20,35,29,41,21], is based on the following property: corresponding points in two frames usually have very similar RGB values. This is especially true when the relative motion between the frames is small, as is the case in consecutive video frames. This property can be used to define a loss function known as the *view synthesis loss*, which combines both the egomotion and the depth estimated by the algorithm.

Concretely, the network consists of several sub-networks, as shown in Fig. 3. The depth estimation network is given as input I_t, the RGB frame of time t and produces the depth image D_t. In addition, given the image pair I_{t-1} and I_t, the pose network produces the relative pose / egomotion. Specifically, the

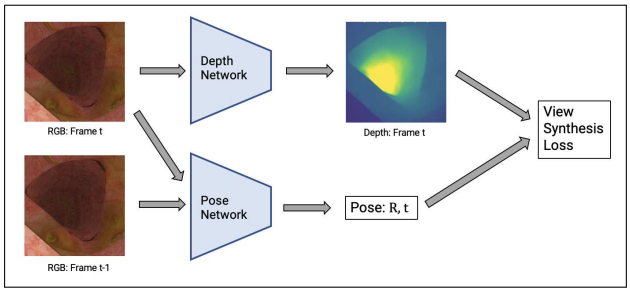

Fig. 3. The view synthesis loss and corresponding network architecture. See accompanying description in the text.

pose is defined as the rigid transformation (rotation matrix R and translation vector t) from the current frame t to the previous frame $t-1$. An additional intrinsics network can be used to produce an estimate of the internal calibration K; alternatively, a pre-learned K can be given as input.

The relationship between the geometric location of corresponding pixels in the two frames may be expressed by combining the depth, pose, and intrinsics information as:

$$z'p' = KRK^{-1}zp + Kt \tag{1}$$

where p and p' are the corresponding pixels in homogeneous coordinates and z and z' are their corresponding depth values. The view synthesis loss then compares the RGB values of the pixels at p in image I_{t-1} and p' in image I_t. In particular, the loss function is the L_1 loss of the RGB difference combined with an analogous loss based on structural similarity (SSIM).

In practice, there may sometimes be pairs of pixels for which Equation (1) does not hold, for example due to self occlusion in one frame or due to non-rigid motion of the colon. In such cases, when the relative transformation is applied to one depth image, there will be a difference in the corresponding depth in the other depth image. These depth differences are incorporated into the loss function, effectively reducing the weight for these pixel pairs.

3.3 The Landmark Prediction Module

As shown in Fig. 1, the cecum has several distinctive features which when detected may indicate the end of the insertion phase and beginning of withdrawal.
Appendiceal Orifice and Triradiate Fold Both of these landmarks (see Fig. 4) reside inside the cecum and may therefore be used as indicators of arrival at the cecum. We train a dual-head binary classification model to predict the presence of these two landmarks within the frame.

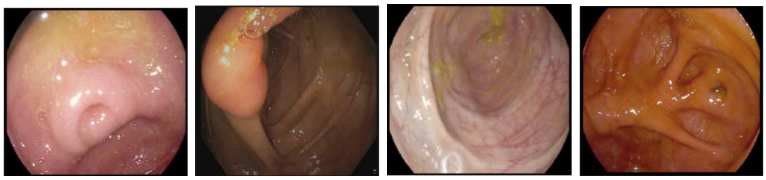

Fig. 4. Anatomical landmarks. Left to right: appendiceal orifice; ileocecal valve; cecum; triradiate fold.

Ileocecal Valve This landmark (see Fig. 4) is located just outside of the cecum and is much less distinctive than the other two. This lack of distinctiveness makes annotation of individual frames in which the ileocecal valve appears quite challenging. We deal with this by labelling the overall temporal region in which the ileocecal value is located; that is, we annotate an initial frame before which the ileocecal valve does not appear, and a final frame after which it does not appear. We then use this temporal region as weak supervision in a Multiple

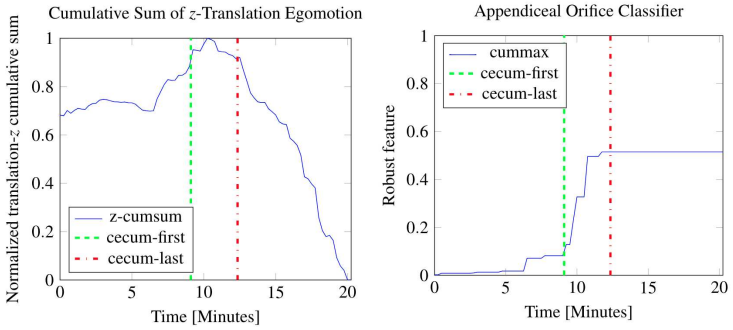

Fig. 5. Illustration of the role of various features in computing CWT. In both graphs, the triradiate fold is first observed at the green vertical line and last observed at the red vertical line; this denotes the boundaries of the time spent in the cecum. Left: egomotion. The cumulative sum of the z-translation egomotion (blue) can be seen to increase while the scope is inserted (before the green line), to remain constant in the cecum (between green and red), and to decrease during withdrawal (after the red line). Right: landmarks. The cumulative maximum of the appendiceal orifice classifier (blue) increases significantly in the cecum. See accompanying description in the text.

Instance Learning (MIL) [15] scheme where the model learns to predict if a frame has any distinctive features that are common to the cecum region.

Both of the networks - the dual head classification network for the appendiceal orifice and triradiate fold, and the MIL classifier for the ileocecal valve – share a common feature extractor backbone, a Resnet-50 CNN [22]. Each then has a separate fully-connected layer mounted on top of the resulting embedding to yield the probabilities for each of the three landmark features. All networks are trained together in an end-to-end fashion.

3.4 Representation Extraction

Thus far, we have described three conceptually different sources of feature representation: egomotion and depth are generated from a module described in Section 3.2, while the landmark detection module is described in Section 3.3. We now summarize the actual representations:

Egomotion: For each pair of consecutive frames, the egomotion estimate has 3 translation coordinates (x, y, z) and 3 rotation coordinates (ϕ_x, ϕ_y, ϕ_z), which are Euler angles. The egomotion estimate reflects the camera motion with respect to the camera coordinates.

Depth: For each frame, the depth map estimate consists of an entire image, where each pixel contains the depth estimate corresponding to that pixel.

Landmarks: For each frame and landmark (ileocecal valve; appendiceal orifice; and triradiate fold), the landmark feature is the classifier's probability estimate of the landmark's presence in the frame.

We now give some intuition as to why each of these features can play a role in the estimation of the Colonoscopic Withdrawal Time, beginnning with egomotion. Logically, a positive z-axis egomotion should indicate forward motion, while negative z-axis motion should indicate backward motion. This idea is illustrated on the left side of Figure 5, which shows a graph of the cumulative sum of the z-translation egomotion, overlaid with a ground truth annotation of the triradiate fold (cecum area). The cumulative sum of the z-translation egomotion, which we call the z-cumsum, can be seen to have the following rough characteristics: (1) there is a small positive z-cumsum while the scope moves forward; (2) there is roughly zero z-cumsum while the scope is in the cecum area; and (3) there is a large negative z-cumsum while the scope moves backward.

The depth maps can be useful in estimating the CWT, in that they help to distinguish between (1) frames in which the camera is adjacent to the colon, which occurs more often in colonoscopic insertion; and (2) frames which see an unobstructed view of the colon implying more pixels with high depth values, which occurs more often in colonoscopic withdrawal.

Finally, by definition the landmark features are extremely indicative of having reached the cecum. A clear view of the appendiceal orifice or the triradiate fold is strong evidence the withdrawal phase has started. The right side of Figure 5 shows that the landmark classifier for the appendiceal orifice is high around the cecum area. The cumulative maximum of the classification probabilities creates a very robust feature, which reaches its maximum once the navigation away from cecum area has begun.

3.5 Feature Filtering

Since raw per-frame features are extremely noisy on their own, we apply smoothing filters to aggregate the values of multiple frames. Specifically, the exponential-weighted-moving-average of a discrete signal s, denoted by $\text{ewma}(s)$, is calculated as follows:

$$\begin{aligned}\text{ewma}(s)[0] &= 0\\ \text{ewma}(s)[i] &= (1-\alpha)\cdot\text{ewma}(s)[i-1] + \alpha\cdot s[i]\end{aligned}$$

Here α determines the effective memory span: a span of m steps uses $\alpha = 2/m$. Given the above definition, we define smoothed versions of each of the features described in Section 3.4 as follows:

Smoothed z-Translation: Exponentially weighted moving average of egomotion's z-translation, where the moving average is taken with a variety of length spans: 1, 2, 3, and 4 minutes.

Smoothed Depth Quantiles: For each depth map, first the $(0.1, 0.25, 0.5, 0.75, 0.9)$-quantiles are extracted. Then, the quantiles are smoothed using exponentially weighted moving average with two different length spans: 2 and 4 minutes.

Peaked Smoothed Landmarks: The running maximum (cumulative maximum) of exponentially weighted moving average of a landmark's probability estimates. For each of the ileocecal valve, appendiceal orifice, and triradiate fold

landmarks, the moving average is taken with a variety of length spans: 8, 15, and 30 seconds.

Note that all filtered features at time t are computed using only representations observed up to time t. Moreover, all filtered features for frame t require only frames $t-1$ and t as well as the the filtered features for frame $t-1$. Therefore, the computation is well suited for an online setup.

3.6 Per-Frame Algorithmic Phase Classification

As the penultimate stage of our algorithm, we learn a per-frame classifier for the phases of the colonoscopy procedure. One of the useful side benefits of learning a per-frame classifier is that doing so greatly simplifies the pipeline, in an engineering sense. We define the following three phases:

1. `forward` marks the navigation from the rectum to cecum area. This phase is characterized by forward motion (positive z-translation); many frames see the colon's walls; and no landmarks are detected.
2. `stagnant` marks the start of the screening process (reaching the cecum area). This phase is characterized by little movement (z-translation is approximately 0) and possibly one or more landmarks have been detected.
3. `backward` marks the withdrawal from the cecum area. This phase is characterized by strong backward movement (negative z-translation); many frames view deep regions (the lumen of the colon); and one or more landmarks have been detected.

Thus, the colonoscopic insertion phase consists of the `forward` phase, while the colonoscopic withdrawal phase consists of the union of the `stagnant` phase and the `backward` phase.

For the classification model we use gradient boosted decision trees [18,19] on the filtered features from Section 3.5. The classifier is trained by minimizing the standard cross-entropy loss. The resulting model generates a probability estimate for each of the three algorithmic phases (which sum up to 1). The reason for choosing a gradient boosted decision trees over a deep-learning model is that at this point we are left with a small number of features and a small number of (relatively independent) samples. Moreover, the separation into modules greatly simplifies the training.

To generate the training data each procedure video is sampled at a fixed rate. Each frame's weight is set to account for duration variability. Specifically, the total weight of samples for a specified video is fixed: longer videos do not influence the training more than shorter videos. At inference time, a probability estimate is generated for every frame.

3.7 Change-Point Detection

The phases in an actual colonoscopy procedure form a sequence of contiguous segments with known order: `forward` $\rightarrow$ `stagnant` $\rightarrow$ `backward`. Using this order

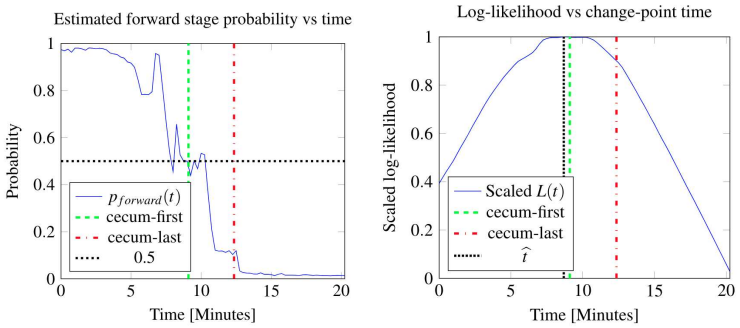

Fig. 6. Illustration of the per-frame classifier and the log likelihood of the change point. In both graphs, the triradiate fold is first observed at the green vertical line and last observed at the red vertical line; this denotes the boundaries of the time spent in the cecum. Left: the per-frame classifier probability of being in the `forward` phase. Observe that around the entrance to the cecum area the probability is very close to 0.5 Right: the log likelihood $L(t)$ of the change point vs. time. The estimated change point, which occurs at the maximum of $L(t)$, is shown with a black vertical line, and is very close to the manually annotated change point shown in green.

constraint, we seek a segmentation of the per-frame probability estimate which yields the maximum likelihood solution.

Let $\widehat{p}_{c,t}$ denote the probability estimate for phase c at frame t. For two phases, denote by $L(t)$ the log-likelihood of the change-point occurring at time t. Then we may write

$$L(t) = \sum_{t' \leq t} \log \widehat{p}_{1,t'} + \sum_{t' > t} \log \widehat{p}_{2,t'}$$

The log-likelihood (and the rest of the analysis) can be similarly extended into two splits and three phases.

The optimal change point, $\hat{t}$, is chosen such that $\hat{t} = \arg\max_t L(t)$. (If there are multiple such points, we take the earliest one.) We now present an online computational method for $\hat{t}$. To this end, we define $V[c,t]$ as the value of the optimal log-likelihood for the $1, 2, \ldots, t$ frames which end with phase c. We have $V[1,0] = V[2,0] = 0$ and

$$\begin{aligned} V[1,t] &= V[1,t-1] + \log \widehat{p}_{1,t} \\ V[2,t] &= \max\left(V[1,t-1], V[2,t-1]\right) + \log \widehat{p}_{2,t} \end{aligned} \tag{2}$$

Finally, $L(\hat{t}) = V[2,T]$ where T is the index of last frame. The value $\hat{t}$ can be retrieved by setting $\hat{t} = t$ where t is the last index for which $V[1,t-1] > V[2,t-1]$. Note that Equation (2) can be easily extended to accommodate 3 (or any) number of phases.

$L(t)$ is visualized for a given sequence in Fig. 6. In this figure, we used the colonoscopic insertion phase as the first segment and the colonoscopic withdrawal phase (union of the `stagnant` phase and `backward` phase) as the second segment.

4 Results

4.1 The Dataset

The dataset consists of real de-identified colonoscopy videos (acquired from Orpheus Medical) of procedures performed at an academic hospital. In total, there are 788 videos. All videos were recorded at 30 frames per second, with a compression rate of 16 mbps. The distribution of frames in each phase is: `forward` - 45.2%; `stagnant` - 12.1%; `backward` - 42.7%. To maximize the data usage, we employ 5-fold cross validation on the entire set of videos.

The videos were annotated offline by gastroenterologist annotators, drawn from a pool of four with 4, 7, 7, and 9 years of experience. For every landmark, the annotators were asked to carefully mark the first and last frame in which the landmark is visible (for every contiguous period separately), as well as the the time point which marks the start of the withdrawal phase.

4.2 Hyperparameters

The hyperparameters for Section 3.2 and Section 3.3 were chosen separately (each for its own sub-task). We focus here on the hyperparameters for the per-frame classifier Section 3.6. The video was sub-sampled at a fixed rate of a frame every 15 seconds. Technically, computational resources allow for the training of the per-frame classifier without sub-sampling. However, consecutive frames are highly correlated and their inclusion does not improve overall metrics. Moreover, sub-sampling is useful in stochastic gradient boosting models to create diversity. The other hyperparameters relate to gradient boosting, namely: the number of trees; the maximum tree depth; the learning rate (0.03); and the subsample probability (0.5). The effect on performance of the choice of both the number of trees and maximum tree depth is presented in a separate discussion in Section 4.4.

4.3 Results

We report various statistics for the absolute error of the CWT. Specifically, if the ground truth time is t and the estimated time is $\hat{t}$, then the absolute error is denoted $\Delta t = |\hat{t} - t|$. To capture the error distribution we use: (1) the mean absolute error (MAE) and (2) the i^{th} percentile / quantile, which we denote as $q_i(\Delta t)$. To reduce variability, each estimate was taken as the median of 9 bootstrap runs. The results reported are attained using the hyperparameter settings described in Section 4.2, along with 1000 trees with maximum depth 1. (The role of the latter hyperparameters is analyzed in Section 4.4.)

The results are reported in Table 1, where times are reported in minutes. The MAE is 1.20 minutes, while the median absolute error is 0.58 minutes, and the 75^{th} percentile is 1.32 minutes. In order to calibrate the size of these errors, we compared them with the disagreement between gastroenterologist experts. We provided 45 videos to be analyzed by 3 gastroenterologists and measured the difference between the earliest and latest of the 3 annotations, which we

Table 1. Statistics for the CWT error in minutes: mean and various percentiles. Notice that the algorithm error and the annotator spread are roughly the same.

	mean	50^{th}	75^{th}
$\lvert\Delta t\rvert$ = Algorithm Error	1.20	0.58	1.32
Annotator Spread	1.17	0.62	1.38

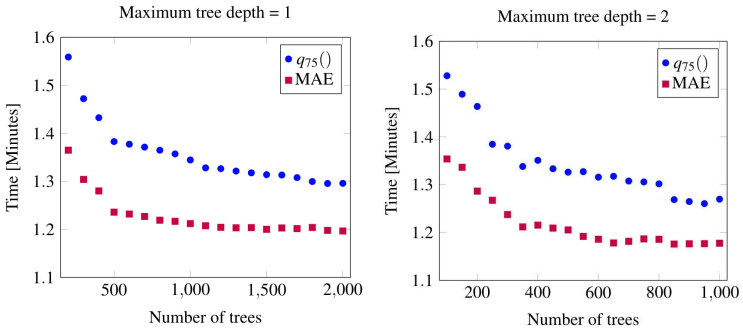

Fig. 7. Sensitivity of the algorithm to hyperparameters: number of trees and maximum tree depth. In each case, both MAE and the 75^{th} percentile error are plotted vs. number of trees. Left: maximum tree depth of 1. Right: maximum tree depth of 2.

refer to as the *annotator spread*. The mean annotator spread is 1.17 minutes, while the median annotator spread is 0.62 minutes; and the 75^{th} percentile is 1.38 minutes. Note that the algorithm error and the annotator spread are quite similar in value. More specifically, in the case of the median and 75^{th} percentile, the annotator spread is higher than the algorithm error; while in the case of the mean, the algorithm's error is slightly higher.

4.4 Sensitivity Analysis

We now study the sensitivity of the estimator to the choice of hyperparameters, specifically the number of trees used in the gradient boosting algorithm, as well as the maximum depth of these trees. Fig. 7 shows graphs of the performance - as measured by both MAE as well as the 75^{th} percentile error $q_{75}(\Delta t)$ - vs. the number of trees. There are two separate graphs, corresponding to maximum tree depths of 1 and 2. We note that the graphs are not entirely smooth due to the optimization performed in the change-point detection algorithm (see Section 3.7). Nonetheless, overall there is little variability in the metrics, especially once the number of trees has increased past 500 (for depth 1) or 300 (for depth 2). Thus, it is strongly evident that the estimate is quite robust to the choice of these hyperparameters.

4.5 Analysis of Feature Importance

We now turn to analyzing the role of each of the features in the algorithm, in particular their role in the per-frame classifier. To assess the role of each feature,

we use the relative importance heuristic described in [18], which we report in Table 2. We see that the egomotion is quite important, accounting for 40.7% of the total contribution. Furthermore, the landmarks are also very important, accounting for a total of 56.3% between the three of them; the triradiate fold is the most important, followed by the ileocecal valve and the appendiceal orifice. Finally, the depth appears to be less important, accounting for only 3.0%.

4.6 Ablation Studies

We continue our analysis of the algorithm by performing an ablation study. In particular, we test the effect on performance of removing each of the features one by one, and then retraining the per-frame classifier. The features we study are: (1) egomotion (2) depth (3) all landmarks considered together. In the case of each retraining, we used maximum tree depth equal to 1 and chose the number of trees which minimized $q_{75}(\Delta t)$ for each ablation test.

The results are shown in Table 3. Note that by removing the landmarks, the performance suffers the most: the MAE increases from 1.20 to 1.85, with a concomitant increase in $q_{75}(\Delta t)$. Next is the egomotion, followed by the depth. Interestingly, the ablation study shows that the depth is still quite important: removing it increases the MAE from 1.20 to 1.33, a 10.8% relative increase.

Table 2. Feature importance Analysis.

Feature	Importance
Egomotion	0.407
Depth	0.030
Appendiceal orifice	0.090
Ileocecal value	0.331
Triradiate fold	0.143

Table 3. Ablation studies. Performance is shown when each feature is removed.

Ablation	MAE	$q_{75}(\Delta t)$
None	1.20	1.32
z-translation	1.42	1.68
Depth	1.33	1.42
Landmarks	1.85	2.08

5 Conclusions

We presented a system for estimating the colonoscopic withdrawal time in colonoscopy procedures. The method is based on combining features from three disparate sources: egomotion, depth, and landmark classification. The resulting algorithm has been validated on a GI-annotated dataset, and has demonstrated an error which is smaller than the inter-rater disagreement. As a result, the algorithm shows promise in a variety of applications, including as a performance metric for GIs, as an add-on to existing polyp detection systems, and as part of a training system for novice endoscopists.

Acknowledgment The authors would like to thank Gaddi Menahem and Yaron Frid of Orpheus Medical Ltd. for helping in the provision of data.

References

1. Cancer Facts & Figures 2019. https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2019.html, accessed: 2019-11-26
2. Colorectal Cancer Fact Sheet 2018. http://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf, accessed: 2020-01-08
3. Aghanouri, M., Ghaffari, A., Serej, N.D., Rabbani, H., Adibi, P.: New image-guided method for localisation of an active capsule endoscope in the stomach. IET Image Processing **13**(12), 2321–2327 (2019)
4. Almalioglu, Y., Saputra, M.R.U., de Gusmao, P.P., Markham, A., Trigoni, N.: GANVO: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 5474–5480. IEEE (2019)
5. Ballesteros, C., Trujillo, M., Mazo, C., Chaves, D., Hoyos, J.: Automatic classification of non-informative frames in colonoscopy videos using texture analysis. In: Iberoamerican Congress on Pattern Recognition. pp. 401–408. Springer (2016)
6. Bar, O., Neimark, D., Zohar, M., Hager, G.D., Girshick, R., Fried, G.M., Wolf, T., Asselmann, D.: Impact of data on generalization of AI for surgical intelligence applications. Scientific reports **10**(1), 1–12 (2020)
7. Barclay, R.L., Vicari, J.J., Doughty, A.S., Johanson, J.F., Greenlaw, R.L.: Colonoscopic withdrawal times and adenoma detection during screening colonoscopy. New England Journal of Medicine **355**(24), 2533–2541 (2006)
8. Barclay, R.L., Vicari, J.J., Doughty, A.S., Johanson, J.F., Greenlaw, R.L.: Colonoscopic withdrawal times and adenoma detection during screening colonoscopy. New England Journal of Medicine **355**(24), 2533–2541 (2006)
9. Casser, V., Pirk, S., Mahjourian, R., Angelova, A.: Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In: Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19) (2019)
10. Casser, V., Pirk, S., Mahjourian, R., Angelova, A.: Unsupervised monocular depth and ego-motion learning with structure and semantics. In: CVPR Workshop on Visual Odometry and Computer Vision Applications Based on Location Cues (VOCVALC) (2019)
11. Cheng, C.L., Kuo, Y.L., Liu, N.J., Tang, J.H., Fan, J.W., Lin, C.H., Tsui, Y.N., Lee, B.P., Hung, H.L.: Comparison of polyp detection during both insertion and withdrawal versus only withdrawal of colonoscopy: A prospective randomized trial. Journal of Gastroenteroly and Hepatology **34**(8), 1377–1383 (2019)
12. Cho, M., Kim, J.H., Hong, K.S., Kim, J.S., Kong, H.J., Kim, S.: Identification of cecum time-location in a colonoscopy video by deep learning analysis of colonoscope movement. PeerJ **7** (2019). https://doi.org/https://doi.org/10.7717/peerj.7256
13. Cho, M., Kim, J.H., Kong, H.J., Hong, K.S., Kim, S.: A novel summary report of colonoscopy: timeline visualization providing meaningful colonoscopy video information. International Journal of Colorectal Disease **33**(5), 549–559 (2018)
14. Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N.: TeCNO: Surgical phase recognition with multi-stage temporal convolutional networks. arXiv preprint arXiv:2003.10751 (2020)
15. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence **89**(1-2), 31–71 (1997)

16. Farha, Y.A., Gall, J.: MS-TCN: Multi-stage temporal convolutional network for action segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3575–3584 (2019)
17. Freedman, D., Blau, Y., Katzir, L., Aides, A., Shimshoni, I., Veikherman, D., Golany, T., Gordon, A., Corrado, G., Matias, Y., Rivlin, E.: Detecting deficient coverage in colonoscopies. IEEE Transactions on Medical Imaging **39**(11), 3451–3462 (2020). https://doi.org/10.1109/TMI.2020.2994221
18. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. The Annals of Statistics pp. 1189–1232 (2001)
19. Friedman, J.H.: Stochastic gradient boosting. Computational Statistics & Data Analysis **38**(4), 367–378 (2002)
20. Garg, R., Carneiro, G., Reid, I.: Unsupervised CNN for single view depth estimation: Geometry to the rescue. ECCV (2016)
21. Gordon, A., Li, H., Jonschkowski, R., Angelova, A.: Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In: Proceedings of the IEEE International Conference on Computer Vision (2019)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
23. Horn, B.K., Schunck, B.G.: Determining optical flow. In: Techniques and Applications of Image Understanding. vol. 281, pp. 319–331. International Society for Optics and Photonics (1981)
24. Kaminski, M.F., Wieszczy, P., Rupinski, M., Wojciechowska, U., Didkowska, J., Kraszewska, E., Kobiela, J., Franczyk, R., Rupinska, M., Kocot, B., Chaber-Ciopinska, A., Pachlewski, J., Polkowski, M., Regula, J.: Increased rate of adenoma detection associates with reduced risk of colorectal cancer and death. Gastroenterology **153**(1), 98–105 (2017)
25. Lee, S.H., Chung, I.K., Kim, S.J., Kim, J.O., Ko, B.M., Hwangbo, Y., Kim, W.H., Park, D.H., Lee, S.K., Park, C.H., Baek, I.H., Park, D.I., Park, S.J., Ji, J.S., Jang, B.I., Jeen, Y.T., Shin, J.E., Byeon, J.S., Eun, C.S., Han, D.S.: An adequate level of training for technical competence in screening and diagnostic colonoscopy: a prospective multicenter evaluation of the learning curve. Gastrointestinal Endoscopy **67**(4), 683–689 (2008)
26. Lee, T., Blanks, R., Rees, C., Wright, K., Nickerson, C., Moss, S., Chilton, A., Goddard, A., Patnick, J., McNally, R., Rutter, M.: Longer mean colonoscopy withdrawal time is associated with increased adenoma detection: evidence from the Bowel Cancer Screening Programme in England. Endoscopy **45**(01), 20–26 (2013)
27. Li, R., Wang, S., Long, Z., Gu, D.: UndeepVO: Monocular visual odometry through unsupervised deep learning. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 7286–7291. IEEE (2018)
28. Li, Y., Ding, A., Cao, Y., Liu, B., Chen, S., Liu, X.: Detection of endoscope withdrawal time in colonoscopy videos. In: IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 67–74 (2021)
29. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised learning of depth and egomotion from monocular video using 3D geometric constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5667–5675 (2018)
30. Pinheiro, G., Coelho, P., Salgado, M., Oliveira, H.P., Cunha, A.: Deep homography based localization on videos of endoscopic capsules. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 724–727. IEEE (2018)

31. Scaramuzza, D., Fraundorfer, F.: Visual odometry [tutorial]. IEEE Robotics & Automation Magazine **18**(4), 80–92 (2011)
32. Scaramuzza, D., Fraundorfer, F.: Visual odometry part ii: Matching, robustness, optimization and applications. IEEE Robotics & Automation Magazine **19**(2), 78–90 (2012)
33. Simmons, D.T., Harewood, G.C., Baron, T.H., Petersen, B.T., Wang, K.K., Boyd-Enders, F., Ott, B.J.: Impact of endoscopist withdrawal speed on polyp yield: implications for optimal colonoscopy withdrawal time. Alimentary Pharmacology & Therapeutics **24**(6), 965–971 (2006)
34. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Transactions on Medical Imaging **36**(1), 86–97 (2016)
35. Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: DeMoN: Depth and motion network for learning monocular stereo. CVPR (2017)
36. Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., Baldi, P.: Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology **155**(4), 1069–1078 (2018)
37. Wang, M., Shi, Q., Song, S., Hu, C., Meng, M.Q.H.: A novel relative position estimation method for capsule robot moving in gastrointestinal tract. Sensors **19**(12), 2746 (2019)
38. Wang, P., Xiao, X., Brown, J.R.G., Berzin, T.M., Tu, M., Xiong, F., Hu, X., Liu, P., Song, Y., Zhang, D., Yang, X., Li, L., He, J., Yi, X., Liu, J., Liu, X.: Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nature Biomedical Engineering **2**(10), 741–748 (2018)
39. Wang, S., Clark, R., Wen, H., Trigoni, N.: DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 2043–2050. IEEE (2017)
40. Yang, N., Wang, R., Stuckler, J., Cremers, D.: Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 817–833 (2018)
41. Yin, Z., Shi, J.: GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1983–1992 (2018)
42. Young, Y.G., Sooand, E.H., Seok, K.J., Seok, J.J., Hyung, K.S., Seok, M.H., Seok, L.E., Hyun, K.S., Kyu, S.J., Seok, L.B., Yong, J.H.: Colonoscopic withdrawal time and adenoma detection in the right colon. Medicine **97** (2018)
43. Zhan, H., Garg, R., Saroj Weerasekera, C., Li, K., Agarwal, H., Reid, I.: Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 340–349 (2018)
44. Zhou, H., Ummenhofer, B., Brox, T.: DeepTAM: Deep tracking and mapping. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 822–838 (2018)
45. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1851–1858 (2017)