

Computer-aided Detection of Diagnostic and Therapeutic Operations in Colonoscopy Videos

Yu Cao, *Student Member, IEEE*, Danyu Liu, *Student Member, IEEE*, Wallapak Tavanapong, *Member, IEEE*, Johnny Wong, *Member, IEEE*, JungHwan Oh, *Member, IEEE*, and Piet C. de Groen

Abstract—Colonoscopy is an endoscopic technique that allows a physician to inspect the inside of the human colon and to perform – if deemed necessary – at the same time a number of diagnostic and therapeutic operations. In order to see the inside of the colon, a video signal of the internal mucosa of the colon is generated by a tiny video camera at the tip of the endoscope and displayed on a monitor for real-time analysis by the physician. We have captured and stored these videos in digital format and call these colonoscopy videos. Based on new algorithms for instrument detection and shot segmentation, we introduce new spatio-temporal analysis techniques to automatically identify an operation shot—a segment of visual data in a colonoscopy video that corresponds to a diagnostic or therapeutic operation. Our experiments on real colonoscopy videos demonstrate the effectiveness of the proposed approach. The proposed techniques and software are useful for (1) post-procedure review for causes of complications due to diagnostic or therapeutic operations; (2) establishment of an effective content-based retrieval system to facilitate endoscopic research and education; and (3) development of a systematic approach to assess and improve the procedural skills of endoscopists.

Index Terms—Biomedical instruments, Video signal processing, Biomedical image processing, Object detection.

I. INTRODUCTION

COLORECTAL cancer is the second leading cause of cancer-related deaths behind lung cancer in the United States [1]. As the name implies, colorectal cancers are malignant tumors that develop in the colon and rectum. The survival rate is higher if the cancer is found and treated early

Copyright (c) 2006 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

Manuscript received April 6, 2006. This work was supported in part by the U.S. National Science Foundation under Grant IIS-0513777, IIS-0513809, and IIS-0513582, and the Mayo Clinic.

Yu Cao, Danyu Liu, Wallapak Tavanapong, Johnny Wong are with the Department of Computer Science, Iowa State University, Ames, IA 50011-1040, U.S.A. (Correspondence to: Dr. Wallapak Tavanapong, Department of Computer Science, Iowa State University, Ames, IA 50011-1040, U.S.A, phone:+1-515-294-2987. fax: +1-515-294-0258; email: impact.isu@cs.iastate.edu)

JungHwan Oh is with the Department of Computer Science and Engineering, University of North Texas, Denton, TX 76203, U.S.A. (email:jhoh@cse.unt.edu)

Piet C. de Groen is with Mayo Clinic College of Medicine, Mayo Clinic, Rochester, MN 55905, U.S.A.

before metastasis to lymph nodes or other organs occurs.

The colon is a hollow, muscular tube about 150 centimeters long. Colonoscopy is an important screening tool for colorectal cancer. It allows for inspection of the entire colon and provides the ability to perform a number of diagnostic and therapeutic operations such as tissue-sampling and polyp removal during a single procedure. A colonoscopic procedure consists of two phases: *insertion phase* and *withdrawal phase*. The video camera generates a video signal of the internal mucosa of the colon during the two phases. The video data are displayed on a monitor for real-time analysis by the endoscopist. In current practice, these video data are not routinely stored for either manual or automated post-procedure analysis.

Recent years have seen research on techniques for guiding a colonoscope during a colonoscopic procedure [2], development of colonoscope hardware [3], analyses of microscopic images from biopsies of colon tissues [4], analyses of images from colonoscopic procedures for tumor detection [5], and virtual colonoscopy [6, 7]. In virtual colonoscopy, a virtual colon is reconstructed from Computer Tomography (CT) cross-sectional images of the abdomen of a patient. CT images are significantly different from those of colonoscopic procedures. To the best of our knowledge, there are no prior techniques that can automatically identify a video segment that corresponds to a diagnostic or therapeutic operation in colonoscopy videos. In this paper, we focus on developing such techniques. They are useful for (1) post-procedure review for causes of complications due to diagnostic or therapeutic operations, i.e., quality assessments; (2) establishment of an effective content-based retrieval system to facilitate endoscopic research and education; and (3) development of a systematic approach to assess and improve the procedural skills of endoscopists, both in training and in practice.

Video segmentation is a necessary first step that divides a video file into smaller meaningful segments. In recent years, many video segmentation techniques have been designed to (1) detect important semantic units such as scenes and shots for movies and news [8, 9], (2) track simple objects with homogeneous visual characteristics [10] or specific objects such as text captions [11] or humans' movements [12], or (3) detect interesting sports events such as field goals or foul events [13, 14]. No video segmentation techniques have been designed for colonoscopy videos. Existing segmentation techniques proposed for other video domains are not suitable

for colonoscopy videos since semantic units of colonoscopy videos are significantly different. Furthermore, colonoscopy videos possess unique characteristics. For example, colonoscopy videos contain many blurry frames due to frequent shifts of the camera position while the camera is moving along the colon. Current endoscopes are equipped with a single, wide-angle lens that cannot be focused. Hence, sharpness, brightness, and contrast of the image are optimized using the endoscopist's skills.

Our contribution is as follows. First, we define a new type of semantic units called "operation shot". An *operation shot* is a *segment of visual and audio data that correspond to a diagnostic or therapeutic operation in a colonoscopy video*. Second, we introduce new techniques to detect operation shots based on detection of diagnostic or therapeutic instruments. We evaluate the effectiveness of the proposed techniques on real colonoscopy videos. This paper is an extended version of our conference paper [15]. The extension includes (1) a new algorithm to detect the *insertion direction*---a direction in which diagnostic or therapeutic instruments appear in the field of view of the camera. The insertion direction is fixed for a given endoscope, but can vary among different endoscopes, (2) our enhanced algorithm that utilizes our new image enhancement technique, the detected insertion direction, and new shape features to detect operation shots more accurately, and (3) a more comprehensive performance study on the effectiveness of the algorithms. The new enhancements make our approach more effective in detecting operation shots yet flexible to handle videos generated from different endoscopes.

The remainder of this paper is organized as follows. In Section II, we provide background on instrument, diagnostic and therapeutic operations. We present new analysis techniques for operation shot detection in Section III. We evaluate the effectiveness of the proposed techniques in Section IV. Finally, we offer our concluding remarks in Section V.

II. BACKGROUND ON DIAGNOSTIC AND THERAPEUTIC OPERATIONS

An endoscope has instrument channels that allow the insertion of flexible accessories such as biopsy forceps, cytology brushes, sclerotherapy needles, and diathermy snares from a port on the endoscope control head through the shaft and into the field of view. These instruments are used for tissue-sampling, other diagnostic and therapeutic procedures. Biopsy forceps used for tissue sampling consist of a pair of sharpened cups, a spiral metal cable, and a control handle. The tissue specimen is used for microscopic examining for its structure or for searching for the presence of infectious agents such as *Helicobacter pylori*. "Hot" biopsy forceps (allowing the passage of current) and diathermy snares are used for polyp removal. Fig. 1(a), Fig. 1(b), and Fig. 1(c) show some examples of such instruments. Fig. 1(d) and Fig. 1(e) depict images from actual colonoscopic procedures when a snare or biopsy forceps are in use. The instruments may appear in the images in a

different position (e.g., from the bottom right corner, bottom left corner) depending on the endoscope model.

Our video capturing process is designed in such way that it ensures patients' privacy and does not disrupt the normal routine of the procedure. No patient identifiable information, whether audio, video or superimposed text, is recorded in videos used in this study. Patients provide verbal consent to capture images and sign a HIPAA "Authorization to Use and Disclose Protected Health Information" form per Mayo Clinic Institutional Review Board guidelines.

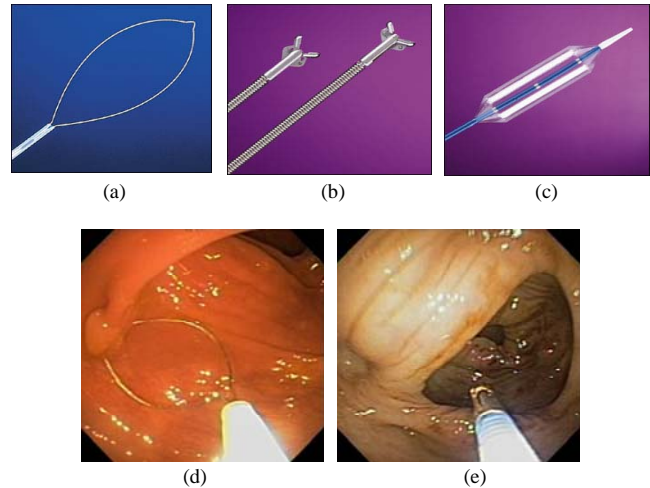


Fig.1. Examples of instruments (a) Snare, (b) Biopsy forceps, (c) Ballon. Example images with instruments during colonoscopy (d) Use of a snare (e) Use of biopsy forceps

III. PROPOSED APPROACH FOR OPERATION SHOT DETECTION

We map the problem of detecting operation shots to the problem of identifying instruments used in diagnostic or therapeutic operations since the operations cannot be performed without these instruments. Given a variety of instruments, we further map the problem of detecting instruments to the problem of detecting the cables of the instruments as the cable is frequently present in an operation regardless of the types of the instruments. The remaining difficulties are as follows. First, the cables come in different directions, colors, and sizes. Second, the cable appears very bright in many frames; this is related to light required to illuminate the colon. The light beam exits the endoscope tip directly adjacent to the instrument channel opening causing any cable exiting this channel to be exposed to undispersed light at maximal intensity which may result in over-exposure of the camera's CCD chip. The same holds true for colon mucosa and contents that are in immediate proximity of the endoscope tip. The intense brightness and resulting over-exposure may mask the actual color information of the cable and adjacent colon wall, making it difficult to utilize color features for operation shot detection. Last, the appearance of the cable in a frame varies from one frame to another during an operation. Depending on the location in the colon, the space between the endoscope tip and the lesion, and the position of the lesion

within the colon, one may see only the head of the instrument (without the cable) or the head of the instrument with a shorter of longer segment of the cable.

To overcome the above difficulties, we propose a new spatio-temporal segmentation approach for operation shot detection. There are six steps (Image Preprocessing, Identification of Insertion Direction of Instruments, Region Filtering, Region Merging, Region Matching, and Shot Segmentation) in this approach. The first five steps together identify the presence of the cable in each of the images extracted from the input colonoscopy video. The first step is image preprocessing. In this step, each selected image is first enhanced by our new light reflection filtering algorithm. The enhanced image is then segmented into a number of regions. Next, we identify the insertion direction of an instrument; this is useful for removing irrelevant regions (i.e., regions that are not part of the cable) in the region filtering step. To remove the case that the instrument is falsely segmented into several regions, we use the region merging step to combine these regions into one potential cable region. Next, the region matching step matches the candidate regions in the image with the pre-defined template of the cables. We use the terms *cable image* and *non-cable image* to refer to an image with the cable and without the cable, respectively. The region matching step outputs a 1 when the image has at least one region sufficiently similar to the cable templates. Otherwise, the image is considered a non-cable image and the region matching step outputs a 0. Based on temporal information, the shot segmentation step utilizes our pre-defined rules to determine the boundaries of operation shots given a series of binary numbers from the region matching step. The details of each step of our algorithm are discussed below.

A. Image Preprocessing

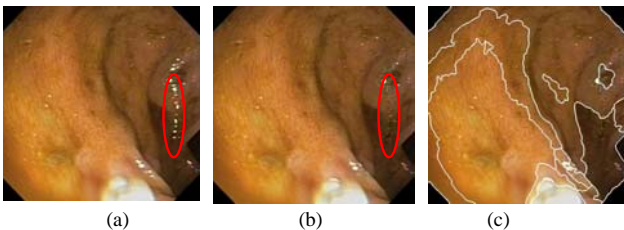


Fig. 2. Image examples in the procedure of light reflection filtering for the image preprocessing step: (a) Original color image, (b) Image after removing the light reflected regions, (c) Segmented image using JSEG

This step includes three stages. Fig. 2(a), Fig. 2(b), and Fig. 2(c) illustrate image examples in these stages. In the first stage, we extract t images per second to reduce the analysis time for the subsequent steps. A sequence of extracted images from the input colonoscopy video is called the reduced colonoscopy video. The smaller the value of t , the larger the reduction in the processing time, but the larger the difference in the actual and the detected boundaries. We explain the selection of the appropriate value of t in Section IV. Fig. 2(a) is an example of the selected image. Fig. 2(a) shows many small over-bright white areas (in an ellipse that we manually draw to indicate the

light reflected areas). They are generated due to light reflected on substances (i.e., mucus, water, cleansing agent, air bubbles, etc.) covering the colon wall, when the light beam hits the reflecting surface at a 90 degree angle. They may considerably disturb subsequent image processing techniques such as edge detection, texture analysis, and segmentation. We include our new light reflection filtering as the second stage of the image preprocessing to address this problem. We observe that many light-reflected areas are small. The majority of the pixels inside a light-reflected area can be identified as edge pixels by commonly used edge detectors. Based on these observations, we develop the following filtering procedure.

- **Step 1:** Using Sobel edge detector and the morphology closing using a flat, disk-shaped structuring element [16] to extract the edge pixels from each image. This step generates a binary image where the white curvilinear structures represent the real edges and small isolated white regions represent small over-bright areas in the original image, respectively.
- **Step 2:** Using a predefined $W \times W$ sliding window to scan the entire image. If we find that more than 85% of the pixels inside the window are edge pixels and more than 90% of the pixels on the boundary of the window are not edge pixels, we claim that the area delineated by the window is an actual over-bright white area. The percentage thresholds (85% and 90%) are derived from experiments on different colonoscopy videos. Our image enhancement technique is not very sensitive to these thresholds since the results did not vary much when we performed experiments with different threshold values between 80% and 95%.
- **Step 3:** For the area inside the sliding window that is determined to be an actual over-bright white area in Step 2, we calculate the average pixel intensity I_{ave} of all the pixels on the boundary of this window. Next, we replace the intensity value of all the pixels inside the sliding window with I_{ave} . As a result, we get an enhanced image without the previously detected over-bright areas.

The generated image is illustrated in Fig. 2(b). We can see that the majority of the over-bright white areas in the indicated ellipse have been removed. Next, each enhanced image is segmented into a number of regions using JSEG [17] since JSEG performs better than the k-mean image segmentation technique and Blobworld [18] for our videos. We experimentally determine the appropriate parameter values for JSEG such that the largest number of desirable segmentation results observed by humans is obtained. Those parameters are effective for endoscopy videos from different endoscope models, for example, colonoscopy and gastroscopy videos.

B. Identification of Insertion Direction of Instruments

This step identifies the insertion direction of instruments. Only one endoscope is used per colonoscopic procedure and standard colonoscopy models have only one working channel, in which instruments can be inserted. The insertion direction is determined by the location of the working channel in relation to the camera lens (see Fig. 3(a)). Therefore, each colonoscopy

video has one insertion direction. The instrument can appear in the field of view of the endoscope in any direction, depending on the model of the endoscope used in the procedure. We classify these directions into eight general directions as shown in Fig. 3(c) and associate insertion direction i with a triangular “Area i ” where $1 \leq i \leq 8$ as shown in Fig. 3(b). The ability to identify the correct triangular area can greatly improve the accuracy and decrease the processing time of subsequent steps.

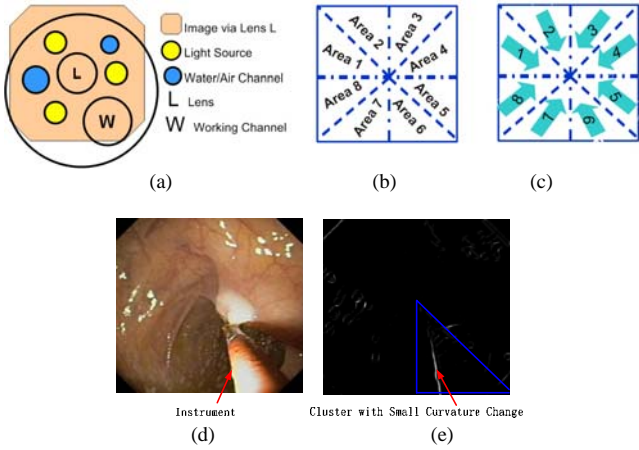


Fig. 3: Possible triangular areas and insertion directions of instruments: (a) Various components of the tip of a current model colonoscope projected on top of the image area; note the position of the working channel in relation to the lens (b) Eight triangular areas, (c) Eight insertion directions corresponding to eight triangular areas; Image examples for insertion direction determination: (d) Original color image, (e) Binary image after edge enhancement and clustering;

We propose an algorithm to identify the insertion direction of the instrument in each image. The cable of the instrument has a tubular shape. The tubular shape has a strong curvilinear structure at the proximal end (most central in the image) with linear line shape of the longitudinal edges of the cable. If we can find this kind of shape in one of the eight triangular areas (for example, “Area 6”) and the orientation of the shape is close to the insertion direction of that triangular area (for example, the angle between the orientation of the object and the insertion direction 6 of triangular area 6 is very small), it is very likely that the insertion direction of this image is the same as the orientation of the shape. For each video, we perform the following algorithm to identify the insertion direction.

Phase 1: Identification of the insertion direction of instruments for each clear image I

- **Step 1:** Calculate the 2-D line filter using the Hessian Matrix [19, 20], which is often used to detect curvilinear structures. Suppose the Hessian Matrix of a pixel X of the 2-D image $I(X)$ (where $X = (x; y)$) is given by $\nabla^2 I(X)$. Let the eigenvalues of $\nabla^2 I(X)$ be $\lambda_1(X)$ and $\lambda_2(X)$ (where $|\lambda_1(X)| > |\lambda_2(X)|$).
- **Step 2:** Generate a binary image $I_B(X')$ and initialize all the pixel values with zero. For any pixel X' (where $X' = (x; y)$) in the binary image I_B , check the corresponding eigenvalue

$\lambda_1(X)$ of pixel X (where $X = (x; y)$) in the original image $I(X)$. If the absolute value of $\lambda_1(X)$ is larger than a predefined threshold value Th_λ , we treat the pixel X' as an edge pixel and set the pixel value to one [17, 19, 20].

- **Step 3:** We apply a hierarchical clustering approach to group all the edge pixels in the binary image I_B into clusters. Next, we consider only clusters with more than MIN_NUM number of pixels in the cluster. For such a cluster, we extract the skeleton of the cluster using the distance transformation-based skeletonization technique [16]. Then we use the polynomial curve fitting method [16] to fit the skeleton pixels and derive the curvature for each pixel based on the coefficients of the polynomial. If the average curvature of the skeleton pixels is below MAX_CURVE , the cluster can be approximated as a linear line and it is a possible candidate for the boundary of the tubular-shape cable. We name the cluster approximated as a linear line as C_{Linear} . Note that we may have several C_{Linear} clusters. In our implementation, we set MIN_NUM at 30 and MAX_CURVE at 0.05. These thresholds are obtained from experiments and are effective for handling different colonoscopy videos.

- **Step 4:** This step finally determines the insertion direction of the instrument in this image. For each triangular $Area_i$ ($1 \leq i \leq 8$) in Fig. 3(b), we identify $C_{Linear}^{Area_i}$ ---the C_{Linear} cluster with more than 90% of its edge pixels in $Area_i$. We select $C_{MaxLinear}^{Area_i}$ cluster---the cluster with the largest

number of edge pixels among all the $C_{Linear}^{Area_i}$ clusters. If the orientation of the $C_{MaxLinear}^{Area_i}$ cluster and insertion direction i is less than 22.5° ($90^\circ/4$), we claim that direction i is a possible insertion direction for this image.

If there are multiple insertion direction candidates in one image, for example i_1 , with the corresponding cluster $C_{MaxLinear}^{Area_i1}$ is in the triangular area i_1 , and i_2 with the corresponding cluster $C_{MaxLinear}^{Area_i2}$ is in the triangular area i_2 , we choose i_1 as the final insertion direction if the number of edge pixels in the corresponding cluster $C_{MaxLinear}^{Area_i1}$ is greater than that in $C_{MaxLinear}^{Area_i2}$. If we cannot find linear line shape clusters with the orientation close to any of the eight insertion directions, we classify the image as not containing an instrument.

Phase 2: Identification of insertion direction of instruments for the entire video:

- **Step 1:** For each insertion direction i (where $1 \leq i \leq 8$), calculate a value Di , where $1 \leq i \leq 8$. Di refers to the number of images determined as insertion direction i over the total number of images in this video.
- **Step 2:** Compare the eight value Di , (where $1 \leq i \leq 8$) and select j , where $j = \arg \max_{1 \leq i \leq 8} (Di)$, as the final insertion

direction for this video.

Fig. 3(d) and Fig. 3(e) show two example images to illustrate the detection of insertion direction in an image. Fig. 3(d) is the original color image. Fig. 3(e) is the binary image after edge enhancement and clustering. In Fig. 3(e), there is a cluster with a small curvature in the triangle area 6, which is surrounded by a triangular. The angle difference between the orientation of this cluster and the insertion direction 6 is less than 22.5° , which means that direction 6 is the insertion direction of the instrument in this image.

C. Region Filtering

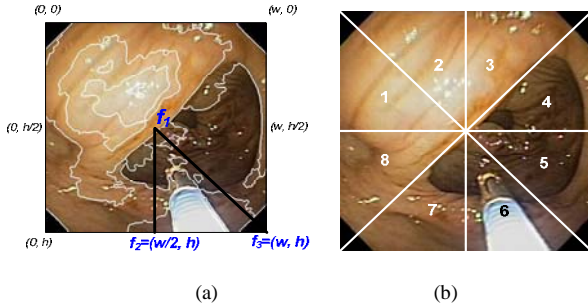


Fig. 4: Triangular filters superimposed on a segmented image: (a) The triangular filter in area 6, (b) Eight triangular filters

This step removes regions outside the triangular area of the corresponding insertion direction detected in the previous step. We do this because the other regions are irrelevant to the cable of the instrument. Recall that in the image preprocessing step, an image is enhanced and segmented using JSEG. Even with our careful selection of important parameters for JSEG, a segmented image still consists of roughly 30 regions on average and over 50 regions in extreme cases. These cases are caused by (1) various degrees of light reflection from the colon wall and (2) complex colon structure in some parts of the colon. Obviously, not all detected regions are part of the cable and therefore should be excluded. The following description of the filtering algorithm assumes that the detected instrument insertion direction is “Direction 6”. By examining a set of segmented cable images from several colonoscopy videos, we find that all the centroids of the desired regions (cable regions) fall in the triangular area shown in Fig. 4(a). To remove irrelevant regions in our colonoscopy videos, we design the triangular filter as follows. Let w and h be the width and the height of the image in pixels, respectively. Given that the top-left corner of the image has the origin coordinate $(0, 0)$, the filter F is a triangle with three vertices: $f_1 = (w/2, h/2)$, $f_2 = (w/2, h)$, and $f_3 = (w, h)$. Let R be a set of regions of a segmented image after the preprocessing step and let $r.centroid$ represent the centroid of region r . The region filtering step identifies the result set C where

$$C = \{r \mid r \in R \wedge r.centroid \in F\}. \quad (1)$$

In order to accommodate instrument detection in cases where the instrument appears in a different position in the field of view (e.g., different type or brand of endoscope), we define eight triangular filters as shown in Fig. 4(b). Based on the triangular filter, we remove all the regions in which the centroid

of the region falls outside the filter.

D. Region Merging

The region merging step identifies the possible instrument regions from the set of candidate regions. Region merging is important since a whole instrument is often over-segmented into several regions. The over-segmentation is not necessarily due to a fault of the image segmentation technique; it may be the effect of very bright light reflections in the image. With the intense light, it is quite common that the different parts of the same instrument have different colors.

Our previous method [15] assumes the instrument insertion direction is “Direction 6” and generates all combinations of regions such that each combination contains at least one bottom region (the region that touches the bottom edge of the image). This method does not miss any region combinations that represent the true instrument. However, it also generates many redundant region combinations for the next step. To overcome this problem without sacrificing the effectiveness of the technique, we propose a new region growing algorithm based on texture features as follows.

- **Feature extraction:** We extract four texture features for each region in the image. These features are StandardDev, Smoothness, Uniform, and Entropy [16].
- **Region clustering:** We apply K-means clustering to classify each region in this image into three categories: smooth region, periodic region, and coarse region based on the above texture features.
- **Region growing:** We pick the bottom region as a seed region. Suppose R_{bottom} is the current bottom region we inspect, we check each neighboring region $R_{neighboring}$ of R_{bottom} . Note that all the regions under inspection are in the triangular area determined by the detected instrument insertion direction. If both $R_{neighboring}$ and R_{bottom} belong to the same category (for example, both of them are *smooth regions*), we combine them into one region. We perform this step recursively until no more neighboring regions can be added to form a new region.

We refer to the set of the combined regions after the merging step as Q . It is composed of merged regions used for the next step.

E. Region Matching

This step matches each of the regions in Q with a manually defined template set of the cable regions. The template set represents the different representative shapes of the cable found commonly in our colonoscopy videos. We manually selected representative cable images and extracted the corresponding cable regions. Instead of using Fourier descriptors as in [15], we use moment invariants [21, 22] as our shape features. These features are not sensitive to linear transformation, making it suitable to handle different insertion directions of the instrument.

Let $shape(q)$ return seven moment invariant features of region q . Let $S = \{S_1, S_2, \dots, S_k\}$ be a set of k feature vectors

where $S_i = \text{shape}(i)$ for the template region i . Let $\text{dist}(i, j)$ return the “city-block” distance between feature vectors i and j [16]. Given a similarity threshold d , the region matching step decides whether image I is a cable image or not as follows.

I is a cable image if there exists an s such that $\text{dist}(s, \text{shape}(q)) < d$ where $s \in S \wedge q \in Q$ (2)

In other words, the image is declared as a cable image if the dissimilarity between one of its regions in Q and one of the template regions is less than the threshold. Otherwise, the image is considered a non-cable image. The appropriate value of the threshold d is found to be 0.025 from experiments. The region matching step outputs a 1 for each detected cable image and a 0 otherwise. The “city-block” distance is used for similarity measure since it has been reported to perform slightly better than other distance metrics for shape matching in [23]. Since the number of distinct cable shapes is small (about ten shapes), these shape feature vectors are loaded in memory once and used during the matching process for the entire video.

F. Shot Segmentation

This step utilizes temporal information and domain knowledge to identify operation shots. This step addresses the fact that the appearances of a cable vary in the same operation shot and corrects the errors introduced by steps prior to this step. The shot segmentation step accepts L , a sequence of 0's and 1's from the region matching step, as an input and locates the boundaries of operation shots as follows.

Step 1: This step aims to correct the misclassification results due to the region matching step. A misclassification result is the case where the image without a cable is classified as a cable image by the region matching step. We found that the detected cable image when surrounded by several non-cable images is very likely a misclassification. We use this observation to correct a misclassification. We first explain the algorithm when one frame per second ($t=1$) is used in the image preprocessing step. Let L' be the output sequence of binary numbers with the same length as the input sequence L . Starting from the beginning of the input sequence L , we slide a sliding window W (covering 5 binary numbers at a time) over the input L to find the *correction pattern* [0,0,1,0,0] in L . When such a pattern is found, we correct the misclassification result by changing the middle 1 to 0 in the corresponding position in the output sequence L' . Except this change, the corresponding position in the output L' has the same binary number as that in L . In other words, we have [0,0,0,0,0] in L' when [0,0,1,0,0] is under the current sliding window in L . Next, we slide the window one number to the right and repeat the same process until the end of the input sequence is reached. Note that we elected to use the pattern [0,0,1,0,0] since in our experiments this pattern removed errors better than other patterns with more zeros surrounding the middle one. We generalize the correction pattern for different values of t as a pattern that has $2*t$ of zeros followed by $1*t$ of one followed by $2*t$ of zeros. For instance, when t is 2 (2 frames per second are used), we use the sliding window of size $5*2$ to search for the correction pattern of [0,0,0,0,1,1,0,0,0,0].

Step 2: We scan L' from the beginning to the end. We declare an operation shot when we find a sequence O of consecutive frames in L' with all of the following properties.

- The sequence O starts with a 1 and ends with a 1 followed by at least $8 \times t$ consecutive zeros. In other words, the first frame and the last frame in the sequence O are cable images. A fixed number ($8 \times t$) of consecutive non-cable images following the last frame of the sequence O captures the withdrawal of the instrument quite well. We have experimented with larger or smaller numbers of trailing zeros. However, we found that the effectiveness of shot segmentation degrades with more or less trailing zeros.
- The number sequence O has more 1's than 0's. That is, the sequence O has more cable images than non-cable images. This rule is developed based on the observation that an actual operation shot typically has more frames with a cable present than without one.

The sequence O lasts at least 4 seconds. Based on our experience, operation shots typically last more than 4 seconds. Only random biopsies (e.g., for tissue studies in patients with diarrhea or for dysplasia screening in patients with ulcerative colitis) may result in operation shots shorter than 4 seconds; sometimes these random, blind biopsies are even very difficult to be observed by the human eye. Hence, for the studies presented in this article we do not consider an operation shot shorter than 4 seconds.

IV. PERFORMANCE STUDY

This section presents experimental results illustrating the effectiveness of our proposed techniques on three test data sets: **(1) Video Set I:** for identification of insertion direction of instruments. The accuracy of subsequent steps, such as region filtering and region merging, depends on this step. We used videos generated from multiple endoscope models used for different endoscopic procedures, including colonoscopy and esophago-gastro-duodenoscopy (EGD), to determine the effectiveness of our technique. We use five colonoscopy videos whose insertion direction is “Direction 6” and four EGD videos whose direction is “Direction 8”. **(2) Image Set:** consists of about 1,000 cable and non-cable images extracted from six colonoscopy videos. We used this set to evaluate the effectiveness of the region filtering, region merging, and region matching steps of the operation shot detection technique. Details are listed in Table 1. **(3) Video Set II:** consists of twenty five colonoscopy videos with and without operation shots. This test set was used to evaluate the effectiveness of the operation shot detection technique. The total number of operation shots in these videos is 117. Among them, there are

TABLE I
CHARACTERISTICS OF THE IMAGE SET

| Video ID | 010 | 015 | 017 | 019 | 024 | 044 | Total |
|-----------------|-----|-----|-----|-----|-----|-----|-------|
| Cable image | 93 | 11 | 25 | 14 | 123 | 163 | 429 |
| Non-cable image | 149 | 92 | 60 | 21 | 142 | 194 | 658 |

96 operation shots contain biopsy forceps, 20 shots have snare, and 1 shot with balloon. The average length of an operation shot in all test videos with diagnostic and therapeutic operations is about 22 seconds.

A. Determining Important Parameters for the Proposed Approach

The first step for operation shot detection is “Image Preprocessing”. In this step, we obtain the reduced colonoscopy video by extracting t frames per second from the input colonoscopy videos. In the experiments, we chose t equal to one, which implies that the maximum temporal distance between the actual boundary and the detected boundary due to temporal sampling is 1 second. This temporal distance is considered very small compared to the average length of an operation shot (22 seconds). This distance can be made smaller with a higher value of t , however, with the expense of a significant increase in the analysis time for operation shot detection. Also in this step, we propose a non-linear filter to remove small over-bright white areas. In this method, we use a $w \times w$ sliding window to scan the entire image to identify these areas. The value of w is mainly determined by the resolution of the colon image because the size of the white area is proportional to the size of the image. In our experiments, the resolution of our colon image is 390×370 and we set the value w at 15. Recall that our algorithm to identify insertion direction extracts the strong curvilinear structure in the colon image and checks the orientation of the tubular shape object to determine the final direction. One important parameter of this method is the predefined threshold Th_2 used in the second step of this algorithm. We set this value relatively high in order to remove most false positive images (images that do not have any information about instrument insertion direction, but detected as images that contain the insertion direction). At the same time, because of the high threshold, our method generates more false negative images (images with a cable detected as images without cable insertion direction information). However, this does not affect the final results since our detection algorithm selects the insertion direction i with the maximal Di (where $1 \leq i \leq 8$) value as our final insertion direction. We will discuss this issue in detail in the next section.

B. Effectiveness of Cable Detection

There are four important steps in detecting whether an image is a cable image or not. They are identification of instrument insertion direction, region filtering, region merging, and region matching steps. We quantify the effectiveness of each step as follows.

1) Effectiveness of Identification of Instrument Insertion Direction

Recall that for each input image, our algorithm either assigns an insertion direction or skips the image. For each video, we use “ Di ” to indicate the number of images with insertion direction i . In the final stage of our method, we compare the eight “ Di ” (where $1 \leq i \leq 8$) values and select the “ i ” with the largest “ Di ” value as the final insertion direction. Based on this

method, values with “ $D6$ ” for the first five colonoscopy videos and values with “ $D8$ ” for the last four EGD videos are selected. This indicates that the insertion directions for the first five videos and last four videos are “*Direction 6*” and “*Direction 8*”, respectively. Hence, our algorithm gives the correct results for all tested videos.

2) Effectiveness of Region Filtering

The purpose of the region filtering step is to remove irrelevant regions from further consideration. We use the image set for performance evaluation. For each image in the image set, we obtain the total number of original regions after image segmentation and the number of *result regions*—regions left after region filtering. We compute the ratio of the number of the result regions to the number of original regions gathered from selected images of each video. For the cable images, only 18% of the original regions remain. For the non-cable images, only 13% of the original regions remain. More regions are left in the cable images due to the presence of the cable. Although 82% of irrelevant regions are removed, no parts of the actual cable region are removed.

3) Effectiveness of Region Merging

To quantify the effectiveness of the region merging step, we evaluated the effectiveness of cable detection with and without region merging. Out of the 429 cable images in the image set, we manually identified all cable images whose cable is fragmented into more than one region by JSEG. We performed region matching with and without prior region merging on this sub-set of cable images. Region matching with region merging correctly identifies 96% of the images in the set as cable images. Without region merging, region matching only correctly identifies 69% of the images in the set as cable images. Therefore, region merging results in 27% improvement in accuracy for cable detection. Note that some fragmented cables can be detected even without region merging because the fragment of the cable happens to have a shape similar to that of the entire cable.

4) Effectiveness of Region Matching

TABLE 2: ACCURACY OF CABLE DETECTION

| Video ID | Cable Images | Non-Cable Images |
|----------------|--------------|------------------|
| 010 | 0.96 | 0.89 |
| 015 | 0.75 | 0.92 |
| 017 | 0.99 | 0.92 |
| 019 | 1.00 | 0.95 |
| 024 | 0.90 | 0.96 |
| 044 | 0.94 | 0.95 |
| Average | 0.92 | 0.93 |

The region matching step is the last step for cable detection. The effectiveness of this step is demonstrated via the effectiveness of the entire cable detection. Given an actual cable image, the cable detection algorithm should indicate that the image is a cable image with a high accuracy. Second, given a non-cable image, the cable detection algorithm should determine that the image is a non-cable image with a high accuracy. Table 2 shows the results of cable detection on the image set. The average accuracy of 92% for cable images and

93% for non-cable images in Table 2 are attributed by the effectiveness of (1) the region filtering step that removes irrelevant regions from the cable images; (2) the region merging step that combines fragmented regions that should have been detected as one region; and (3) the use of moment invariants as shape features for matching the candidate region with the template regions. Nevertheless, the cable detection algorithm still has 8% inaccuracy and we rely on the shot segmentation step to correct these small errors due to the following reasons: (1) The JSEG image segmentation algorithm sometimes merges parts of the colon wall and the cable together, which results in a shape different from the template cable shapes; (2) In a non-cable image, the shapes of one or more regions of the colon and the cable may be similar by chance. This error is inevitable associated with our detection method. In addition, the region merging step introduces the possibility that a combination of colon regions is similar to one of the template cable regions. However, this case happens rarely.

C. Effectiveness of Operation Shot Detection

To evaluate the effectiveness of the entire operation shot detection algorithm, we measured the number of false operation shots, the number of missed operation shots, the true positive fraction, the false positive fraction, and the boundary precision and recall. False operation shots are software detected shots that are not actual operation shots determined manually. A missed operation shot is an actual operation shot for which the software failed to detect both boundaries. Note that if one of the two boundaries of an operation shot is incorrect, the detected operation shot still captures part of the actual operation. In such cases we did not treat the detected operation shot as a false or a missed operation shot, but we quantified it using the following metrics. The true Positive Fraction (TPF) is the ratio of the total number of correctly detected images as part of actual operation shots (true positives) to the total number of images of actual operation shots. High TPF is desirable. The False Positive Fraction is defined as the ratio of incorrectly detected images as part of operation shots (false positives) to the total number of images of actual operation shots. Low FPF indicates that a small fraction of a detected operation shot is not part of an actual operation shot. Note that our definition of FPF is different from the traditional FPF that uses the ratio of false positives to real negatives. We chose a different definition for FPF because the number of real negatives in general is much larger than the number of false positives our algorithm produces. Using the traditional definition, we have around 0.006 FPF on our test data set. To also quantify the percentage of correctly detected boundaries, we use boundary precision and recall. Boundary precision is the ratio of the number of correctly detected boundaries to the total number of detected boundaries. Boundary recall is the ratio of the number of correctly detected boundaries to the number of actual boundaries determined manually by humans. High boundary precision and recall are desirable.

We applied our method to 25 colonoscopy videos (compared

to 20 videos used in our previous paper [15]). Only seven false shots are detected and the number of missed shot is zero. In our opinion this number is a very small number given that any pair of frames in the videos can form a false operation shot. Averages of true positive and false positive fractions are 94% and 10%, respectively. The majority of the videos have a perfect true positive fraction of 1.0. For some videos, the detected operation shots are shorter than the actual operation shots by a couple of frames in the beginning or the end of an actual operation shot. The false positive fraction is due to the case that some detected shots are false; in addition, some detected shots are slightly longer than the actual operation shots. A boundary recall of 97% is very high. Only 3% of the actual boundaries are missed by the algorithm due to the following reasons. First, intense brightness causes JSEG to combine parts of the cable and colon wall. Second, in rare cases only the head of the biopsy forceps (without the cable) is presented in the video during the starting or the ending of an operation. The head of the forceps remains open for several seconds. Since the shape of the head of the open forceps is different from the shape of the cable, we cannot detect the correct boundary in this case. Note that the cable detection algorithm still declares a forceps

TABLE 3: EFFECTIVENESS OF OPERATION SHOT DETECTION

| | |
|---------------------------------|------|
| Number of false shots | 7 |
| Number of missed shots | 0 |
| Average true positive fraction | 0.94 |
| Average false positive fraction | 0.10 |
| Average boundary precision | 0.88 |
| Average boundary recall | 0.97 |

| <i>C1</i> | <i>C2</i> | <i>C3</i> | <i>C4</i> | <i>C5</i> | <i>C6</i> | <i>C7</i> | <i>C8</i> |
|----------------|------------|------------|------------|-------------|-------------|-------------|-------------|
| 002 | 2 | 4 | 2 | 1.00 | 0.38 | 0.50 | 1.00 |
| 009 | 2 | 4 | 2 | 1.00 | 0.45 | 0.50 | 1.00 |
| 010 | 12 | 14 | 11 | 0.71 | 0.04 | 0.79 | 0.92 |
| 012 | 12 | 12 | 12 | 1.00 | 0.00 | 1.00 | 1.00 |
| 014 | 2 | 2 | 2 | 1.00 | 0.00 | 1.00 | 1.00 |
| 024 | 24 | 26 | 24 | 1.00 | 0.11 | 0.92 | 1.00 |
| 044 | 18 | 20 | 17 | 0.72 | 0.08 | 0.85 | 0.94 |
| 047 | 18 | 19 | 16 | 0.86 | 0.17 | 0.84 | 0.89 |
| 053 | 14 | 15 | 13 | 0.82 | 0.09 | 0.80 | 0.93 |
| 097 | 6 | 8 | 6 | 1.00 | 0.31 | 0.75 | 1.00 |
| 102 | 10 | 10 | 10 | 1.00 | 0.00 | 1.00 | 1.00 |
| 111 | 4 | 4 | 4 | 1.00 | 0.00 | 1.00 | 1.00 |
| 114 | 12 | 14 | 12 | 1.00 | 0.03 | 0.86 | 1.00 |
| 116 | 16 | 16 | 16 | 1.00 | 0.00 | 1.00 | 1.00 |
| 133 | 8 | 8 | 8 | 1.00 | 0.00 | 1.00 | 1.00 |
| 134 | 10 | 10 | 8 | 0.85 | 0.08 | 0.80 | 0.80 |
| 148 | 6 | 7 | 6 | 1.00 | 0.10 | 0.86 | 1.00 |
| 156 | 4 | 6 | 4 | 1.00 | 0.16 | 0.67 | 1.00 |
| 165 | 4 | 6 | 4 | 1.00 | 0.14 | 0.67 | 1.00 |
| 168 | 2 | 2 | 2 | 1.00 | 0.00 | 1.00 | 1.00 |
| 174 | 0 | 0 | 0 | - | - | - | - |
| 183 | 0 | 0 | 0 | - | - | - | - |
| 186 | 20 | 22 | 20 | 1.00 | 0.17 | 0.91 | 1.00 |
| 192 | 20 | 20 | 19 | 0.82 | 0.00 | 0.95 | 0.95 |
| 202 | 8 | 8 | 8 | 1.00 | 0.00 | 1.00 | 1.00 |
| Total | 234 | 257 | 226 | - | - | - | - |
| Average | - | - | - | 0.94 | 0.10 | 0.88 | 0.97 |

Name of each column: *C1* (Video ID), *C2* (#Actual Boundaries), *C3* (#Detected Boundaries), *C4* (#Correctly Detected Boundaries), *C5* (True Positive Fraction), *C6* (False Positive Fraction), *C7* (Boundary Precision), *C8* (Boundary Recall)

head a cable image if the head of the forceps is closed since the closed head shape is very similar to the cable shape. The boundary precision is lower compared with the boundary recall. Our shot detection method introduces $257-234-7=16$ false boundaries, of which 14 boundaries are due to 7 false shots.

All the experiments were conducted on a PC with 3.40 GHz Pentium(R) 4 and 1GB of RAM. The processing time for each video frame once the insertion direction has been identified is about 7 seconds on average, of which 6 seconds are spent by JSEG to perform region segmentation. Better performance can be achieved with the more efficient implementation of JSEG.

V. CONCLUSIONS

We have introduced a new operation shot detection technique. Our first set of experiments on real colonoscopy videos demonstrates that the operation shot detection technique does not miss any meaningful operation shots and produces only a very small number of false video segments that do not correspond to actual diagnostic or therapeutic operations. The algorithm that we present is procedure, endoscope brand and operation equipment independent. Indeed, it does not matter what type of endoscope, what brand of equipment or what type of procedure or surgery are being used or done as our method is simply based on recognizing a specific shape entering the image from the periphery and directed towards the center of the image. That shape in most cases will be that of a conical cylinder similar to the shape of our cable in colonoscopy videos. Thus we have created a software module that can be applied to EGD, bronchoscopy, arthroscopy, laparoscopy, and other forms of endoscopic procedures.

As our future work, we will investigate a better shot segmentation step to further improve precision and recall of the software-detected shot boundaries. The current approach is based on observations on our video sets and some domain knowledge. Our future investigation is to use statistical hypothesis testing to decide whether to include an image as part of an operation shot or not based on derived spatial features and temporal features of the image and its preceding images. We will also evaluate the operation shot detection technique on other types of endoscopic procedures such as upper gastrointestinal endoscopy, enteroscopy, cystoscopy, and laparoscopy.

REFERENCES

- [1] S. Kuwada, "Colorectal cancer 2000," *Postgraduate Medicine*, vol. 107, pp. 96-107, May 2000.
- [2] S. J. Phee and W. S. Ng, "Automatic of colonoscopy: visual control aspects," in *IEEE Engineering in Medicine and Biology Magazine*, vol. 17, 1998, pp. 81-88.
- [3] Y. M. Lim and J. H. Lee, "A self-propelling endoscopic system," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Maui, HI, USA, 2001, pp. 1117-1122.
- [4] N. Esgiar, R. N. G. Naguib, B. S. Sharif, M. K. Bennett, and A. Murray, "Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonic mucosa," *IEEE Transactions on Information Technology in Biomedicine*, vol. 2, pp. 197-203, September 1998.
- [5] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features," *IEEE Transactions on Information Technology in Biomedicine*, vol. 7, pp. 141-152, September 2003.
- [6] S. Lakare, D. Chen, L. Li, A. Kaufman, and J. Liang, "Electronic colon cleansing using segmentation rays for virtual colonoscopy," in *Proc. of SPIE 2002 Symposium on Medical Imaging*, San Diego, CA, USA, 2002, pp. 412-418.
- [7] S. B. Gokturk, C. Tomasi, B. Acar, C. F. Beaulieu, D. S. Paik, R. B. J. Jr., J. Lee, and S. Napel, "A statistical 3-D pattern processing method for computer-aided detection of polyps in CT colonography," *IEEE Transactions on Medical Imaging*, vol. 20, pp. 1251-1260, December 2001.
- [8] H. Zhang, "Content-Based Video Browsing and Retrieval," in *Handbook of Multimedia Computing*, vol. 5, B. Furht, Ed.: CRC Press, 1998, pp. 255-280.
- [9] Y. Rui, T. S. Huang, and S. Mehrotra, "Exploring video structure beyond the shots," in *Proc. of IEEE International Conference on Multimedia Computing and Systems*, Austin, TX, USA, 1998, pp. 237-240.
- [10] J. Mao and K.-K. Ma, "Semantic spatial-temporal segmentation for extracting video objects," in *Proc. of IEEE International Conference on Multimedia Computing and Systems*, Florence, Italy, 1999, pp. 48-55.
- [11] X. Tang, X. Gao, J. Liu, and H. Zhang, "A spatial-temporal approach for video caption detection and recognition," *IEEE Transactions on Neural Networks*, vol. 13, pp. 961-971, July 2002.
- [12] R. Green, "Spatial and temporal segmentation of human from monocular video images," in *Proc. of Image and Vision Computing*, Palmerston North, New Zealand, 2003, pp. 163-169.
- [13] L. Xie, S. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer videos with Hidden Markov Models," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, 2002, pp. 767-775.
- [14] L. Duan, M. Xu, T. Chua, Q. Tian, and C. Xu, "A mid-Level representation framework for semantic sports video analysis," in *Proc. of ACM Multimedia*, Berkeley, CA, USA, 2003, pp. 33-44.
- [15] Y. Cao, D. Li, W. Tavanapong, J. Oh, J. Wong, and P. C. de Groen, "Parsing and browsing tools for colonoscopy videos," in *Proc. of ACM Multimedia*, New York, NY, USA, 2004, pp. 844-851.
- [16] M. Sonka and V. Hlavac, *Image Processing, Analysis, and Machine Vision*. New York City, NY, USA: Thomson-Engineering, 2000.
- [17] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 800-810, August 2001.
- [18] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: image segmentation using expectation-maximization and its application to image querying," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1026-1038, August 2002.
- [19] D. H. Eberly, *Ridges in image and data analysis*. Norwell, MA: Kluwer Academic Publishers, 1996.
- [20] T. Lindeberg, "Edge detection and ridge detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, pp. 117--156, November 1998.
- [21] M.-K. Hu, "Visual pattern recognition by moment invariants," *IEEE Transactions on Information Theory*, vol. 8, pp. 179-187, 1962.
- [22] N. Sebe and M. S. Lew, "Robust Shape Matching," in *Proc. of IEEE International Conference on Image and Video Retrieval*, London, UK, 2002, pp. 17-28.
- [23] J. P. Eakins, K. J. Riley, and J. D. Edwards, "Shape feature matching for trademark image retrieval," in *IEEE International Conference on Image and Video Retrieval*, Urbana-Champaign, IL, USA, 2003, pp. 28-38.