MINING COLONOSCOPY VIDEOS TO MEASURE QUALITY OF COLONOSCOPIC PROCEDURES

Danyu Liu\textsuperscript{a}, Yu Cao\textsuperscript{a}, Wallapak Tavanapong\textsuperscript{a}, Johnny Wong\textsuperscript{a}, JungHwan Oh\textsuperscript{b}, and Piet C. de Groene\textsuperscript{c}

\textsuperscript{a}Department of Computer Science, Iowa State University, Ames, IA 50011, USA
\textsuperscript{b}Department of Computer Science & Engineering, University of North Texas, Denton, TX 76203, USA, \textsuperscript{c}Mayo Clinic College of Medicine, Rochester, MN, USA

impact.isu@cs.iastate.edu

ABSTRACT
Colonoscopy is an endoscopic technique that allows a physician to inspect the inside of the human colon. Colonoscopy is the accepted screening method for detection of colorectal cancer or its precursor lesions, colorectal polyps. Indeed, colonoscopy has contributed to a decline in the number of colorectal cancer related deaths. However, not all cancers or large polyps are detected at the time of colonoscopy, and studies of why this occurs are needed. Currently, there is no objective way to measure in detail what exactly is achieved during the procedure (i.e., quality of the colonoscopic procedure). In this paper, we present new algorithms that analyze a video file created during colonoscopy and derive quality measurements of how the colon mucosa is inspected. The proposed algorithms are unique applications of existing data mining techniques: decision tree and support vector machine classifiers applied to videos from medical domain. The algorithms are to be integrated into a novel system aimed at automatic analysis for quality measures of colonoscopy.

1. Introduction

Colorectal cancer is the second leading cause of cancer-related deaths behind lung cancer in the United States [1]. Colonoscopy is currently the preferred screening modality for prevention of colorectal cancer. A colonoscopic procedure consists of two phases: an \textit{insertion phase} and a \textit{withdrawal phase}. During the insertion phase, a flexible endoscope (a flexible tube with a tiny video camera at the tip) is advanced under direct vision via the anus into the rectum and then gradually into the most proximal part of the colon (signified by the appearance of the appendiceal orifice or the terminal ileum). In the withdrawal phase, the endoscope is gradually withdrawn. Careful mucosa inspection and diagnostic or therapeutic interventions such as biopsy, polyp removal, etc., are performed in the withdrawal phase. The video camera generates a sequence of images (frames) of the internal mucosa of the colon. These images are displayed on a monitor for real-time manual analysis by the endoscopist. In current practice, images of the entire procedure are not routinely captured for post-procedure review or analysis.

Colonoscopy is performed over 10 million times per year [2]. However, in current practice there is no objective way to measure in detail what exactly is achieved during the procedure although a number of indirect markers of quality have been proposed. These include duration of the withdrawal phase and average number of polyps detected per screening colonoscopy. Thoroughness of inspection of the colon mucosa, i.e., by looking off-axial to the mucosa or behind mucosal folds, currently cannot be measured.

As part of a novel quality measurement system for colonoscopy, we have recently developed (i) a system to automatically capture all images from a colonoscopic procedure into a colonoscopy video file and upload the file to an analysis server; no identifiable patient information is captured; (ii) image analysis techniques to identify biopsy and therapeutic operations from colonoscopy videos [3]; (iii) image analysis techniques that output objective measures of quality of colonoscopic procedures [4].

In this paper, we introduce new algorithms to obtain estimates of view mode from a colonoscopy video. The view mode is the distance of the camera at the tip of the endoscope to the most distant colon wall. We classify the view mode into two types: \textit{global inspection} (more distant examination in which more than one side of the colon wall is seen) in Figure 1(a-c) and \textit{close inspection} (close examination of the colon mucosa) in Figure 1(d-f). Both close and global inspections should be present in a good colon examination. Since most commonly used endoscopes are not able to provide measurements of the view mode, we can only obtain these measures through analysis of images in a colonoscopy video.

Estimating the view mode is a challenging problem due to the complexity and individual variation of the colon structure, colonic motility, changes in mucosal brightness related to light source to lens position, presence of out-of-focus (blurry) images, and presence of other objects in the image such as instruments during biopsy and therapeutic operations, stools, or water injected...
to clean the colon. Due to the above complexities, determining the view mode based on image analysis is not as trivial as it seems. Simply training a well known classifier with common global features such as color, texture, and generic shape descriptors or with salient features detected using a recent salient feature detection technique for generic objects \cite{5} is not able to address this problem satisfactorily. Hence, we propose a new technique for estimating the view mode.

The remainder of this paper is organized as follows. Section 2 provides background and related work. In Sections 3, we present the proposed techniques in detail along with the experimental results. Finally, we conclude the paper in Section 4.

2. Background and Related Work

2.1. Background and Challenges

To estimate the view mode given an image, we rely on the presence of the lumen in the image. In our previous work \cite{4}, we defined a lumen view as a clear (non-blurred) frame in which the distant colon lumen is seen. That means that the line of view is along the longitudinal axis of the colon proximal to the endoscope tip. If the distant lumen is central in the image, the view is axial; if the distant lumen is in the periphery of the image, the view is off-axial. A clear frame without the distant colon lumen is called a wall view. A wall view most often occurs as a result of a close inspection of the lateral colon wall whereas the lumen view indicates a more global inspection where more than one side of the colonic wall is seen. Both local and global inspections are important. Once the correct classification of the images is made, we can derive different quality metrics such as (i) the ratio of close inspection to global inspection \cite{4}, (ii) the duration of a sequence of wall views or a sequence of lumen views, or (iii) the interleaving pattern between close inspection and global inspection.

To illustrate the challenges in estimating the view mode from each image, we show examples of lumen views in Figure 1(a-c) and wall views in Figure 1(d-f). For each image in Figure 1(a-c), we superimposed an arrow originating from the center of the image to indicate the view direction. This representation of the view direction is based on our endoscopic experience. The arrow approximately points to the lumen area. In Figure 1(a-b), the lumen is very dark whereas in Figure 1(c), the lumen is not very dark, but is relatively darker compared to the other part of the colon wall. Figure 1(a) shows that the lumen is seen slightly on the lower right side whereas in Figure 1(b) the lumen is seen in the bottom and is partially blocked by a tube-like instrument. For the wall view examples, Figure 1(d) shows the wall view in which the very dark area similar to the appearance of the lumen is on the top-left corner. Figure 1(e) depicts another wall view in which the relatively dark area is seen in the bottom of the image. Figure 1(f) shows an interesting wall view with protruding lesions and the dark shadow behind some lesions. The head of a biopsy forceps (silver color) is also seen. These images partially illustrate the challenges that we need to overcome when measuring view mode and direction.

2.2. Related Work

The most related research efforts are in the area of micro robotic endoscopy \cite{6, 7}. They focus on the following problem: given an endoscopic image with the lumen, identify the lumen boundary. The work in \cite{6, 7} do not discuss how to determine whether the lumen is seen in the image or not. Unlike the aforementioned techniques, our previous work \cite{4} determines whether the lumen is seen in the image or not. The technique employs the relative darkness of the lumen coupled with the following facts. First, more than one bilateral convex colon wall is seen around the colon lumen. Second, the intensity difference between consecutive colon walls is small. We refer to this technique as “Grayscale Shape-based View Mode Classification (GSVM)”

Existing techniques including our previous GSVM have the following drawbacks. First, they do not utilize chrominance information of the pixels; we will show in this paper that the chrominance information is very useful. Second, the adaptive threshold methods like APT-Iris on pixel intensity alone may misclassify many wall view images as lumen view images. For instance, Figure 1(e) may be misclassified as a lumen view due to the shadow of the colon fold. With shape features, our GSVM can address some drawbacks of the adaptive threshold methods. However, GSVM requires the appropriate manual setting of several thresholds.

3. Proposed View Mode Detection
First, lumen pixel classification uses a decision-tree classifier to classify each pixel in an image into either a lumen pixel or a wall pixel. Lumen pixels are in the image area where the distant lumen is seen. See the very dark areas in Figure 1(a-b) as examples. A wall pixel is not part of the lumen seen in the image. The lumen pixel classification outputs an intermediate image called “red-green image.” The pixel of this image is either green or red. The green pixel indicates that the corresponding pixel in the original image is classified as the lumen pixel. The red pixel corresponds to the pixel in the original image classified as the wall pixel. Second, feature extraction extracts seven image features from the red-green image as well as the corresponding original image. These features are carefully designed. Third, image classification uses a decision-tree classifier to determine whether the image is a wall view or lumen view. We also investigated the application of SVM for the image classification. Last, the view mode is used to compute novel objective quality measurements. In this paper, we focus our discussion on the lumen pixel classification, the image feature extraction and classification for view model.

3.1. Lumen Pixel Classification

Pixel-based classification methods have gained popularity as pre-processing steps in many image processing applications, such as face recognition and skin detection. To classify each pixel into a lumen pixel or a wall pixel, we need to select (i) a proper color representation (i.e., a color space and color planes) of pixels and (ii) an effective classifier.

3.1.1. Selection of Color Spaces and Color Planes.

To determine which color space is appropriate, we consider seven major color spaces: RGB, normalized rgb, HSV, YCrCb, CIE La*ub*, CIE xyY, and YIQ.

The distant colon lumen is seen relatively darker compared to the proximal colon wall. In fact, both lumen view and wall view represent views of the colonic mucosa, and the real color of the lumen is similar to the color of the wall; however, it is seen as very dark beige or black due to the long distance from the light source to the distant wall, the dispersion of the light bundle and the small of light reflected back to the camera of the endoscope. Using only chrominance components as typically used for skin pixel detection may not give the best results. To test our hypothesis, we investigated the discriminating power of chrominance components only for lumen pixel classification. Preliminary experiments show that techniques using chrominance components alone or a simple threshold do not possess the discriminating power to distinguish lumen pixels from wall pixels.

We then investigated the possibility of using RGB color space since it combines both luminance and chrominance components. Based these investigations, we have the following hypotheses. First, the effectiveness of lumen pixel classification using RGB color space may outperform that using grayscale space. Second, the effectiveness of lumen pixel classification using RGB may be comparable to those of popular color spaces used in skin detection. We report the results of our investigation in Section 3.1.3.

3.1.2. Selection of Classifiers and Training/Testing Data Sets.

Several machine learning techniques are available in the literature. We chose the decision tree training and classifying model for two reasons. First, recent papers [8] concluded that a decision-tree based classifier is one of the two best performers for skin pixel classification. Hence, the decision tree classifier may also lend itself well for lumen pixel classification. Second, we can examine the rule set obtained from the decision tree learning to gain more insights on the ranges of values of color planes that impact the classification. We use C4.5 [9] as the classifier. Since C4.5 is a supervised learning algorithm, labeled training samples are required.

Because there is no established data set of labeled lumen and non-lumen pixels, we built our own data sets as follows. We chose sixty non-blurred images. Out of these images, we manually classified each of these images as a lumen view or a wall view. Then, we built the data sets for lumen pixels and wall pixels as follows.

To create the data set for lumen pixels, we used a semi-automatic method to mark the lumen area in a lumen view. This is to avoid human errors of including small bright spots due to strong light reflection in the data set for lumen pixels. The semi-automatic method works as follows. We manually selected a seed point in a lumen region and indicated a dissimilarity threshold as input to the existing region growing software. The software takes the seed point as the initial region and iteratively includes in the region a new pixel that satisfies two constraints. The first constraint is that the pixel must be a neighbor of one of the pixels already included in the region. Second, the
color dissimilarity of this pixel and one of the pixels in the region must be within the given threshold. We generated a binary mask image to indicate the lumen region identified by the software. We selected a rather small dissimilarity threshold to prevent the region growing from expanding outside of the real lumen area. Next, we manually checked the correctness of the mask image. Figure 2(a) shows an image with the lumen. Figure 2(b) shows the white area corresponding to the pixels marked as lumen pixels by the region growing software. A small black spot inside the white area in Figure 2(b) corresponds to a bright spot in Figure 2(a).

We generated the data set for wall pixels as follows. We manually marked regions that are not part of the lumen as wall pixels. Figure 2(d, f) shows the mask image with the regions in white corresponding to some wall regions in the original image. We also marked the pixels that have similar appearance to the dark lumen, but actually not part of the lumen as wall pixels. The total number of lumen pixels is 251,791 and the total number of wall pixels is 1,619,503. The number of wall pixels is larger, which is typically the case for colonoscopy videos.

3.1.3. Effectiveness of Lumen Pixel Classification under Various Color Spaces.

We used the training data sets of lumen and wall pixels discussed in Section 3.1.2 to build a lumen pixel classifier. For each color space, we used the values of the three basic color components (e.g., R, G, and B for RGB color space) as input to the C4.5 classifier to build a decision tree model. For grayscale spaces, we only use the luminance component of the color spaces as input to C4.5.

We compared the effectiveness of the classifiers built from training data sets using RGB, normalized rgb, HSV, YIQ, YCrCb, CIE La*b*, CIE xyY, and the grayscale spaces: we used the luminance component V from HSV, and the luminance components Y from YIQ, Y from YCbCr, L from CIE La*b*, and Y from CIE xyY. We performed ten-fold cross validation on these data sets. That is, we divided images into ten subsets with approximately same size and similar distribution between the two classes (lumen pixel and wall pixels). Each time we trained the decision tree classifier using one of the ten subsets and used the other nine subsets for testing.

Four different performance metrics are used for evaluation. Precision is the proportion of all detected lumen pixels that are real lumen pixels. Recall is the proportion of all lumen pixels identified correctly by the classifier. Specificity is the proportion of all wall pixels identified correctly by the classifier. Accuracy is the proportion of all pixels (both lumen and wall) identified correctly by the classifier. Table 1(a) illustrates that the decision tree classifier trained using pixel values in RGB color space outperforms the decision tree classifiers trained using the grayscale space. Table 1(b) shows that RGB performs comparably with HSV, YCbCr, CIE La*b* and CIE xyY. The worst performance is given by the classifier using normalized rgb. This result markedly differs from the reported results for skin detection in which normalized rgb gives good performance [13]. Based on these results, we conclude the following: (1) Dropping the chrominance components degrades the ability to separate lumen pixels and wall pixels; (2) RGB is also a good color space for lumen pixel classification.

3.2. Image Classification

To determine whether an image is a lumen view or a wall view, we investigated applications of two well known machine learning methods: decision tree and support vector machine (SVM). For our image classification, we use C4.5 variant implemented by WEKA [10] as well as SVM with the Gaussian RBF as the kernel function. The kernel function plays an important role for SVM. In this study, the parameters of the Gaussian RBF are found optimally using a “grid search”.

Feature selection is critical to the effectiveness of the classification. To obtain the features, each color image is filtered by a corner mask that can help to remove non-mucosa pixels on the four corners of image; the image generated by the endoscopes used for these studies have an octagonal shape (see Figure 3(a)). We introduce the following features.

1. Number of lumen pixels identified by the lumen pixel classifier. Recall that the lumen pixel classifier outputs an intermediate red-green image. We count the number of lumen pixels (green pixels) as the feature. Figure 3 (b) shows the generated intermediate image from Figure 3(a). The red-green image removes the complex background in
the original image that is not useful for lumen or wall view classification. We call the green pixels as foreground pixels hereafter.

(2) **Number of foreground pixels in the filtered red-green image with vertical filter.** The morphological operation opening is applied on the red-green image to smooth the contour of foreground objects (in green) and to remove thin and small isolated objects. The opening of image \( I \) by structuring element \( B \) is defined as the erosion of \( I \) by \( B \) followed by the dilation of the result of the first erosion. Here, we use a column vector (vertical filter) of length 3 as the structuring element \( B \). The area of the filtered red-green image is computed as the second image feature. Unlike the first feature, this feature does not take into account small isolated spots in the red-green image (Figure 3(b)).

(3) **Difference of foreground pixels in the filtered red-green images with horizontal and vertical filters.** This feature is introduced to indicate that the image may be partially blurred. This type of image is not filtered out by our non-informative frame filtration to avoid missing any important information in the clear part of the image. For a partially blurred image, other features may not be as pronounced as in a clear image. We apply the morphological opening to the red-green image with another structuring element \( B' \). This structuring element is a row vector of length 3 (horizontal filter). Our third image feature is the difference between the number of the foreground pixels of the filtered image and the second feature (the number of foreground pixels of the filtered image using the vertical filter). For a partially blurred image, the difference is noticeably larger than that of the clear image. The amount of the difference depends on how large the blurred part of the image is.

(4) **Number of white pixels in the binary image generated from the original color image.** First, the original color image is converted to a grayscale image. Second, a binary mask image is created from the red-green image as follows. A 0 is assigned to the mask image at each location corresponding to the location of a red pixel in the red-green image. A 1 is assigned to the rest of the pixels. Third, the filtered grayscale image is obtained by filtering the grayscale image in the first step using the mask image in the second step. Figure 3(c) is generated by filtering the grayscale representation (not shown) of the original image in Figure 3(a) with the mask image created from Figure 3(b). Figure 3(c) shows the lumen area as well as the most distant part (darkest) of the lumen seen in the image. The new filtered grayscale image is converted to a binary image using one iteration of Otsu’s adaptive threshold method [11]. The number of pixels with the lowest pixel intensity is used as the fourth image feature. For a lumen view, this feature indicates the area of the distant lumen.

### Table 2. Effectiveness of Image Classification Between CVM (Color-based View Mode Classification) and GSVM (Grayscale Shape-based View Mode Classification)

<table>
<thead>
<tr>
<th>Video</th>
<th>Pre</th>
<th>Rec</th>
<th>Spe</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSVM</td>
<td>0.90</td>
<td>0.81</td>
<td>0.84</td>
<td>0.82</td>
</tr>
<tr>
<td>CVM-S</td>
<td>0.91</td>
<td>0.85</td>
<td>0.84</td>
<td>0.85</td>
</tr>
<tr>
<td>CVM-D</td>
<td>0.93</td>
<td>0.85</td>
<td>0.89</td>
<td>0.90</td>
</tr>
</tbody>
</table>

(5) **Area of the largest foreground object in a reconstructed image.** To obtain this feature, we apply opening-by-reconstruction followed by closing-by-reconstruction [12] to the filtered grayscale image generated to get the previous feature. These two operations were shown to address the sensitivity of morphological opening due to the choice of the structuring element and the shape of interest. Figure 3(d) is the reconstructed image from Figure 3(b). A connected component labeling algorithm [13] is then applied to find the largest connected cluster in the image and the area of this largest cluster is returned as the fifth image feature.

(6) **Distance of the centroid of the largest cluster from the image boundary.** The sixth feature is the shortest distance from the image boundary to the centroid of the largest cluster identified during the computation of the previous feature. This feature is helpful when a small distant lumen is seen in a lumen view.

(7) **Ratio of the length of the minor axis to the length of the major axis of the ellipse that fits the detected largest cluster.** We compute this feature using the lengths of the minor and major axes of the ellipse that has the same normalized second-order moments as the detected largest cluster. This shape feature helps discriminating the real lumen from the non-lumen in the image when the lumen is seen at different angle in the image.

### 3.3. Data Sets and Classification Results

The ground truth classification process was very time consuming (at least 10 hours per video) due to the complexity of the images. We created three image data sets from three randomly selected videos as follows. For each video, we extracted one frame per second from the video and removed blurry frames from the data set. We manually classified the remaining images into lumen views and wall views.

We performed ten-fold cross validation on these data sets. Four performance metrics, precision, recall, specificity, and accuracy, were gathered to evaluate our new technique CVM (Color-based View Mode Classification) and our previous technique GSVM (Grayscale Shape-based View Mode Classification). We denote CVM using the decision tree classifier as CVM-D and CVM using the SVM classifier as CVM-S. For GSVM, we first determined the thresholds that give the best results via experiments for each video. The performance comparison between CVM and GSVM is shown in Table 2.
Table 2 clearly shows that CVM with either classifier outperforms GSVM in all four metrics. The overall average accuracy for the three videos is slightly below 0.90 for CVM-D. The recall (correct lumen classification) is noticeably lower than the other performance metrics for video043 and video046. This is because CVM misclassified some lumen views as wall views for images in which the appearance of the lumen is not very dark. CVM misclassified wall views as lumen views due to dark non-lumen pixels in these wall views. Unlike the other two videos, the specificity for video045 is lower than the recall. The wall views in this video have different characteristics. In many wall view images parts of the colon wall protrude and create shadows similar to the appearance of the lumen pixels (see Figure 1(f)). Such images affect the effectiveness in correctly identifying wall views. CVM with the decision tree classifier slightly outperforms the SVM classifier. When examining the rule set obtained from the decision tree classifier, we found that all features were used in the rules. Although performance of GSVM as shown in Table 2 using manually obtained optimized thresholds for each video is quite good, manual thresholding is not practical.

4. Conclusion

We have developed novel algorithms based on machine learning techniques that give a first quantitative estimate of the thoroughness of mucosa inspection during a colonoscopic procedure. Colonoscopic procedures that have mostly global inspection views during the withdrawal phase may imply a less desirable quality since close inspection of the colon wall is not seen and some lesions, such as those behind a fold, may be missed. Indeed, optimal withdrawal during colonoscopy probably is defined by a continual interchange between global and wall views. These measurements can be used as building blocks for further development of profiles representing different levels of procedure quality; prospective studies will have to provide inside into what automated profile is associated with best patient outcome. Our methods are adaptable to other types of endoscopic procedures.

References


