HHS Public Access

Author manuscript

J Biophotonics. Author manuscript; available in PMC 2022 January 01.

Published in final edited form as:

J Biophotonics. 2021 January ; 14(1): e202000276. doi:10.1002/jbio.202000276.

Diagnosing colorectal abnormalities using scattering coefficient maps acquired from optical coherence tomography

Yifeng Zeng[1], **William C. Chapman Jr.**[2], **Yixiao Lin**[1], **Shuying Li**[1], **Matthew Mutch**[2], **Quing Zhu**[1,3]

[1]Department of Biomedical Engineering, Washington University, St. Louis, Missouri

[2]Department of Surgery, Section of Colon and Rectal Surgery, Washington University School of Medicine, St. Louis, Missouri

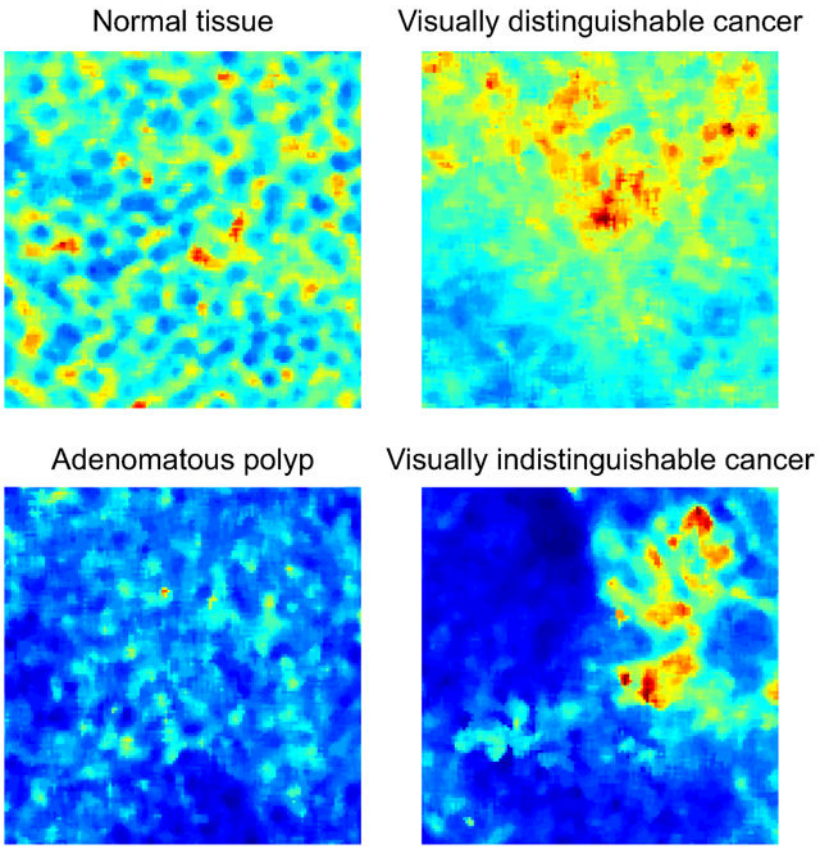
[3]Department of Radiology, Washington University School of Medicine, St. Louis, Missouri

Abstract

Optical coherence tomography (OCT) has shown potential in differentiating normal colonic mucosa from neoplasia. In this study of 33 fresh human colon specimens, we report the first use of texture features and computer-vision-based imaging features acquired from *en face* scattering coefficient maps to characterize colorectal tissue. *En face* scattering coefficient maps were generated automatically using a new fast integral imaging algorithm. From these maps, a gray-level co-occurrence matrix algorithm was used to extract texture features, and a scale-invariant feature transform algorithm was used to derive novel computer-vision-based features. In total, 25 features were obtained, and the importance of each feature in diagnosis was evaluated using a random forest model. Two classifiers were assessed on two different classification tasks. A support vector machine model was found to be optimal for distinguishing normal from abnormal tissue, with 94.7% sensitivity and 94.0% specificity, while a random forest model performed optimally in further differentiating abnormal tissues (i.e., cancerous tissue and adenomatous polyp) with 86.9% sensitivity and 85.0% specificity. These results demonstrated the potential of using OCT to aid the diagnosis of human colorectal disease.

Graphical Abstract

Corresponding author: zhu.q@wustl.edu.



Keywords

optical coherence tomography; colorectal cancer; scattering coefficient map; feature engineering; machine learning

Introduction

As of 2020, colorectal cancer (CRC) is estimated to be the third most prevalent type of cancer and to have the third highest mortality rate among all cancer types in the US[1]. CRC typically starts in the mucosa layer, which is within 1 mm from the surface, in the form of polyps. It has been proposed that the progression of CRC follows the adenoma-carcinoma sequence[2,3]. Although the vast majority of cancers arise from a polyp beginning with an aberrant crypt, deterministically predicting the outcome of a polyp remains a medical challenge[4]. As the polyp develops, it penetrates through deeper layers, and, left untreated, the disease is fatal. Currently, screening for colorectal abnormality is performed by flexible endoscopy, which relies on a camera for visual inspection of the colon and rectum[5,6]. Although colonoscopy is considered the gold standard for accuracy, it has limitations, such as the lack of quantitative justification beyond visual inspection and the challenge of detecting diminutive, flat, or subsurface neoplastic growths[7,8]. Often, early malignancies can be missed[9], but early intervention can provide significant survival advantages[10]. Therefore, we hypothesized that knowledge of subsurface tissue optical properties would

improve screening and surveillance of CRC and its potential risk factors when they are still undetectable under direct visual examination.

Optical coherence tomography (OCT) is a high-resolution imaging technique that can probe about 1 mm into the surface[11-13], and it has been commercialized in ophthalmology and cardiology[14-17]. OCT has also been extensively studied as an “optical biopsy” tool for differentiating malignant tissue from abnormal/normal tissue in multiple organ systems[18-21]. Recently, Yu et al. has used *en face* images acquired from *ex vivo* colorectal tissue using micro-OCT for differentiating adenomas and non-neoplastic polyps[22]. A 94.83% accuracy was achieved when evaluating the micro-OCT images. This study used tissue morphology and professional readers. Tissue optical scattering coefficient maps can be computed from OCT volumetric data, enabling quantification of early cancer morphological changes for diagnostic purposes. Yi et al. used inverse spectroscopic OCT to evaluate the optical scattering coefficient and ultrastructural properties of *ex vivo* colorectal biopsy samples[23]. They discovered an alternation in optical and ultrastructural properties in cancerous tissues. However, no further quantitative evaluation of diagnostic accuracy or machine learning capability was introduced. However, the OCT technique generates an enormous volume of data, which is slow and laborious to process manually. That’s why computer-aided diagnosis from OCT data, especially from within the GI tract, has garnered increasing interest in recent years[24-26].

Computer-aided diagnosis based on radiographic images requires feature extraction and texture analysis, and most features/textures are difficult to register under visual inspection. Texture analysis assumes that textural information is contained in the local gray-scale variations of an image[27]. Currently, its major medical applications are in oncology, such as automated tumor segmentation and grading, as well as in characterization of tumor heterogeneity[28]. Relevant oncological studies on PET[29,30], MRI[31], and CT[32] have demonstrated diagnostic results that are comparable to diagnoses from expert radiologists. Specifically, texture analysis has been applied in OCT for feature extraction. Almog et al. found more success using texture features which enabled them to differentiate between homogeneous gray matter and other brain regions with more complex structures[33]. Chen et al. used retinal vessel OCT images and their corresponding GLCM features for anemia screening. Experimental results demonstrated an 83.6% accuracy, suggesting potential for future clinical utilization[34]. Ashok et al. combined Raman spectroscopy and OCT for colorectal cancer diagnosis[35]. With the help of texture features, a sensitivity and specificity of ~94% was achieved for cancer versus normal snap-frozen tissue samples. Scale-invariant feature transform (SIFT) is an algorithm used in computer vision for object detection[36]. Unlike texture analysis, this algorithm detects local image descriptors that can be visually identified. It then quantifies these features by finding interest points, using Gaussian kernels with different scales. This algorithm has been adopted in genetic analysis of colorectal cancer[37] and segmentation of kidney lesion areas in CT images[38]. Sun et al. proposed a method using SIFT descriptors and multiclass linear SVM for OCT image classification between diseased and normal retina OCT scans[39]. Using this method, 100% of 30 OCT volumes under the diseased classes were correctly classified and 93.33% of 15 OCT volumes under the normal class were correctly classified.

In this *ex vivo* study of human colorectal cancer, we used swept-source OCT (SS-OCT) to acquire volumetric structural information about the colorectal tissue. Then the tissue scattering coefficient was computed using an algorithm developed in-house, generating scattering coefficient maps of the entire imaged region. An integral image algorithm, which reduced the processing time by 25%, was employed for image preprocessing. Regions of interest (ROIs) were manually selected from scattering maps and analyzed using a set of statistical texture features along with computer-vision related features. Significant features were selected based on feature importance, and a model with the reduced feature set was constructed to classify a tissue sample into normal tissue, cancerous tissue, or polyp. The results demonstrated the feasibility and potential for an alternative and improved way to differentiate colorectal tissue. To the best of our knowledge, this is the first report on using texture features and computer vision-based image features acquired from scattering coefficient maps to differentiate malignant, polypoid, and normal colorectal tissues.

Materials & Methods

Colon Specimen Preparation

Thirty-three patients (mean age, 66 years; range, 42-91; detailed characterizations in Table 1) undergoing extirpative colonic resection at Washington University School of Medicine were recruited to our initial study from August 2017 to February 2020. We studied one resected colorectal specimen from each patient. Among these specimens twenty-five were cancerous and four contained adenomatous polyps. We imaged one area per abnormality, i.e., twenty-five cancer areas and four adenomatous polyp areas. For imaging normal colorectal tissue regions, we used two criteria to select the imaging area. First, if there were any abnormal growth in the resected tissue, the normal area needed to be at least 5 cm far from it. Second, only a single normal area per patient was evaluated. Using these selection criteria, twenty-six normal areas were imaged. The study protocol was approved by the Intuitional Review Board and informed consents were obtained from all patients. All samples were imaged within one hour after resection, and diagnoses were ascertained by subsequent pathology examination of the surgical specimen. Twenty-five cancer areas, twenty-six normal areas, and four adenomatous polyp areas were imaged.

OCT System Setup

The SS-OCT system was based on a 1310 nm center-wavelength swept source (HSL-2000, Santec Corp., Japan) with a 110 nm full-width-at-half-maximum bandwidth and a 20 kHz scan rate. A balanced detector (Thorlabs PDB450C) detected the interference signal and sent to a data acquisition board (ATS9462, Alazartec Technologies Inc). The lateral resolution of the system in air was 10 μm, and the axial resolution was 6 μm. Details of the imaging system and experimental setup can be found in our previous work[40].

en face Scattering Coefficient Mapping

To generate the *en face* scattering coefficient map, we first automatically located the epithelium layer without any human intervention. Our surface detection algorithm will first read in a B-scan image, and then output the coordinates of the surface of the epithelium layer. Adding a 1 mm depth, which is the typical mucosa thickness, the mucosa layer was

extracted automatically. Details can be found in our previous publication[41]. Formerly, this could take up to 14 hours for a 5 mm x 10 mm area, which corresponds to 500 B-scans with 1000 pixels by 1000 pixels per B-scan. In short, we formulated the surface delineation job as a global optimization problem. We created a matrix representation of each B-scan image named I, in which each entry $I(z, x)$ represented the intensity of pixel (z, x). Here, x represented the lateral dimension and z was the depth dimension. We let $\boldsymbol{l}$ represented a vector, with each entry $l(i)$ representing the imaged surface depth in each column of the image matrix. We optimized $V(\boldsymbol{l}(i)) = C(i) + \alpha * Diff(i) + V(\boldsymbol{l}(i-1))$ to find the epithelium surface $\boldsymbol{l}$. The most computationally expensive aspect was calculating $Diff(i) = \sum_{j=l(i)-w}^{l(i)-1} \boldsymbol{I}(j,i) - \sum_{j=l(i)}^{l(i)+w-1} \boldsymbol{I}(j,i)$ for every pixel, where I is the OCT signal intensity and w is a custom-defined window size, for which we chose 10 pixels. Hence, for global optimization, we needed to calculate $n * n * 2w$ (i.e., $n * n * 20$) times, where n is the number of pixels within each B-scan. In this work, we introduce a preprocessing technique named integral image. As the name suggests, the value at any pixel (z,x) in the integral image is the sum of all the pixel intensities above and to the left of the pixel (z,x), expressed as $II(z,x) = \sum_{z' \le z, x' \le x} I(z', x')$, where $I(z', x')$ is the intensity of the pixel (z', x') in the original image and $II(z,x)$ is the intensity of the pixel (z,x) in the integral image. Figure 1 shows an example of the integral image technique, proceeding from the original image to the integral image. Fig. 1A is the original image intensity distribution: $I(z', x')$. Fig. 1B shows the calculated values for the first four elements in the integral image. For example: $II(1, 2) = I(1, 1) + I(1, 2) = 1 + 1 = 2$; $II(2, 1) = I(1, 1) + I(2, 1) = 1 + 2 = 3$; $II(2, 2) = I(1, 1) + I(1, 2) + I(2, 1) + I(2, 2) = 1 + 1 + 2 + 3 = 7$. Fig. 1C is one more step from Fig. 1B: $II(2, 3) = I(1, 1) + I(1, 2) + I(1, 3) + I(2, 1) + I(2, 2) + I(2, 3) = 1 + 1 + 3 + 2 + 3 + 4 = 14$. Since we have information in those pixels which were already derived in the integral image, we can further simplify the calculation of $II(2, 3)$ to: $II(2, 3) = II(2, 2) + II(1, 3) - II(1, 2) + I(2, 3) = 7 + 5 - 2 + 4 = 14$. In general, it will take $n * n * 4$ computations to generate the entire integral image (Fig. 1D). Once the integral image has been computed, evaluating the sum of intensities over any rectangular area requires only four pixels in the integral image, regardless of the area's size. Say the coordinates of the four vertices are $A = (z,x)$, $B = (z', x)$, $C = (z, x')$, and $D = (z', x')$. The sum of the pixel intensities over the rectangle $ABCD$ is $II(D = (z', x')) + II(A' = (z-1, x-1)), - II(B' = (z', x-1)) - II(C' = (z-1, x'))$ For example, to calculate the sum of the green area (vertices: $A = (3,3)$, $B = (4,3)$, $C = (3,5)$, $D = (4,5)$) in Fig. 1(A), we need just the four red pixels (vertices: $A' = (2,2)$, $B' = (4,2)$, $C' = (2,5)$, $D = (4,5)$) in Fig. 1(D): $8 + 6 + 2 + 1 + 2 + 3 = 22 = 80 + 7 - 28 - 37$. In practice, we performed zero padding to avoid boundary conditions. Therefore, the total computation time needed to calculate the $Diff(i)$ for every pixel is $n * n * 4 + n * n * (4 + 4 + 1) = n * n * 13$. Theoretically, a 35% reduction of computation time for calculating $Diff$ is expected. In practice, the reduction is 25%, due to the involvement of other calculations (Table 2).

After localization of the epithelium layer, scattering coefficients were extracted using Beer's Law and an *en face* scattering coefficient map was generated by applying this method to an OCT image volume. On each constructed *en face* scattering coefficient map, ROIs were manually selected for further quantitative analysis. Each ROI was 128 by 128 pixels, corresponding to 1.28 mm by 1.28 mm in physical dimensions. The final set of ROIs consisted of 121 normal regions, 84 malignant regions, and 24 adenomatous polyp regions.

Feature Extraction

Three sets of features were extracted from the ROIs: (1) six features extracted from the scattering maps, i.e., the mean scattering coefficient, median scattering coefficient, image entropy of the scattering map, and the10^{th}, 25^{th}, and 75^{th} percentiles; (2) nineteen texture parameters derived from the gray-level co-occurrence matrix (GLCM) of the scattering maps; and (3) four computer-vision based features acquired using SIFT, and one feature called the angular spectrum index (ASI), constructed as described in our previous work[41]. All feature extraction procedures were done in MATLAB 2019b.

The GLCM allows the calculation of texture features by describing the relationship among neighboring pixels within an image[30]. It measures how often different combinations of pixel intensities occur among neighboring pixels. To calculate the GLCM for each ROI, we first converted each ROI to a grayscale image. Then the GLCMs were generated for four directions, $\theta = \{0°, 45°, 90°, 135°\}$, assuming the distance between two neighboring pixels is 1. Then these four matrices were averaged to generate the final GLCM, which is rotationally invariant to the intensity distribution.

We used the SIFT algorithm, a feature detection algorithm in computer vision, to locate points of interest[36]. The original image $I(x,y)$ was first convolved with a Gaussian kernel $G(x,y,k\sigma)$ at scale $k\sigma$:

$$L(x, y, k\sigma) = G(x, y, k\sigma) * I(x, y).$$

Then the differences of the Gaussians that occurred at multiple scales were calculated by

$$D(x, y, k\sigma) = L(x, y, k_i\sigma) - L(x, y, k_j\sigma).$$

Next, the algorithm used the differences as templates for pattern matching across each ROI, searching for interest points. Possible candidates were further screened using gradient and Hessian tests to remove edge points, and the number of interest points on each ROI was tallied. To characterize their distribution, the coordinate matrix of interest points was analyzed using principal component analysis, and the eigenvalues of the first two principal components were extracted. The relative difference between them was calculated to reflect the regularity of the interest points' distribution.

ASI is a feature which can evaluate whether there is a periodic structural pattern within an image. It first calculates the 2-dimensional fast Fourier transformation (2D FFT) of the image. After 2D FFT, a frequency spectrum of the image is generated. If there is a periodic structural pattern within the image, the frequency spectrum will show in higher spatial frequency band. ASI measures the ratio between higher spatial frequencies and all spatial frequencies. In general, normal colorectal tissues have a well-organized crypt pattern and the ASI is higher; while cancerous tissues are heterogeneous even in histopathological level and have a lower ASI.

Feature Selection & Image Classification

Feature selection is essential to avoid overfitting and to provide meaningful information from 25 features. Random forest (RF) is one of the most popular machine learning algorithms, as well as being a well-established feature selection algorithm[42]. Feature selection using RF is categorized as an embedded method, offering the advantages of accuracy and generalizability. A random forest consists of hundreds of decision trees, and each node within a decision tree represents a rule for splitting data by using a single feature. The rule is based on Gini impurity (or information gain). When training a tree, we can compute how much each feature contributes to decrease the weighted impurity. In the sense of a random forest, we average the decrease in impurity caused by a single feature over all the trees to evaluate the feature's importance. According to their importance, we add features one-by-one to machine learning classifiers until an optimal dataset is found. A Python module, Scikit-Learn, was used for generating the feature importance.

Image classification was done in two phases. First, different machine learning classifiers were evaluated for differentiating abnormal tissues (cancers and adenomatous polyps) from normal tissues. Second, those classifiers were further tested on distinguishing cancerous tissues and adenomatous polyps. Two classifiers were evaluated, support vector machine (SVM) and RF. All features were normalized to avoid systematic biases. The evaluation of each model was based on the average performance from 100 repetitions of random train-test splits to minimize the randomness of single train-test splits. A train-test split for model fitting was defined as follows: the training set size was defined as $\frac{2}{3}$ of the smallest sample set, and then training data were chosen randomly from each diagnosis, while the rest were used for testing. Finally, the area under the receiver operating characteristic (ROC) curve (AUC) was used for determining both the optimal performance and the optimal feature sets of each model.

Results

en face Scattering Coefficient Maps

En face scattering coefficient mapping was performed on all OCT 3-D volumes. Figure 2(A-C) shows representative scattering map ROIs from three different diagnoses. (A) is a ROI from normal tissue, within which a dotted pattern can be found. This pattern appears because the normal crypt pattern in the colon mucosa layer results in a crater structure in *en face* scattering maps. (B) and (C) are ROIs from an adenomatous polyp and cancerous tissue, respectively. Since abnormality growth breaks the crypt pattern and result in heterogeneous tissue distribution, no clear dotted pattern is found.

Figure 2(D-F) comes from a special colorectal cancer case. Figure 2(D) is a photograph of this imaged tissue. The cancer area (green box) is flat and almost indistinguishable under visual inspection. This area was discovered using biopsy during colonoscopy since it was suspicious to an experienced endoscopist. It was also confirmed with following histopathology examination after OCT imaging. The histopathology slide is shown in Fig. 2(G). Figure 2(E) is a scattering map of the imaged area (blue box in Fig. 2(D). The distribution is heterogeneous and no dotted pattern can be found. The red box is a

representative ROI, and Fig. 2(F) shows an enlarged view of this area. Both the regular cancer (Fig. 2(C)) and the flat cancer (Fig. 2(F)) show a heterogeneous scattering coefficient distribution.

Feature Importance

Figure 3 and Figure 4 summarize the degrees of importance for each feature derived from the random forest classifier using Gini impurity. Figure 3 shows the relative individual importance of each variable included in the model differentiating malignant from normal tissue. Two computer-vision based features, ASI and SIFT interest points, are the two most important features. This result is expected from the scattering map (Fig. 2) since there is a unique image pattern within normal tissues. Likewise, Fig. 4 displays the relative importance of the variables used in the model differentiating polypoid from cancerous tissues. Due to the lack of specific image patterns, texture features show higher importance among all features.

Image Classification

Table 3 shows the testing AUC of RF and SVM trained by different feature sets for distinguishing abnormal from normal tissue. The feature set starts with the most important feature, and adds other features one by one according to their importance rank. RF and SVM show similar trends, and they achieve optimal performance when three features are used. Adding more features does not increase the AUC. Trained by the optimal feature set, RF achieves an AUC, sensitivity, and specificity of 0.973, 90.0%, and 94.4%, whereas SVM does marginally better, with an AUC, sensitivity, and specificity of 0.984, 94.7%, and 94.0%. We conclude that ASI, SIFT interest points, and IMC 1 form the optimal feature set, while SVM performs better in distinguishing abnormal from normal tissue.

Table 4 shows the testing AUC of RF and SVM trained by different feature sets for distinguishing adenomatous polyp from cancerous tissue. RF achieves an optimal performance when four features are used, and SVM achieves an optimal performance when six features are used. Trained by the optimal feature set, RF can achieve an AUC, sensitivity, and specificity of 0.913, 86.9%, and 85.0%, whereas SVM can achieve an AUC, sensitivity, and specificity of 0.892, 81.0%, and 84.8%. Therefore, we conclude that RF has a better performance in distinguishing adenomatous polyp from abnormal tissue, with an optimal feature set of Correlation, MCC, ASI, and Image entropy.

Discussion

This is the first report using texture features and computer vision-based image features acquired from scattering coefficient maps to differentiate malignant, polypoid, and normal colorectal tissue types. From 33 patients, 121 normal, 84 cancer, and 24 polyp ROIs were processed, and 25 features were then derived. Two classifications were assigned: abnormal tissue vs. normal tissue and adenomatous polyp vs. cancerous tissue. Based on the RF classifier using Gini impurity, the feature importance ranking for each task was calculated. Two classifiers, RF and SVM, were trained on different feature sets according to the feature importance, and the optimal feature set was found based on the AUC. The results indicate

that SVM with computer vision-based features (including ASI, SIFT interest points, and IMC1) is suitable for distinguishing abnormal and normal tissues, while RF with texture features (including Correlation, MCC, ASI, and Image entropy) shows better performance in identifying benign polyps.

When identifying abnormal and normal tissues, we found a specific dotted pattern related to the well-organized crypt pattern in the mucosa layer of the colorectal tissue. Studies have shown that changes in crypt size and appearance are associated with the earliest forms of colorectal cancer[43]. Fig. 2(D)-(F) provide evidence that our scattering map approach can detect early colorectal cancer before it becomes visible with a normal endoscopic camera. Since this dotted pattern occurs only in normal tissues, computer vision-based features show great predictive importance because these features fit the morphology best. Both RF and SVM perform well. SVM is slightly better, possibly because separating abnormal and normal tissue is a relatively easier task, i.e., only three features are essential for a high accuracy.

In classifying adenomatous polyp and cancerous tissue, no distinguishing imaging pattern was found. Therefore, texture features are more important for accurate differentiation. Abnormalities in colorectal cancer are heterogeneous even at the histology level, which makes separating different types of abnormality a difficult non-linear problem. Texture features are more important for accurate differentiation under this situation because it provides a statistical measure of the intensity variation in space by evaluating a pixel's intensity with respect to its neighbors. By evaluating contrast, uniformity of energy, correlation, and homogeneity, texture features can reveal tissue functional properties beyond morphology. Since RF is designed for non-linear problems while SVM needs a suitable non-linear kernel to solve such problems, RF yields a better result in this task. In a recent study using a deep-learning pattern-recognition method to classify OCT B-scans[44], we found that a very accurate diagnosis can be achieved for normal vs. cancer. However, there were limitations in effectively distinguishing polyp from cancer. Future study will focus on combining a feature-based method and a deep learning approach for a more accurate model.

Clinical translation of the scattering map requires integration of the probe into the colonoscopy for "optical biopsy" in real time during endoscopic evaluation. The application of OCT ancillary to endoscopy has been gaining momentum in recent years[45-47]. Because endoscopic OCT provides 3D structural information, it is particularly suited for inspecting diseases arising from the mucosa. Ahsen *et al.* investigated Barrett's esophagus with volumetric *en face* OCT[48]. One investigator developed a reading criteria for the imaged volumes. Three readers with different endoscopy/OCT experience were recruited to use the criteria to read the OCT datasets while blinded to the histopathological diagnoses. They discovered an atypical gland pattern under the mucosa in dysplasia tissues. This irregular pattern occurred in 100% of neoplasia datasets, however, as the authors stated, there was a selection bias due to the unbalance of dataset. Li *et al.* demonstrated a multimodal endoscopy for colorectal cancer detection, using OCT and near-infrared fluorescence imaging in a rat model[49]. They used OCT and NIR to monitor the growth of colorectal abnormalities, and concluded that NIR fluorescence imaging can identify the suspect lesions rapidly, and OCT can help visualize the microanatomy of the subsurface layer. This study

was qualitative and injection of contrast agent was required. Mora *et al.* developed a steerable OCT catheter for real-time assistance during teleoperated endoscopic treatment of colorectal cancer[50]. The catheter was steerable which could be used for enhancing the performance of the robotically controlled flexible colonoscope. This study demonstrated their work using swine *in vivo* images. With these promising implementation of OCT catheters, our method can implement fully automated data quantification and diagnosis based on a large imaging volume. This prospective can also help foster machine learning and OCT for clinical translation.

Integration of this technology into a colonoscope can facilitate investigation of several areas of study. First, such a device would enable *in vivo* imaging and testing. Since the eventual application of such technology would be in perfused tissue, testing it *in vivo* is key. However, we suspect that with minimal modifications, the machine learning algorithm will also function well, so long as it is trained with *in vivo* images. Second, colonoscopic imaging would allow investigation of the device's performance with other benign pathologies of the colon, such as inflammatory bowel disease and hyperplastic polyps. The ability to differentiate adenomatous polyps from hyperplastic polyps would have a significant clinical impact. Because hyperplastic polyps are a benign growth and usually not resected, an *in vivo* approach is essential to imaging such abnormalities. Certainly, more training samples from all types of abnormalities can improve the machine learning model. A large data base can enable us exploring more complicated training models that best suit the colorectal cancer diagnosis task. Additionally, we believe combining texture features from *en face* scattering maps and B-scan images can potentially help discriminate polyps and cancer, because scattering maps contain functional information of macro-structures (i.e. from an image volume) and B-scan images carry morphology micro-structure information.

In addition, real-time data processing is also crucial for "optical biopsy". We improved our image processing speed by 25% with the usage of integral image. At present, it takes around 11 minutes to generate an *en face* scattering map for a 1 mm x 1 mm area. Certainly, this is faster than obtaining biopsy results (at least one day), but it remains too slow to facilitate bedside decision-making. Migrating the data processing platform to a GPU using parallel processing is one possible solution that should be investigated in the future. Another possible solution is using a deep learning based surface detection method. This method can predict the surface within seconds after the deep learning model is well-trained. However, the ground truth has to be labeled manually and the training process is in general time-consuming.

Based on these results, we conclude that the scattering map derived from OCT images can provide qualitative and quantitative information which demonstrates the potential for aiding the diagnosis of human colorectal tissues. *In vivo* study is needed to validate the performance of our machine learning model. With further improvement, the scattering map may guide physicians during colonoscopy for early cancer screening and biopsy site selection. Future efforts will focus on the real-time image processing algorithm and integrating the OCT system into a clinical endoscope.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

The authors thank Michelle Cusumano, the research coordinator of the colorectal surgery division, for consenting patients and coordinating specimen studies. The authors appreciate pathology fellows Rehan Rais, Iván González, Zahra Alipour, and Heba Abdelal for helping with specimens and providing specimen diagnosis. The authors thank Prof. Deyali Chatterjee, Department of Pathology and Immunology, for her valuable help on pathology related issues. The authors appreciate the partial funding support of this work from National Cancer Institute (RO1 R01 CA228047, T32CA009621, and R01CA237664).

References

[1]. Siegel RL, Miller KD, and Jemal A, CA. Cancer J. Clin 2020, 70, 7. [PubMed: 31912902]

[2]. Kuipers EJ, Grady WM, Lieberman D, Seufferlein T, Sung JJ, Boelens PG, van de Velde CJH, and Watanabe T, Nat. Rev. Dis. Prim 2015, 1, 15065. [PubMed: 27189416]

[3]. Leslie A, Carey FA, Pratt NR, and Steele RJC, Br. J. Surg 2002, 89, 845. [PubMed: 12081733]

[4]. Levine JS, and Ahnen DJ, N. Engl. J. Med 2006, 355, 2551. [PubMed: 17167138]

[5]. Issa IA, and Noureddine M, World J. Gastroenterol 2017, 23, 5086. [PubMed: 28811705]

[6]. Church JM, Clin. Colon Rectal Surg 2005, 18, 141. [PubMed: 20011297]

[7]. Young PE, and Womeldorph CM, J. Cancer 2013, 4, 217. [PubMed: 23459594]

[8]. Li S, Zeng Y, Chapman WC, Erfanzadeh M, Nandy S, Mutch M, and Zhu Q, J. Biophotonics 2020, e201960241. [PubMed: 32125775]

[9]. Than M, Witherspoon J, Shami J, Patil P, and Saklani A, Ann. Gastroenterol 2015, 28, 94. [PubMed: 25609386]

[10]. Bibbins-Domingo K, Grossman DC, Curry SJ, Davidson KW, Epling JW, García FAR, Gillman MW, Harper DM, Kemper AR, Krist AH, Kurth AE, Landefeld CS, Mangione CM, Owens DK, Phillips WR, Phipps MG, Pignone MP, and Siu AL, JAMA - J. Am. Med. Assoc 2016, 315, 2564.

[11]. Zhou KC, Qian R, Degan S, Farsiu S, and Izatt JA, Nat. Photonics 2019, 13, 794.

[12]. Dong Z, Liu G, Ni G, Jerwick J, Duan L, and Zhou C, J. Biophotonics 2020, 13, e201960135. [PubMed: 31970879]

[13]. Zhou H, Dai Y, Gregori G, Rosenfeld PR, Duncan JL, Schwartz DM, and Wang RK, Biomed. Opt. Express 2020, 11, 1834. [PubMed: 32341851]

[14]. Bouma BE, Villiger M, Otsuka K, and Oh W-Y, Biomed. Opt. Express 2017, 8, 2660. [PubMed: 28663897]

[15]. Fujimoto J, and Swanson E, Investig. Ophthalmol. Vis. Sci 2016, 57, OCT1. [PubMed: 27409459]

[16]. Gan Y, Lye TH, Marboe CC, and Hendon CP, J. Biophotonics 2019, 12, e201900094. [PubMed: 31400074]

[17]. Zhang X, Beckmann L, Miller DA, Shao G, Cai Z, Sun C, Sheibani N, Liu X, Schuman J, Johnson M, Kume T, and Zhang HF, Invest. Ophthalmol. Vis. Sci 2020, 61, 23.

[18]. Shostak E, Hariri LP, Cheng GZ, Adams DC, and Suter MJ, J. Bronchol. Interv. Pulmonol 2018, 25, 189.

[19]. Meiburger KM, Chen Z, Sinz C, Hoover E, Minneman M, Ensher J, Kittler H, Leitgeb RA, Drexler W, and Liu M, J. Biophotonics 2019, 12, e201900131. [PubMed: 31100191]

[20]. Zeng Y, Nandy S, Rao B, Li S, Hagemann AR, Kuroki LK, McCourt C, Mutch DG, Powell MA, Hagemann IS, and Zhu Q, J. Biophotonics 2019, 12, e201900115. [PubMed: 31304678]

[21]. Juarez-Chambi RM, Kut C, Rico-Jimenez JJ, Chaichana KL, Xi J, Campos-Delgado DU, Rodriguez FJ, Quinones-Hinojosa A, Li X, and Jo JA, Clin. Cancer Res 2019, 25, 6329. [PubMed: 31315883]

[22]. Yu X, Wang X, Yang T, Li N, Ding Q, and Liu L, IST 2019 - IEEE Int. Conf. Imaging Syst. Tech. Proc 2019, 1.

[23]. Yi J, Radosevich AJ, Stypula-Cyrus Y, Mutyal NN, Azarin SM, Horcher E, Goldberg MJ, Bianchi LK, Bajaj S, Roy HK, and Backman V, J. Biomed. Opt 2014, 19, 36013. [PubMed: 24643530]

[24]. Ahsen OO, Liang K, Lee HC, Wang Z, Fujimoto JG, and Mashimo H, World J. Gastroenterol 2019, 25, 1997. [PubMed: 31086467]

[25]. Wang Z, Lee HC, Ahsen OO, Liang K, Figueiredo M, Huang Q, Fujimoto JG, and Mashimo H, Appl. Sci 2018, 8, 2420.

[26]. Liu T, Lu M, Chen B, Zhong Q, Li J, He H, Mao H, and Ma H, J. Biophotonics 2019, 12, e201900151. [PubMed: 31465142]

[27]. Haralick RM, Shanmugam K, and Dinstein I, IEEE Trans. Syst. man Cybern 1973, 6, 610.

[28]. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout RGPM, Granton P, Zegers CML, Gillies R, Boellard R, Dekker A, and Aerts HJWL, Eur. J. Cancer 2012, 48, 441. [PubMed: 22257792]

[29]. Doumou G, Siddique M, Tsoumpas C, Goh V, and Cook GJ, Eur. Radiol 2015, 25, 2805. [PubMed: 25994189]

[30]. Giannini V, Mazzetti S, Bertotto I, Chiarenza C, Cauda S, Delmastro E, Bracco C, Di Dia A, Leone F, Medico E, Pisacane A, Ribero D, Stasi M, and Regge D, Eur. J. Nucl. Med. Mol. Imaging 2019, 46, 878. [PubMed: 30637502]

[31]. Wibmer A, Hricak H, Gondo T, Matsumoto K, Veeraraghavan H, Fehr D, Zheng J, Goldman D, Moskowitz C, Fine SW, Reuter VE, Eastham J, Sala E, and Vargas HA, Eur. Radiol 2015, 25, 2840. [PubMed: 25991476]

[32]. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Cavalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D, Hoebers F, Rietbergen MM, Leemans CR, Dekker A, Quackenbush J, Gillies RJ, and Lambin P, Nat. Commun 2014, 5, 4006. [PubMed: 24892406]

[33]. Almog IF, Der Chen F, Senova S, Fomenko A, Gondard E, Sacher WD, Lozano AM, and Poon JKS, J. Biophotonics 2020, 13, e201960083. [PubMed: 31710771]

[34]. Chen Z, Mo Y, Ouyang P, Shen H, Li D, and Zhao R, Med. Biol. Eng. Comput 2019, 57, 953. [PubMed: 30506116]

[35]. Ashok PC, Praveen BB, Bellini N, Riches A, Dholakia K, and Herrington CS, Biomed. Opt. Express 2013, 4, 2179. [PubMed: 24156073]

[36]. Cruz-Mota J, Bogdanova I, Paquier B, Bierlaire M, and Thiran JP, Int. J. Comput. Vis 2012, 98, 217.

[37]. Broderick P, Bagratuni T, Vijayakrishnan J, Lubbe S, Chandler I, and Houlston RS, BMC Cancer 2006, 6, 243. [PubMed: 17029639]

[38]. jian Xia K, sheng Yin H, and dong Zhang Y, J. Med. Syst 2019, 43, 2.

[39]. Sun Y, Li S, and Sun Z, J. Biomed. Opt 2017, 22, 16012. [PubMed: 28114453]

[40]. Zeng Y, Rao B, Nandy S, Hagemann I, Siegel C, Powell M, and Zhu Q, Proc. SPIE 2018, 1048338.

[41]. Zeng Y, Rao B, Chapman WC, Nandy S, Rais R, González I, Chatterjee D, Mutch M, and Zhu Q, Sci. Rep 2019, 9, 2998. [PubMed: 30816153]

[42]. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, and Hamprecht FA, BMC Bioinformatics 2009, 10, 213. [PubMed: 19591666]

[43]. Tanaka T, J. Carcinog 2009, 8, 5. [PubMed: 19332896]

[44]. Zeng Y, Xu S, Chapman WC, Li S, Alipour Z, Abdelal H, Chatterjee D, Mutch M, and Zhu Q, Theranostics 2020, 10, 2587. [PubMed: 32194821]

[45]. Gora MJ, Suter MJ, Tearney GJ, and Li X, Biomed. Opt. Express 2017, 8, 2405. [PubMed: 28663882]

[46]. Li K, Liang W, Mavadia-Shukla J, Park HC, Li D, Yuan W, Wan S, and Li X, J. Biophotonics 2019, 12, e201800205. [PubMed: 30302923]

[47]. Welge WA, and Barton JK, Lasers Surg. Med 2017, 49, 249. [PubMed: 27546786]

[48]. Ahsen OO, Liang K, Lee HC, Giacomelli MG, Wang Z, Potsaid B, Figueiredo M, Huang Q, Jayaraman V, Fujimoto JG, and Mashimo H, Endoscopy 2019, 51, 355. [PubMed: 30261534]

[49]. Li Y, Zhu Z, Chen JJ, Jing JC, Sun C-H, Kim S, Chung P-S, and Chen Z, Biomed. Opt. Express 2019, 10, 2419. [PubMed: 31143497]

[50]. Mora OC, Zanne P, Zorn L, Nageotte F, Zulina N, Gravelyn S, Montgomery P, de Mathelin M, Dallemagne B, and Gora MJ, Biomed. Opt. Express 2020, 11, 1231. [PubMed: 32206405]

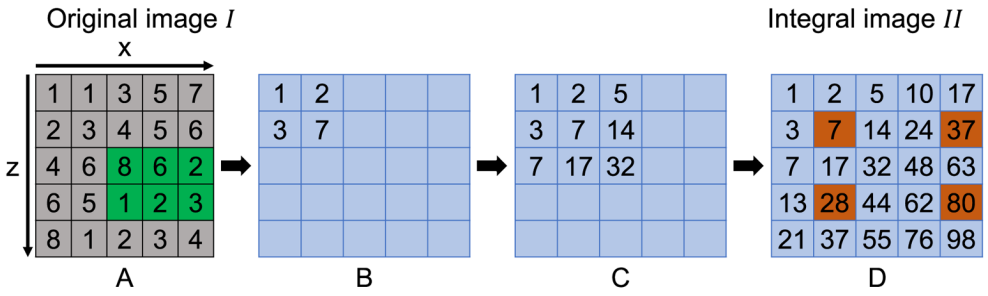


Figure 1.

Integral image example. A. Original image intensity distribution. B and C. Intermediate calculating processes of the integral image. D. The final integral image. To calculate the sum of the green area in the original image (8+6+2+1+2+3=22), we need only four pixels in the integral image (80+7−28−37=22).

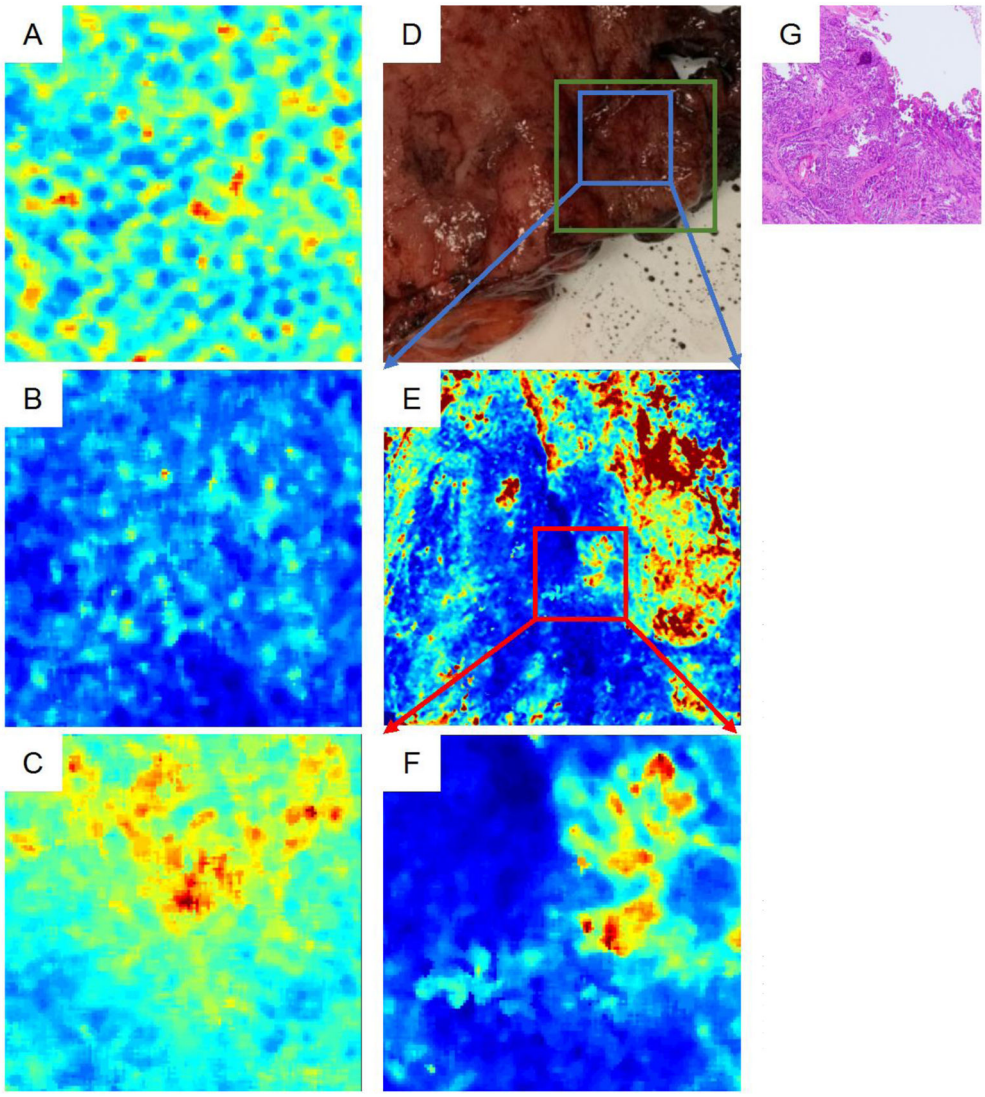


Figure 2 (revised manuscript).

Scattering coefficient maps. ROI of (A) a normal scattering map, (B) an adenomatous polyp scattering map, and (C) a cancerous scattering map. (D) Photograph of a small and almost indistinguishable cancer (green box). The blue box is the imaged area. (E) Scattering coefficient map of the imaged small tumor area. The red box is a representative ROI. (F) Enlarged ROI. (G) Histopathology slide of the cancer area.

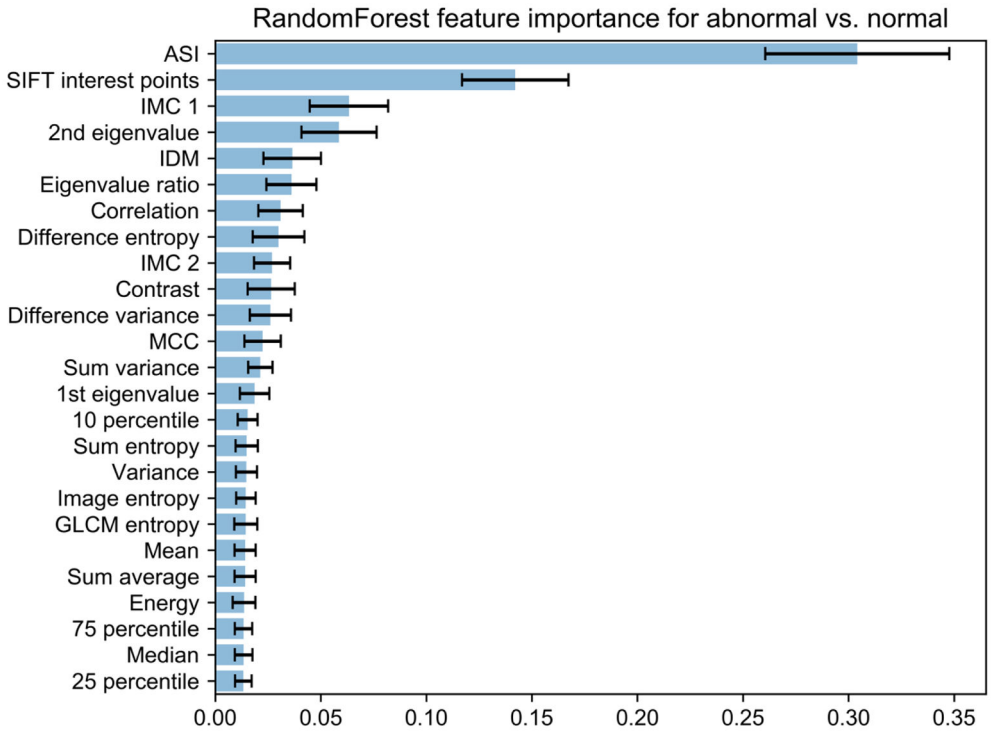


Figure 3.

Feature importance for identifying abnormal tissue from normal tissue. ASI, angular spectrum index; IMC, information measure of correlation; IDM, inverse difference moment; MCC, maximal correlation coefficient.

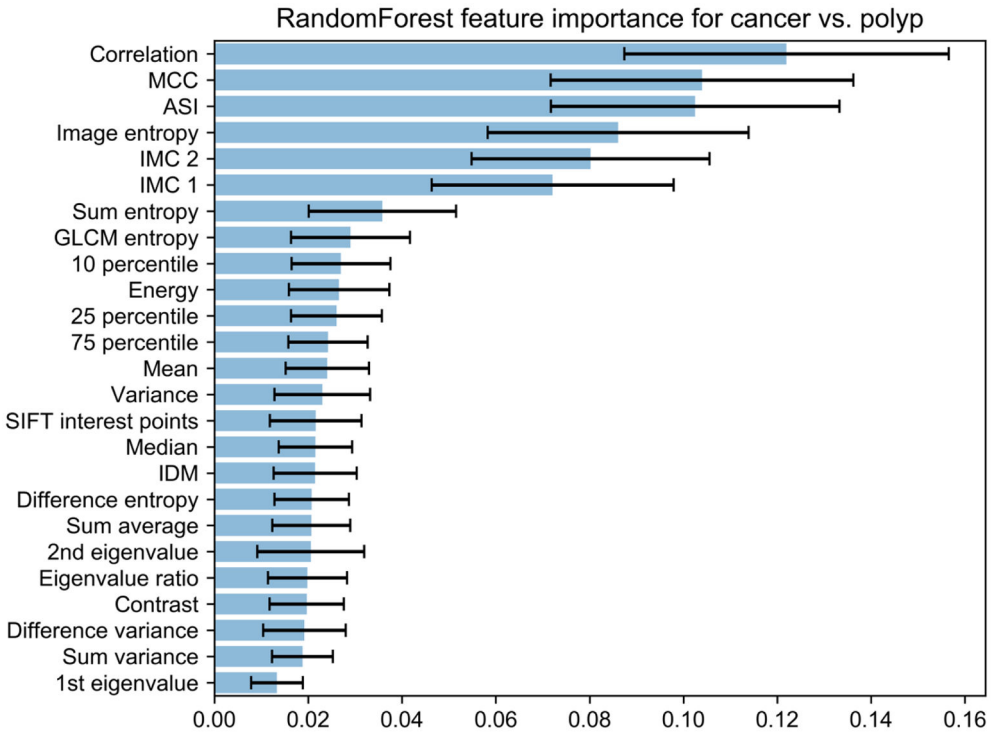


Figure 4.

Feature importance for distinguishing adenomatous polyp from cancerous tissue. MCC, maximal correlation coefficient; ASI, angular spectrum index; IMC, information measure of correlation; IDM, inverse difference moment.

Table 1.

Characteristics of the studied patients

Histologic Examination	Number of patients	Age (mean ± std)	Sex (% male)
Cancer	25	65 ± 12	72 %
T1 adenocarcinoma	2	63 ± 1	50 %
T2 adenocarcinoma	7	69 ± 10	57 %
T3 adenocarcinoma	15	64 ± 14	80 %
T4 adenocarcinoma	1	71	100%
Adenomatous polyp	4	70 ± 8	50 %
Tubular adenoma	2	74 ± 7	100 %
Tubulovillous adenoma	2	68 ± 10	0 %
Normal	26	64 ± 11	73 %

Table 2.

Surface detection time using different algorithms on various volumes of interest

Without integral image on a volume of 5 mm x 10 mm x 3 mm	**With integral image on a volume of 5 mm x 10 mm x 3 mm**	**With integral image on a volume of 1 mm x 1 mm x 3 mm**
47,989 s (13 hrs 19 mins 49 s)	35,831 s (9 hrs 57 mins 11 s)	700 s (11 mins 40 s)

Table 3.

Testing AUCs for distinguishing abnormal from normal tissue.

Features	Random Forest AUC	SVM AUC
ASI	0.938	0.945
+SIFT interest points	0.970	0.981
+IMC 1	0.973	0.984
+2nd eigenvalue	0.971	0.983
+IDM	0.972	0.985
+Eigenvalue ratio	0.972	0.984
All features	0.966	0.978

Table 4.

Testing AUCs for distinguishing adenomatous polyp from cancerous tissue.

Features	Random Forest AUC	SVM AUC
Correlation	0.753	0.860
+MCC	0.836	0.863
+ASI	0.882	0.879
+Image entropy	0.913	0.882
+IMC 2	0.905	0.887
+IMC 1	0.906	0.892
+ Sum entropy	0.906	0.888
+GLCM entropy	0.905	0.878
All features	0.895	0.878