Automatic Detection and Segmentation of Crohn’s Disease Tissues from Abdominal MRI

Dwarikanath Mahapatra*, Peter J. Schüßler, Jeroen A.W. Tielbeek, Jesica Makanyanga, Jaap Stoker, Stuart A. Taylor, Franciscus M. Vos, Joachim M. Buhmann

Abstract—We propose an information processing pipeline for segmenting parts of the bowel in abdominal magnetic resonance (MR) images that are affected with Crohn’s disease (CD). Given a MRI test volume, it is first oversegmented into supervoxels and each supervoxel is analyzed to detect presence of Crohn’s disease using random forest (RF) classifiers. The supervoxels identified as containing diseased tissues define the volume of interest (VOI). All voxels within the VOI are further investigated to segment the diseased region. Probability maps are generated for each voxel using a second set of RF classifiers which give the probabilities of each voxel being diseased, normal or background. The negative log-likelihood of these maps are used as penalty costs in a graph cut segmentation framework. Low level features like intensity statistics, texture anisotropy and curvature asymmetry, and high level context features are used at different stages. Smoothness constraints are imposed based on semantic information (importance of each feature to the classification task) derived from the second set of learned RF classifiers. Experimental results show that our method achieves high segmentation accuracy with Dice metric values of $0.90 \pm 0.01$ and Hausdorff distance of $7.3 \pm 0.8$ mm. Semantic information and context features are an integral part of our method and are robust to different levels of added noise.

Index Terms—Crohn’s disease, semantic information, graph cut, segmentation, random forests, probability maps, supervoxels, image features, context.

I. INTRODUCTION

Inflammatory bowel disease (IBD) constitutes a significant health care problem in the Western world affecting over 1 million European citizens. A significant percentage of the patients suffer from Crohn’s disease (CD) [1, 2, 3]. CD is considered to be caused by an auto-immune response and affects the digestive tract leading to abdominal pain, diarrhea, weight loss, fatigue and anemia. CD can be complicated by the development of bowel strictures, fistulae, and perianal disease. Although the disease can occur in people of all ages, the typical onset is in adolescence or young adulthood.

CD is characterized by episodes of exacerbation interspersed with periods of remission. Because there is no known cure for this relapsing and chronic disease, most patients require long-term medications, and many require hospitalization and surgery. This currently unfavorable prognosis negatively impacts the health and quality of life of affected patients, with severe dietary restrictions and substantial loss in economic productivity (accruing annual per person costs of thousands of Euros and billions of Euros at a larger scale) [4]. Accurate automatic detection of CD can help in rapid diagnosis and possibly reduce the time and cost involved in therapy planning and patient prognosis.

Magnetic resonance (MR) images of the diseased region show thickening of the bowel wall which can be identified by radiologists with sufficient training. For a meaningful contribution to CD therapy planning, automated methods must be able to segment the CD affected area with sufficient precision. This information in turn helps to calculate the length of diseased part, thickness of bowel wall and accurate visualization. Rate of contrast enhancement of the affected region can be determined from pre-registered dynamic contrast enhanced (DCE) MR images of the same patient. These are important parameters to detect presence of disease and grade its severity using a standard scale called Crohn’s Disease Endoscopic Index of Severity (CDEIS). Ultimately, our goal is to offer an alternative to the invasive CDEIS for assessment of disease severity by endoscopy. Before segmenting the diseased region, a volume of interest (VOI) is identified which is suspected to most likely contain diseased tissues. Each voxel within this VOI is then analyzed for disease activity. In this paper we propose an integrated framework for detecting and segmenting regions in abdominal MR images of the bowel which are affected by CD.

A. Related Work on Disease Classification and Crohn’s Disease

Colonoscopy is the current reference standard for CD diagnosis [5] in which a trained gastroenterologist examines the bowel and grades the severity and extent of inflammatory lesions using the CDEIS scale. Several drawbacks of colonoscopy like invasiveness, procedure related discomfort and risk of bowel perforation has fostered the exploration of non-invasive imaging techniques to assess extent and severity of CD [6]. In recent years, sonography and computed tomography (CT) have been explored as alternatives to colonoscopy [7]. For young patients, exposure to ionising radiations seriously limits the use of CT. Assessment in sonography is limited due to gas interposition where the gastrointestinal tract cannot be visualised properly. MRI has the potential to overcome these limitations because of high tissue contrast, lack of ionising radiations and lower incidence of adverse events related to intravenous contrast employed in CT [8]. Rimola et al. in [6] showed the relation between rate of contrast enhancement and bowel wall thickness with the presence of endoscopically active disease. However, reliance on explicit segmentation of the bowel wall and extensive manual scoring limits the ease of usage of their method.

There is very limited work related to computational aspects of CD detection. In our earlier work Vos et al. [9] introduced a workflow and tools for assessment of CD severity. It summarizes different stages like data collection, basic image analysis methodologies for registration, segmentation, classification and visualization of the bowel wall. The conclusions are based on preliminary results from different modules, and do not explore each module (e.g., disease detection and segmentation) in detail. The segmentation and classification work
in [9] addresses colon wall segmentation, and prediction of disease severity, which both exceed the scope of this work. In [10] we used different image features like image statistics, texture anisotropy, and shape asymmetry in a preliminary supervised learning framework to first identify voxels belonging to the bowel and then differentiate between diseased and normal samples using three different classifiers namely random forests (RF), support vector machines (SVM) and a Bayesian classifier (BC). RF and SVM showed similar performance which was significantly superior to BC. The two stage classification approach results in high computation time and requires additional post processing to remove outliers. Therefore it is essential to develop a system which can analyze abdominal MRI rapidly. We did not attempt to design a complete pipeline for localizing and segmenting the diseased regions in [10]. Based on the results of our previous works [9, 10] we propose a fully automated pipeline for automatic detection and segmentation of CD tissues in abdominal MR images in this paper.

We review some works related to disease detection and classification which are relevant to our work. Bhushan et al. in [11] used machine learning techniques on dynamic contrast enhanced (DCE) MRI data for identifying colorectal cancer. Schunk in [12] analyze MR images for their suitability in analyzing IBD, including CD. However they do not explore computational aspects and instead focus on the clinical part. Atasoy et al. in [13] addressed the tasks of localization, annotation or classification of optical biopsy videos in colonoscopy. These works highlight the challenges of analyzing colon and bowel MRI using machine learning techniques.

Pereyra et al. [14] jointly estimate the statistical distribution and segment lesions from skin ultrasound images. Multiple tissue images were modeled as Rayleigh distributions and the mixture parameters are estimated using Bayes method combined with a Markov chain Monte Carlo method. Madabhushi et al. have employed active contours on texture and boundary features to segment breast lesions [15]. They try to mathematically formulate the empirical rules used by radiologists in detecting ultrasonic breast lesions, popularly known as the “Stavros Criteria”. Song et al. [16] devised a multistage method for tumor and lymph node detection in PET-CT thoracic images. All potential abnormal regions are detected and differentiated as tumor or lymph nodes using SVMs and conditional random fields. Shepherd et al. [17] propose statistical shape models for lesion segmentation. They combine radial shape parameterization with machine learning techniques to develop dynamic contour models for lesion and tumor segmentation. The above works use shape information because the lesions are expected to have similar shapes in different images. However, the same is not true for CD affected regions. The affected area can be in any part and have any shape. Therefore prior shape information cannot be effectively used in our work.

Machine learning techniques are being increasingly used for detecting, localizing and grading diseases in medical applications. Cheng et al. [18] advocate the use of biologically inspired features for classification of different glaucoma types. The biological features are based on saliency maps which has been used for registration and segmentation of medical images [19, 20, 21, 22, 23]. Pauly et al. [24] employed a supervised regression technique to detect and localize different parts in whole body MR sequences. They use 3D local binary patterns along with random set of binary features, called Ferns to achieve high segmentation accuracy. Keln et al. in [25] propose a random regression method for detecting and grading coronary stenoses in CT angiography data. Their motivation is to provide an automated system that can rule out clinically relevant stenoses in the coronary arteries and serve as a second reader in the absence of an expert physician. These works highlight the importance of a two stage method for classification and detection. Machine learning methods have been used for classification of conditions of the brain, particularly Alzheimer’s disease [26, 27] and brain tumour classification [28]. Chest images have also generated a lot of interest with methods proposed for localization of chest pathologies in [29], paediatric tuberculosis in [30] and diffuse lung disease [31]. None of the above methods are directly applicable to CD tissue segmentation due to the complex intensity distributions of bowel tissues and the lack of shape information for CD regions.

B. Our Contribution

In this paper we propose a method to detect and segment CD affected tissues from abdominal MRI. The problem has two main challenges. First, we need to identify part of the bowel wall likely to show disease activity and second, differentiate between diseased and normal tissues in the bowel. Identifying disease activity in the bowel wall is challenging because of similar intensities of neighboring regions. Automatic bowel wall segmentation suffers from ambiguities in image data due to presence of image inhomogeneities, and other substances in the bowel wall (like fecal remains, gas, etc.). Moreover CD affected tissues do not have a fixed shape which makes it difficult to include shape information for segmentation.

Therefore we use specially designed features like entropy based texture and curvature anisotropy, and context features to obtain discriminative information. Analysis of entropy based image anisotropy has been discussed in [32]. Although our approach uses entropy, we partition an image patch into sectors which is similar to calculating anisotropy along specific directions. Template based steerable features for capturing context information were used in [33]. We employ a similar scheme based on templates but use different and fewer features.

To overcome the above challenges we develop a machine learning based approach using RF classifiers that identify regions having relevant bowel tissues. The test volume is first oversegmented into supervoxels and each supervoxel is classified for the presence of diseased tissue. The supervoxels containing diseased tissues constitute the relevant region of the bowel or the VOI most likely to show disease activity. A second set of RF classifiers output probability maps for every VOI voxel indicating its likelihood of being diseased or normal. They are integrated into a second order Markov random field (MRF) cost function and the final labels are obtained using graph cut optimization. RF classifiers are used for the following reason: (i) in [10] we demonstrate the superior performance of RF over SVM and BC classifiers in detecting CD samples. This result is relevant in identifying diseased supervoxels. (ii) RF classifiers allow us to quantify the relative importance of different features to the classification task [34]. These importance measures contribute to a novel smoothness cost and higher segmentation accuracy of diseased tissues. (iii) RF classifiers allow for a probabilistic interpretation of the classification of test samples which aids in designing an appropriate cost function for segmentation.

This paper makes the following contributions in terms of technical novelty. First we develop a hierarchical framework using supervoxel segmentation and trained RF classifiers to define a VOI that detects the relevant bowel areas. The RF classifiers also discard medically inconspicuous regions that do not show presence of CD activity. Our second contribution lies in the use of variable importance measures (or semantic information) from trained RF classifiers for medical image segmentation. Importance measures quantify the importance of different features in generating probability maps and are used to weight corresponding features in a novel smoothness cost for obtaining greater segmentation accuracy. The rest of the paper is structured as follows: Sections II-VI describe different parts of our method.
We describe our dataset and discuss our results in Sections VII-VIII. Section IX concludes our paper and outlines possible future work.

II. DESIGN OF THE CD ANALYSIS PIPELINE

Abnormal tissue segmentation is similar to assigning a relevant class to each voxel. Let \( V = \{ v_i : i = 1 \cdots N \} \) be a 3D MR image volume. We define the set of possible labels (or tissue classes) as \( \{ D, N, B \} \) representing diseased bowel, normal bowel and background (non-bowel) tissue. The label set for volume \( V \) is given by \( Y = \{ y_i : i = 1 \cdots N \} \) with one label \( y_i \) for each voxel \( v_i \). Feature extraction for each voxel in a 3D volume is very time consuming. Therefore, we automatically identify a smaller VOI using supervoxel segmentation and trained RF classifiers. Probability values of each voxel within the VOI indicate its likelihood of having different labels. A second set of trained RF classifiers are used for this purpose. We do not directly use intensity information for segmentation because different tissues usually have very similar intensity characteristics. The negative log-likelihood of the probability map is used as the penalty cost in a second order MRF cost function. Furthermore, the second set of trained RF classifiers provide semantic information about the importance of different features in the classification task. These importance measures are used to weight each image feature differently in the smoothness cost. The final class labels are obtained by optimizing the cost function using graph cuts to obtain a regularized solution. An overview of our method is given in Algorithm 1.

Algorithm 1 Segmentation of Crohn’s Disease regions

**Input** Image with \( N \) voxels.

**Output** Segmented diseased region

**Sequence of Steps:**

1. Bias correction and intensity normalization
2. Supervoxel segmentation using Algorithm 2
3. Classification of supervoxels to get VOI using RF classifiers
4. Generate probability maps of VOI using a second set of RF classifiers.
5. Calculate penalty cost and smoothness cost of voxels using Eqs 13, 14.
6. Obtain final segmentation labels using graph cuts

III. RANDOM FOREST CLASSIFIERS

Random forests (RF) [35] are being used increasingly by many medical applications like cancer classification, tissue segmentation, [36, 37, 38], to detect abnormalities in mammograms [39] and identify coronary artery stenoses [25]. They are computationally efficient for large training data, can solve multiclass classification problems, and the learned knowledge can be extracted and interpreted to get a deeper insight into the training procedure. RF also shows higher classification accuracy than SVMs for classifying CD samples [10]. An RF is an ensemble of decision trees, where each tree is typically trained with a different subset of the training set ("bagging"), thereby improving the generalization ability of the classifier. Samples are processed along a path from the root to a leaf in each tree by performing a binary test at each internal node along this path. A test compares a certain feature with a threshold. Training a forest amounts to identifying the set of tests that best separate the data into the different training classes. At each internal node, the feature space is searched for a test that maximizes the reduction of class impurity, typically measured with the class entropy. Rather than inspecting the full space of features at each node, a random subset is probed, and the best one is selected. Even if this choice renders the individual trees weaker, it decreases the correlation between their outputs, increasing the performance of the forest as a whole. Each training sample is sent to the corresponding child depending on the result of the test. Comparison of a feature subset with a threshold continues iteratively till convergence. The convergence criteria for stopping the recursive comparison of feature values to a threshold are: 1) the number of samples in a node falls below a threshold; 2) a predefined maximum tree depth is reached; or 3) all the samples belong to the same class. In that case, the node becomes a leaf, and the most frequent class of the training data at the node is stored for testing.

During testing, a new sample is processed by applying respective tests according to the path from the root node to the leaf it traverses. When a leaf node is reached, the tree casts a vote corresponding to the class assigned to this node in the training stage. The final decision for a test sample is obtained by selecting the class with the majority of votes. Moreover, the probability that a test sample belongs to a class can be estimated as the fraction of votes for that class cast by all trees.

IV. IMAGE FEATURES

This section describes the features used in our method - intensity statistics, texture and curvature anisotropy, and spatial context features. Context features are a combination of intensity, texture and curvature values. VOI identification requires classification of supervoxels for which we use intensity, texture and curvature features, (excluding context information) to ensure fast feature extraction and subsequent classification, as well as good generalization of the classifier. For generating VOI probability maps we employ the complete set of features (including context information).

A. Intensity Statistics

A trained clinician can identify CD affected regions by examining T1 MR images due to their high resolution and high signal to noise ratio (SNR). Psychophysical experiments have established that the human visual system (HVS) is sensitive only to image features of the first and second order (mean and variance) [40]. However, MR images commonly contain regions that do not form distinct spatial patterns but differ in their higher order statistics, e.g. boundaries of some malignant tumours are diffuse and invisible to the naked eye [41]. In addition to features processed by the HVS, we propose to investigate features that are not discernible by the human eye but may provide discriminating information for our task. For every sub-volume (supervoxel or pixel neighborhood) we calculate the mean, variance, skewness and kurtosis of the intensity values.

B. Texture Anisotropy

Texture is modeled as patterns distinguished by a high concentration of localized spatial frequencies. To reduce the computational load, 2-D (instead of 3D) Gabor filter banks are used to generate texture maps for each slice. Gabor filters have optimal joint localization in the spatial and frequency domains. Their multi-scale and multi orientation structure conforms to the receptive fields profiles of simple cortical cells [42], and capture rich visual properties such as spatial localization, orientation selection and spatial frequency characteristics. Since Gabor filters incorporate Gaussian smoothing they are robust to noise. The Gabor filter bank can be represented as

\[
G_{\gamma,\omega}(x, y) = a^\gamma g(a^\gamma(x \cos(\omega) + y \sin(\omega))) \cdot \frac{a^\gamma(-x \sin(\omega) + y \cos(\omega))}{a^\gamma(x \cos(\omega) + y \sin(\omega))}
\]  

(1)
where $\gamma = 0, \cdots, \Gamma - 1$, $\omega = 0, \cdots, \Omega - 1$. The mother function of a Gaussian filter is defined as:

$$g(x, y) = \left(\frac{1}{2\pi\sigma_x\sigma_y}\right) \exp \left[-\frac{1}{2} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right) + 2\pi j(Wx + Vy)\right]$$

(2)

$\Omega = 6$ is the total number of orientations $(0^\circ, 30^\circ, 60^\circ, 90^\circ, 120^\circ, 150^\circ)$, $\Gamma = 2$ is the number of scales $(1, 0.5)$. The rotation factors $\psi = \pi/\Gamma$ and the scaling factor $a = (U_h/U_l)^{1/\Gamma - 1}$. $U_h$ and $U_l$ are parameters that determine the frequency range of the filter bank and $W, V$ are shifting parameters in the frequency domain. Experimentally we set $U_h = 0.1, U_l = 0.025, W = V = U_h, \sigma_x = 2, \sigma_y = 4$. The frequency of the sinusoidal function was set to 16.

In [43], texture anisotropy was used to identify tumorous regions in brain images. Normal tissues exhibit a regular pattern in their appearance while diseased regions show areas of anisotropy where mal regions in a VOI. While anisotropy in [43] was calculated using experimental results indicate that the normal is in the direction of the gradient, i.e., $\mathbf{n} = \nabla f = (f_x, f_y, f_z)$, which defines direction cosines of a local spherical polar coordinate system as

$$\mathbf{\hat{n}} = \left(\sin \theta \cos \phi \quad \sin \theta \sin \phi \quad \cos \theta\right) = \frac{1}{\sqrt{f_x^2 + f_y^2 + f_z^2}} \left(f_x \quad f_y \quad f_z\right)$$

(5)

An arbitrary direction $\mathbf{p}$ is chosen by taking the vector product of $\mathbf{\hat{n}}$ with the Cartesian basis direction $\mathbf{\hat{z}} = (0, 0, 1)$,

$$\mathbf{\hat{p}} = \frac{\mathbf{\hat{n}} \times \mathbf{\hat{z}}}{\mathbf{\hat{n}} \cdot \mathbf{\hat{z}}} = \left(-\cos \phi \sin \theta \quad -\sin \phi \sin \theta \quad \sin \theta \right) = \frac{1}{\sqrt{f_x^2 + f_y^2 + f_z^2}} \left(-f_x \quad f_y \quad f_z\right)$$

(6)

This direction lies on the $xy$ plane. A second tangent direction is calculated by

$$\mathbf{\hat{q}} = \frac{\mathbf{\hat{n}} \times \mathbf{\hat{p}}}{\mathbf{\hat{n}} \cdot \mathbf{\hat{p}}} = \left(-\cos \theta \cos \phi \quad -\sin \theta \cos \phi \quad \sin \theta \right)$$

(7)

Since $\mathbf{\hat{p}}$ and $\mathbf{\hat{q}}$ are orthonormal by construction, the first fundamental form $(F_1)$ is the identity matrix. The second fundamental form $(F_2)$ is identical to the Weingarten mapping matrix $H$. We calculate the following terms as

$$f_{pp} = \mathbf{\hat{n}} \cdot \frac{\partial \mathbf{\hat{p}}}{\partial \psi} f_{pq} = \mathbf{\hat{n}} \cdot \frac{\partial \mathbf{\hat{q}}}{\partial \psi} f_{qq} = \mathbf{\hat{n}} \cdot \frac{\partial \mathbf{\hat{q}}}{\partial \psi}$$

(8)

Each of the above terms is a curvature and $f_{pq} = f_{qp}$. The Weingarten mapping matrix $H$ is given by

$$H = F_2 = \left(\begin{array}{cc} f_{pp} & f_{pq} \\ f_{qp} & f_{qq}\end{array}\right)$$

(9)

The more commonly used Gaussian and Mean curvatures are given by

$$C_{curvGauss} = \text{Det}(H)$$

$$C_{curvMean} = \text{Trace}(H)$$

(10)

When healthy tissues are affected by progression of disease, it affects spatial arrangement of voxels and hence their curvature (and shape). Our aim is to exploit this irregularity for distinguishing between diseased and normal tissues. Curvature anisotropy is calculated in a similar manner as texture anisotropy. The entropy of curvature values is determined from 9 sectors of all slices. If the curvature values have a wide distribution it indicates greater anisotropy, leading to a higher entropy value. On the other hand low entropy values indicates less anisotropy. The anisotropy measure for a region $r$ is given by

$$\text{curv}_{ani} = -\sum_{\theta} p_{\theta} \log p_{\theta}$$

(11)

$p_{\theta}$ denotes the probability distribution of curvature values in sector $r, \theta$ denotes the curvature values. Similar to texture anisotropy, the curvature asymmetry measure is also a 9 dimensional feature vector for a region. The curvature asymmetry vector is denoted as $\text{Curv}$. Intensity, texture and curvature features combined give a 121 dimensional feature vector.
behavior is indicative of the fact that the curvature in diseased regions becomes distorted due to ulcerations or other abnormalities. Thus they lose the regularity observed in healthy tissues. This finding is corroborated by the plot in Fig. 1 (f), where patches of diseased tissue show higher entropy indicating greater randomness than patches of healthy tissue, and by the histogram of curvature values for the two patches in Fig. 1 (g).

The plots in Figs. 1 (e), (f) show examples where features of the two classes have good separability. It illustrates the effectiveness of our proposed features in distinguishing between different tissue types. However such clear separation in the feature space does not exist for all cases and a threshold value of one feature cannot easily separate the different classes. The use of multiple feature sets helps us to distinguish between different classes and Random forest classifiers are particularly effective in this scenario. The numbers in Table II also reflect this situation. Classification accuracy is low for individual features, but improves by combining different features.

**D. Spatial Context Features:**

Since the human anatomy displays a high degree of regularity with only moderate variations, presence of one organ provides a strong clue about the presence of another organ in medical images. Through appropriately designed features we aim to capture the contextual relationship between bowel tissues and other organs. Context features have been used to segment brain structures in MRI [45], the prostate from CT images [46], cardiac structures from MRI [47, 23], localizing anatomical structures [48] and segmenting the cardiac chamber [33]. Basically context features derive information of one set of objects from another set of objects. Figure 2 (a)-(c) shows slices from abdominal MRI. The diseased region is manually annotated in red, normal regions in green and background areas in yellow. Apart from the bowel, we also observe other anatomical structures like kidneys and liver, providing us with a rich set of sources for calculating context features.

Since contextual information depends on relative orientation and distance we sample regions at fixed positions from a pixel. Figure 2 (d) shows an illustration of the sampling scheme where the circle center is the pixel in question and the sampled points are identified by a red ‘X’. At each point corresponding to a ‘X’ we extract a $3 \times 3 \times 3$ region and calculate the mean intensity, texture and curvature values. The texture values were derived from the texture maps at $90^\circ$ orientation and scale 1. The ‘X’s are located at distances of 3, 8, 15, 22 pixels from the center, and the angle between consecutive rays is $45^\circ$. The values from the 32 regions are concatenated into a 96 dimensional feature vector.

**Template based steerable features for capturing context information** were used in [33]. Context features are important in medical images and it is quite intuitive to extract features from points at fixed positions from a given sample. Our approach is different from [33] in the use of features. While we use mean intensity, curvature and texture values from each sampled neighborhood, [33] use a 24 dimensional feature vector from the transformations of intensity and gradients. With fewer features than [33] we obtain similarly high segmentation accuracy (See Table IV and Section VIII-D).

**V. VOLUME OF INTEREST IDENTIFICATION**

The desired VOI is obtained by first oversegmenting the volume into supervoxels using the *simple linear iterative clustering* (SLIC) supervoxel algorithm [49], and classifying each supervoxel as either diseased or normal. This procedure returns a set of adjacent supervoxels containing the diseased VOI. The advantages of VOI selection are: 1) reduced computation time without analyzing every voxel for disease activity; and 2) initial selection of prospective diseased voxels reduces false positives in subsequent analysis. Supervoxels refer to subvolumes of the original image which are homogenous with respect to certain image features (intensity, texture, etc.). A comparison between different supervoxel methods as well as quantitative evaluation metrics can be found in [50].

**A. Supervoxel Segmentation**

Voxels are clustered according to texture similarity (instead of intensity as in the original method) and proximity in the image domain. Texture maps from oriented Gabor filters (e.g. $45^\circ$) give rise to more compact supervoxels than intensity maps. Since spatial distance and texture have different range of values, a modified distance measure ($d_{va}$ in Algorithm 2, Appendix A) is used to rescale them to the same range. The measure also considers the size of supervoxels. A schematic of the algorithm is given in Algorithm 2 (Appendix A).

*Modifications:* Prior to supervoxel segmentation, the images are bias-corrected using the method in [51]. Pixels with the lowest 5% intensities indicate noise, and are all set to zero. Similarly, higher intensity tails indicate artifacts and outliers. Hence the mean of the 5% brightest intensity values is taken to be the maximum image intensity. All intensity values are divided by this threshold value and voxels with intensity above this threshold are set to 1. This also increases the image contrast and highlights finer structures in the volume.

In the original algorithm, the intervals between supervoxels along each spatial dimension are equal so that the initial supervoxels are

---

**Fig. 1.** (a)-(b) 2D patch around diseased voxel and corresponding curvature map; (c)-(d) normal patch and corresponding curvature map; plot of entropy values for (e) texture; (f) curvature of slices shown in (a),(c); and (g) histogram of curvature values for diseased and normal patch.

**Fig. 2.** Example annotations and locations for deriving context information. (a)-(c) show different slices from the same patient with various annotations. Red contours show diseased tissues, green contours show normal tissues and yellow contours show background parts consisting of various organs. (d) shows locations from which context information is derived.
Fig. 3. Probability maps for ROI voxels. (a) Cropped showing ROI (red) and diseased annotations (green). Probability maps for (b) background; (c) normal; and (d) diseased (with colormap).

cube-shaped with equal side lengths. This choice does not work well in our application since the spatial extent in the \( z \) direction is short, has low resolution, and the structure of organs also changes significantly along \( z \)-axis. If supervoxels have too many slices, a single supervoxel may contain more than one organ, which complicates further analysis. For better spatial coherence we restrict the number of slices which supervoxels may cover. Instead of applying the same step size along three spatial axes, the average number of supervoxels on the \( xy \)-plane and along \( z \)-axis, denoted \( K_{xy} \) and \( K_z \), respectively, are specified separately so that the step sizes are calculated independently.

B. Supervoxel Classification for VOI identification

Supervoxel classification for VOI identification needs to be fast and accurate. In [10] we identified the importance of different features for detecting CD as described in Section IV. Out of the whole feature set we use intensity statistics, texture anisotropy and curvature asymmetry for identifying supervoxels with diseased tissue. Context features (Section IV-D) are used to generate probability maps as described in Section VI. From the manual annotations we can identify supervoxels that contain diseased, normal and background tissues. For every class of supervoxels we extract features and train an RF classifier with 50 trees which identifies supervoxels with diseased tissues.

VI. Segmenting Diseased Tissues

A probabilistic classifier like RF outputs three probability values (corresponding to three classes) for each VOI voxel. The RF classifier is trained using all the features described in Section IV - intensity, texture, curvature and context features - and is different from the classifiers used to classify supervoxels. Figure 3 (a) shows a cropped ROI (red border) of a slice with the manually identified diseased region shown in green. Note that the rectangular ROI is shown for illustration purposes, although the actual VOI need not have such a regular shape. Figures 3 (b)-(d) show the probability maps, respectively, for background, normal bowel and diseased bowel for the region of interest (ROI) (red outline) of Fig. 3 (a). The diseased area in Fig 3 (a) shows a corresponding high probability value in Fig 3 (d) which indicates the effectiveness of our features in classifying diseased regions.

If we use the probability maps directly for obtaining the final labels - i.e., highest probability value of a class is the desired label for a voxel - we get isolated clusters of voxels. Therefore we need to impose spatial smoothness constraints for spatial continuity using standard cost functions for segmentation. The final segmentation of diseased regions is obtained by formulating the segmentation as a labeling problem within a second order MRF cost function. The labels are obtained by optimizing the cost function using graph cuts [52]. A second order MRF energy function is written as

\[
E(L) = \sum_{s \in P} D(L_s) + \lambda \sum_{(s,t) \in N} V(L_s, L_t),
\]

where \( P \) denotes the set of pixels and \( N \) is the set of neighboring pixels for pixel \( s \). \( \lambda \) is a weight that determines the relative contribution of penalty cost \( (D) \) and smoothness cost \( (V) \), \( D(L_s) \) is given by

\[
D(L_s) = -\log (Pr(L_s) + \epsilon),
\]

where \( Pr \) is the likelihood (from probability maps) previously obtained using RF classifiers and \( \epsilon = 0.000001 \) is a very small value to ensure that the cost is a real number.

Semantic Information for Smoothness Cost: \( V \) ensures a spatially smooth solution by penalizing discontinuities. We formulate the smoothness cost by using semantic information from the second RF classifier. The RF classifier returns a measure of the importance of each dimension in the feature vector to the classification task. Despite of the multiple dimensional feature vector, the features are of three types - intensity, texture and curvature (context information is a combination of the three). By aggregating the importance values of each feature type and normalizing them we obtain their relative importance in the classification task. This relative importance is the necessary semantic information for voxel classification. Let the normalized weight (importance measures) of the different features be \( w_I \) (intensity), \( w_T \) (texture) and \( w_C \) (curvature), where \( w_I + w_T + w_C = 1 \). The smoothness cost \( V \) is given by

\[
V(L_s, L_t) = \begin{cases} 
    w_I V_I + w_T V_T + w_C V_C, & L_s \neq L_t, \\
    0, & L_s = L_t, 
\end{cases}
\]

where \( V_I, V_T, V_C \) are the individual contributions to the smoothness by intensity, texture and curvature. \( V_I \) is defined as

\[
V_I(L_s, L_t) = \frac{1}{{|r|}} \cdot \frac{1}{|t|},
\]

where \( r \) and \( t \) are the spatial coordinates. \( V_T \) and \( V_C \) are similarly defined using texture and curvature instead of intensity. Smoothness cost is determined over a 3 neighborhood system. The weights assume different values after every training stage because for every group of test patients we use a different set of training patients (as in 5 fold cross validation explained in Section VII). The weights take the following range of values: \( w_I = [0.19, 0.24] \), \( w_T = [0.31, 0.35] \), \( w_C = [0.41, 0.47] \). Thus there are no fixed values of \( w_I, w_T, w_C \) because the importance measures depend upon the training data. However we observe no significant change in the weights, and the relative importance of different features remains the same.

VII. Datasets

We use datasets from two different sources, one from the Amsterdam Medical Center (AMC) which we denote as AMC, and the other from University College of London Hospital (UCL) denoted as UCL. The datasets are described as follows:

AMC: The data used in this paper were taken from a prior study of 30 consecutively included patients with luminal Crohn’s disease that had been approved by AMC’s Medical Ethics Committee. All 30 patients had given informed consent to the prior study. Furthermore, 26 out of 30 patients gave written consent to usage of their data.
for future investigations. The data of the latter 26 patients (mean age 37 years, range, 23.4 – 58.8 years, 16 females) were used for our work. Patients fasted four hours before a scan and drank 1600 ml of Mannitol (2.5%) (Baxter, Utrecht, the Netherlands) one hour before a scan. Images were acquired with patients in supine position using a 3-T MR imaging unit (Intera, Philips Healthcare, Best, The Netherlands) with a 16-channel torso phased array body coil. The protocol consists of axial and coronal single shot fast spin echo (SSFSE) sequences followed by a coronal fat-saturated SSFSE sequence and coronal 3D T1-weighted spoiled gradient echo sequence (SPGE). The pixel spacing of the images was 1.02 mm × 1.02 mm × 2 mm, and the volume dimension was 400 × 400 × 100 pixels.

UCL: The data were acquired from 19 patients (mean age, 31.0 years, range, 16.4 – 51.1 years, 9 females) diagnosed with small bowel Crohn’s disease who had undergone MRI enterography within 2 weeks (mean 4 days) of elective small bowel surgical resection. Patients had to fast 4 hours before a scan and they underwent MR enterography following ingestion of 1000 – 1500 ml of 0.2% locust bean gum and 2% mannitol solution in order to produce more detailed images of the small intestine. Images were acquired with patients in prone position after 20 mg of spasmolytic (Buscopan; Boehringer Ingelheim, Ingelheim, Germany) using a 1.5T MR imaging unit (Avanto; Siemens, Erlangen) with manufacturer’s body and spine array coils. Coronal volumetric interpolated breath-hold examination (VIBE) acquisitions were performed at 90 and 70 s post injection of 10 ml gadopentetate dimeglumine (Magnevist; Berlex Laboratories, Montville, NJ). The spatial resolution of the images was 1.02 mm × 1.02 mm × 2 mm. The images have various dimensions. Two of them have dimension of 512 × 416 × 48, one 512 × 416 × 64, one 512 × 512 × 56 and the rest 512 × 512 × 48. Ethical permission was given by the University College London Hospital ethics committee, and informed written consent was obtained from all participants. Various sized annotations on the images by research physicians specialized in abdominal MRI indicate whether the covered regions are diseased or normal.

7 separate patient datasets (3 from AMC and 4 from UCL) were used to determine the value of λ (Eqn. 12). These volumes have not been part of the test dataset described above. On an average each patient had 10 – 20 slices of annotations and each slice had 2 – 4 annotated regions. Our whole pipeline was implemented in MATLAB on a 2.66 GHz quad core CPU running Windows 7. However our code was not optimized to take advantage of parallel processing. The random forest code was a MATLAB interface to the code in [53] written in the R programming language. Details on computation time are given in Section VIII-H. The 45 patients were divided into 5 groups of nine patients each, and we employ a 5 fold cross validation approach where 4 groups are used for training and the remaining patient group is used for testing the algorithm. Each patient was part of the test group exactly once. The tree depth was fixed at 20.

A. Evaluation Metrics

The quality of our segmentations was evaluated using two measures: 1) Dice Metric (DM) and 2) Hausdorff Distance (HD). DM measures the overlap between the segmented diseased region obtained by our algorithm and reference manual annotations. It is given by

\[ DM = \frac{2 |A \cap M|}{|A| + |M|}. \]  

where \( A \) - segmentation from our algorithm and \( M \) - manual annotations. The DM measure yields values between 0 and 1 where high DM corresponds to a good segmentation [54].

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>10</th>
<th>5</th>
<th>1</th>
<th>0.5</th>
<th>0.1</th>
<th>0.01</th>
<th>0.02</th>
<th>0.05</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>71.3</td>
<td>71.4</td>
<td>76.1</td>
<td>79.3</td>
<td>82.2</td>
<td>83.9</td>
<td>88.4</td>
<td>88.4</td>
<td>88.4</td>
</tr>
</tbody>
</table>

Table I

CHANGE IN SEGMENTATION ACCURACY WITH DIFFERENT VALUES OF \( \lambda \) (Eqn. 12).

Hausdorff Distance (HD): The DM gives a measure of how much the actual manual segmentation was recovered by the automatic segmentation. But the boundaries of the segmented regions may be far apart. The HD aims to measure the distance between the contours corresponding to different segmentations. If two curves are represented as sets of points \( A = \{a_1, a_2, \ldots \} \) and \( M = \{m_1, m_2, \ldots \} \), where each \( a_i \) and \( m_j \) is an ordered pair of the \( x \) and \( y \) coordinates on a point on the curve, the distance to the closest point (DCP) for \( a_i \) to the curve \( M \) is calculated. The HD, defined as the maximum of the DCP’s between the two curves, is:

\[ HD(A, M) = \max(\max_i \{DCP(a_i, M)\}, \max_j \{DCP(m_j, A)\}). \]  

VIII. RESULTS AND DISCUSSION

A. MRF regularization strength \( \lambda \) (Eqn. 12)

To choose the MRF regularization strength \( \lambda \) we adopt the following steps. We choose a separate group of 7 patient volumes, and perform segmentation using our method but with \( \lambda \) taking different values from 10 to 0.001. The results are summarized in Table I. The maximum average segmentation accuracy using Dice Metric (DM) was obtained for \( \lambda = 0.02 \) which was fixed for subsequent experiments. Note that these 7 datasets were a mix of patients from the two hospitals and were not part of the test dataset used for evaluating our algorithm.

B. VOI identification

Accurate VOI identification is an important part of our pipeline. Errors in this step often cause inaccurate segmentations of diseased regions. VOI detection can be seen as a classification problem where we need to identify those supervoxels that contain diseased tissues. As described previously the entire dataset is divided into 5 groups of 9 patients each to facilitate 5 fold cross validation. Each patient volume is segmented into supervoxels and features are extracted from supervoxels corresponding to the annotated regions for three classes - diseased bowel, normal bowel and background. Since we know the voxels included in the annotations we can easily identify each supervoxel as either diseased, normal or background. If a supervoxel has even one diseased voxel it is denoted as diseased, while normal (background) supervoxels have all normal (background) voxels.

Over all the 45 patients we have 791 diseased, 743 normal and 732 background supervoxels. For each training cycle approximately 4/5 of the samples are used while the rest are used for testing. The number of samples of each class is approximately equal in the training and testing stage. Note that each sample denotes a supervoxel and is greater than the number of annotations. The sample size can be varied by varying the size of supervoxels. However, the supervoxel size also influences the classification accuracy. A discussion about the dependence of supervoxel size and classification performance is presented in Section VIII-C.

Table II summarizes the classification performance for different feature combinations. All Features indicates the combination of intensity, texture and curvature features. Note that context features were not used for supervoxels to reduce computation time and ensure good generalization of the RF classifiers. AccD refers to
the percentage of diseased samples that were correctly classified. In this stage we require a higher $\text{Acc}_D$ even at the expense of lower classification accuracy for normal ($\text{Acc}_N$) and background ($\text{Acc}_B$). A large number of ‘false positives’, i.e. samples from other classes identified as diseased is not a major hindrance because they get labeled as normal or background in the subsequent segmentation stage.

As expected the accuracy for the individual features are lower than their combinations. The combination of texture and curvature features produces results closest to All Features. However, this does not indicate that intensity information is unimportant. Conducting a $t$-test on the values for $\text{Tex}$ + $\text{Curv}$ and All Features gives $p < 0.012$ which indicates statistically different results. Further, we also conduct $t$-tests for features $\text{Tex}$ versus $\text{Int}$, and $\text{Curv}$ versus $\text{Curv} + \text{Int}$. In all cases we find that $p < 0.3$, thus clearly showing that inclusion of intensity features improves classification accuracy without extra computational cost.

In any classification scheme it is difficult to get 100% classification/detection accuracy. A similar situation in this scenario can have adverse effects in CD diagnosis. Misclassification of diseased supervoxels generally occurs when the number of diseased voxels in a supervoxel is very low. Hence the extracted features are more representative of other classes.

To overcome this shortcoming, we adopt the following strategy in VOI detection. Since diseased tissues in an image are mostly contiguous, the diseased supervoxels are bunched together in clusters of two or more. After supervoxel classification we choose those clusters with two or more supervoxels thus excluding false positives away from the region of interest. In some cases the diseased colon may be smaller than two supervoxels (first row of Fig. 4) and hence none of the candidate diseased regions have two or more supervoxels. In that case we choose the largest supervoxel. Then we assign corresponding labels to the voxels within it. We change the labels of all supervoxels adjoining such a cluster to diseased (irrespective of their originally assigned labels). This label propagation allows us to include some ‘diseased’ supervoxels that may have been missed in the initial classification.

Figure 4 shows an example where this strategy is particularly effective. The first column shows consecutive slices from a volume where the diseased colon is shown in red and the supervoxels are shown in green. The second column shows the initially selected diseased supervoxels (in yellow) and the third column shows the detected ROI in each slice after selecting the largest supervoxel or the clusters with two or more supervoxels overlaid on the manually annotated diseased colon (in red), thus removing false positives. Small diseased parts are missed by the supervoxel classification scheme. However when we include all the neighboring supervoxels, the ROI encompasses all possible diseased voxels. The fourth column of Fig. 4 shows the new ROIs for each slice which now includes all possible diseased voxels. An effect of including neighboring supervoxels is an increase in the number of voxels in the VOI leading to extra computation time. However this is a small price to pay for an accurate VOI identification which leads to accurate disease diagnosis.

Table II (AllFeat) gives quantitative measures for supervoxel classification before changing the labels of neighboring supervoxels. With the change of labels the following values are obtained: $\text{Acc}_D = 98.8\%$, $\text{Acc}_N = 83.7\%$ and $\text{Acc}_B = 85.7\%$. The labeling of neighboring supervoxels increases ‘false positives’ i.e. more non-diseased (primarily normal) supervoxels being labeled diseased. It reduces number of diseased supervoxels labeled non-diseased (‘false negatives’), and increases number of diseased labeled as diseased (‘sensitivity’). However the number of non-diseased correctly labeled as non-diseased (‘specificity’) reduces. Classification accuracy of background supervoxels is not significantly affected because there are not many neighboring the VOI.

For all the 45 patients we did not encounter a single case where the diseased tissues were completely undetected. A part of the diseased colon was always detected by our supervoxel classifier. Inclusion of the neighboring supervoxels helps to reduce the ‘false negatives’ (diseased regions identified as non-diseased). Although the false positives increase (since many normal voxels are included in the VOI), they are correctly labeled in the subsequent segmentation stage. The diseased regions are found in clusters of more than three, while false positives consist usually of one or two supervoxels. Thus, selecting the single largest cluster excludes false positives away from the diseased colon.

To incorporate contextual information for VOI classification the previously described sampling template for pixels is not very effective. First, we need not use all supervoxels (corresponding to the red ‘X’ in Fig. 2 (d)) because many of them will fall outside the image domain. In that case we may use only the neighboring supervoxels. The supervoxel segmentation does not give a regular lattice of supervoxels, i.e. the centers of neighboring supervoxels are not necessarily at fixed angles. Even if we were to consider supervoxels whose centers lie at fixed positions from the center of the current supervoxel, in many cases more than one center lies within the same neighboring supervoxel. Consequently the extracted features will provide redundant information. A variable template relying upon distinct neighboring supervoxel centers may give inconsistent contextual information. Thus in order to include consistent contextual information we need to devise a complex heuristic that takes into account distance from the current supervoxel, distinct neighboring supervoxel centers, and other factors which would only increase the computational complexity. Moreover our current scheme gives high accuracy in identifying the relevant supervoxels, and hence we do not adopt a time consuming approach.

### C. Effect of Supervoxel Size

Small supervoxels tend to be more homogenous and the extracted features are more representative of a single class than large supervoxels. However they may not always provide sufficient number of voxels to estimate stable features. Large supervoxels mostly contain enough voxels to calculate statistically stable features but they may contain voxels from more than one class which generates label ambiguities. Consequently, the extracted features may not be representative of one class. Thus training with features from very large supervoxels leads to low classification accuracy. Table III summarizes the performance for different supervoxel sizes in terms of $\text{Acc}_D$, $\text{Acc}_N$ and $\text{Acc}_B$. Our experiments clearly demonstrate that an empirically optimal tradeoff between accuracy and homogenous samples is achieved when the

### Table II

| Feature | $\text{Acc}_D$ (%) | $\text{Tex}$ | $\text{Curv}$ | $\text{Tex} + \text{Int}$ |
|---------|------------------|-------------|-------------|----------------|----------------|
| AllFeat | 72.1±2.3         | 76.4±2.1    | 78.8±2.7    | 78.7±1.3       |
| Tex     | 70.3±1.3         | 73.9±1.2    | 74.2±1.9    | 76.2±1.1       |
| Curv    | 70.4±1.6         | 73.8±1.7    | 74.0±2.3    | 75.8±2.3       |
| $\text{Curv} + \text{Int}$ | 79.2±2.4 | 82.8±1.3 | 90.3±2.9 |
| $\text{Curv} + \text{Tex}$ | 76.4±2.3 | 80.1±2.8 | 87.9±2.2 |
| AllFeatures | 76.1±1.8 | 79.8±2.2 | 87.1±2.9 |

**TABLE II**

Quantitative measures for supervoxel classification under different feature combinations. $\text{Acc}_D$ - classification accuracy for normal supervoxels. $\text{Acc}_N$ - classification accuracy for normal supervoxels. $\text{Acc}_B$ - classification accuracy for background supervoxels. $\text{Tex}$-texture, $\text{Curv}$-curvature.
number of voxels in a supervoxel is in the range 1800 – 2200. Depending upon the volume dimensions we set supervoxel parameters \( K_{xy} \) and \( K_z \) such that the resulting supervoxels contain appropriate number of voxels.

### D. Segmentation of Diseased Colon

Once the VOI has been identified, the contained diseased colon is segmented using graph cut segmentation. Table IV shows segmentation results for the following methods 1) \( \text{Our} \): our proposed method; 2) \( \text{Our}_{nc} \): \( \text{Our} \) without context information from images; \( w_I, w_T, w_C \) are weighted according to their importance measures learned from the training step. 3) \( \text{Our}_{S} \): \( \text{Our} \) without semantic information in \( V \) but using context information; \( w_I = w_T = w_C = 0.33 \) 4) \( \text{Our}_{S} \): \( \text{Our} \) with \( V_I = 0 \); 5) \( \text{Our}_{SC} \): \( \text{Our} \) with \( V_T = 0 \); 6) \( \text{Our}_{SC} \): \( \text{Our} \) with \( V_C = 0 \). For 4, 5, 6 above the weights are normalized by the sum of the two values. Figure 5 (a), (b) shows the DM and HD values of the individual datasets from \( \text{AMC} \) (red bars) and \( \text{UCL} \) (blue bars). The lowest DM value is 0.81 (for \( \text{AMC} \) Dataset 5, discussed in Section VIII-G) which was a particularly challenging case because of noise and complex structures. However we achieve higher DMs for all other datasets with the values ranging between 0.85 – 0.95. Similarly, the HD values vary between 6 – 9 mm for all datasets.

\( \text{Our} \) performs the best among all methods followed by \( \text{Our}_{S} \) and \( \text{Our}_{SC} \). From the results in Table IV we draw the following conclusions. First, comparing the results of \( \text{Our} \) and \( \text{Our}_{nc} \) we observe that context features are integral to our method’s performance and their exclusion maximally degrades the classification performance. \( \text{Our}_{SC} \) does not include any context information and hence misses out on important context cues that go a long way in distinguishing between different organs. In the absence of context information many voxels get biased probability values, leading to lower segmentation accuracy. If we use the 24 dimensional feature vector obtained from intensity and gradient transformations as in [33] we obtain an average DM = 90.6. In comparison to \( \text{Our} \) these results are statistically similar \((p > 0.16)\). Thus we are able to obtain similar accuracy with fewer features that capture important discriminative information.

Second, semantic information is also important. \( p < 0.025 \) from \( t \)-tests between \( \text{Our} \) and \( \text{Our}_{S} \) supports this observation. \( \text{Our}_{S} \) weights intensity, texture and curvature equally in the smoothness cost, thereby neglecting the semantic information learned from the trained classifier. This lowers the DM and increases the HD compared

<table>
<thead>
<tr>
<th>( N )</th>
<th>500 – 1000</th>
<th>1000 – 1500</th>
<th>1500 – 1800</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Acc}_{D} ) (%)</td>
<td>77.9±2.1</td>
<td>81.1±2.4</td>
<td>83.5±2.7</td>
</tr>
<tr>
<td>( \text{Acc}_{N} ) (%)</td>
<td>75.3±2.8</td>
<td>78.0±2.1</td>
<td>80.8±3.2</td>
</tr>
<tr>
<td>( \text{Acc}_{I} ) (%)</td>
<td>74.7±2.8</td>
<td>77.6±2.6</td>
<td>80.9±2.7</td>
</tr>
<tr>
<td>( \text{Acc}_{D} ) (%)</td>
<td>90.3±2.9</td>
<td>81.2±1.4</td>
<td>79.4±3.3</td>
</tr>
<tr>
<td>( \text{Acc}_{N} ) (%)</td>
<td>87.9±2.2</td>
<td>81.1±2.9</td>
<td>77.9±3.4</td>
</tr>
<tr>
<td>( \text{Acc}_{I} ) (%)</td>
<td>87.1±2.9</td>
<td>80.5±3.2</td>
<td>77.2±2.6</td>
</tr>
</tbody>
</table>

**TABLE III**

**Quantitative measures for different supervoxel sizes.**

Fig. 4. VOI detection. First column shows the supervoxels (green) and manually annotated diseased colon (red). Second column shows those supervoxels initially identified as diseased (yellow). Third column shows diseased supervoxels after choosing the largest cluster or clusters with two or more supervoxels (in yellow), and fourth column shows the final VOI after neighboring supervoxels of Column 3 are changed labels. The three rows show results of consecutive slices.
to Our, although the magnitude is less than Our$_{nC}$. Particularly for the bowel it is difficult to distinguish different tissues from only the intensity values in MR images. Learned semantic information helps to distinguish between neighboring voxels from different categories.

The two stage classification results in [10] helped us identify the relevant features and the classifier (RF) best suited for CD detection. We also employed ad-hoc post processing to remove outliers and obtain a segmented mask. However it is not a segmentation method as such. Our proposed method in this paper obtains higher DM values than in [10] because of many additional features like automatic VOI identification, inclusion of semantic information and graph cut segmentation.

Comparing between Our$_{nV}$, Our$_{nVC}$ and Our$_{nVC}$ highlights the relative importance of different features. Curvature and texture features are deemed most important in the classification, although intensity also plays a significant role. t-tests between the results yields p-values $p < 0.035$ for all cases indicating that none of the features can be discarded without a significant drop in performance, and their combination achieves the best results. Table V summarizes the relative performance of different methods in terms of the statistical significance of their results.

Figures 6,7 show segmentation results on Patient 11 (UCL) and Patient 20 (AMC) which were particularly challenging because: 1) these patients suffer from multiple diseased colon sections; and 2) in both cases there exist very narrow diseased colon; 3) in AMC Patient 20 the diseased region has similar intensity values as the neighboring normal regions. It poses challenges primarily for graph cut segmentation. Our method is able to accurately segment the region of diseased colon sections because of the inclusion of context information and semantic information. Excluding this information leads to inferior segmentation as demonstrated in the corresponding figures.

We show results for Our (green), Our$_{nC}$ (yellow), Our$_{nV}$ (magenta), Our$_{nVC}$ (cyan). Using our method we are able to achieve an accurate segmentation (DM= 0.92 and HD = 7.0). Although only a single slice is shown the measures are for the whole volume. Our$_{nC}$ shows the maximum over segmentation, i.e., many healthy tissues are labeled as diseased because of the absence of context information. The segmentation of the different methods is consistent with the values observed in Table IV. Since there are no other studies specific to Crohn’s disease we are unable to compare our algorithms performance with others. However we obtain similar DM values as of recent works on lesion segmentation [17, 16].

<table>
<thead>
<tr>
<th></th>
<th>Our</th>
<th>Our$_{nC}$</th>
<th>Our$_{nV}$</th>
<th>Our$_{nVC}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>90.4</td>
<td>80.6</td>
<td>83.7</td>
<td>84.4</td>
</tr>
<tr>
<td>HD</td>
<td>7.4</td>
<td>12.8</td>
<td>8.6</td>
<td>9.2</td>
</tr>
<tr>
<td></td>
<td>Our$_{nV}$</td>
<td>Our$_{nVC}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DM</td>
<td>83.0</td>
<td>82.8</td>
<td>86.5</td>
<td></td>
</tr>
<tr>
<td>HD</td>
<td>10.2</td>
<td>11.0</td>
<td>9.2</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE IV**

Quantitative measures for segmentation accuracy. DM: Dice Metric in % and HD: Hausdorff Distance in mm.

<table>
<thead>
<tr>
<th></th>
<th>Our - Our$_{nC}$</th>
<th>Our - Our$_{nV}$</th>
<th>Our - Our$_{nVC}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td></td>
<td>Our - Our$_{nV}$</td>
<td>Our - Our$_{nVC}$</td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>$&lt; 0.01$</td>
<td>$&lt; 0.001$</td>
<td>$&lt; 0.005$</td>
</tr>
</tbody>
</table>

**TABLE V**

Comparison between different methods for statistical significance of their results.

E. Varying Number of Training Patients

We also investigate the influence of the training data size on the segmentation performance. In one set of experiments (Expt 1) we separated the AMC and UCL data. The two RF classifiers were trained on datasets varying from one to $N_T - 1$ ($N_T = 26$ for AMC, $N_T = 19$ for UCL) and tested on the remaining datasets. Figure 8 shows the corresponding DM values on AMC (Expt 1 AMC) and UCL (Expt 1 UCL). The DM values stabilize after a certain $N_T$, although there is no fixed fraction of the total number of datasets. The purpose of training data is to capture the variations in the real world data. In the case of AMC 17 datasets were sufficient to capture the variations in the real world data, while for UCL 15 datasets proved sufficient. However, it is possible that the classifiers trained on one dataset may not perform very well on another dataset.

In a second set of experiments (Expt 2) we train on one hospital data and test on the other hospital dataset. The segmentation result from this experiment is shown in Table VI. The classification accuracy is lower than what is reported in Table II, although $Acc_D$ and DM both are greater than 85%. Training on datasets from both scanners ensures that the classifier learns certain inherent characteristics of the imaging process. On the other hand training on images from one scanner may not always generalize well to other scanners. This experiment is more representative of the real world scenario where different hospitals may use different sets of scanners to obtain their data. Nevertheless we are able to achieve DM values between $85 – 89\%$ in such cases which indicates that our feature extraction methodology is robust to different scanner protocols.

The DM values in Expt 1 (after stability) are closer to the values reported in Table IV although significantly different ($0.037 < p < 0.059$), while Expt 2 gives much lower values than Table IV ($p < 0.027$). This deviation can be explained by the fact that the results in Table IV are an outcome of a mixed training set which better captures the variation in disease anatomy and scanner protocols than the original training sets. Expt 1 and Expt 2 are able to partially capture these factors, which combined with lesser training...
Fig. 6. Segmentation results for Patient 11 (UCL). The manual annotations are shown in red with other colours showing the following annotations: green-Our (Column 1), yellow-Our\_nC (Column 2), magenta-Our\_nV (Column 3), cyan-Our\_nVC (Column 4). Each row shows results for different slices of the same volume.

<table>
<thead>
<tr>
<th>Acc_D</th>
<th>Acc_N</th>
<th>Acc_B</th>
<th>DM</th>
<th>HD</th>
</tr>
</thead>
<tbody>
<tr>
<td>87.8</td>
<td>64.8</td>
<td>84.4</td>
<td>87.9</td>
<td>8.0</td>
</tr>
</tbody>
</table>

**TABLE VI**
Quantitative measures for different tests. Training data consists of datasets from one hospital and test data comprises of datasets from the other hospitals.

data contributes to inferior performance compared to the original experiments using all 45 datasets.

In a third set of experiments, we randomly choose 15 patients each from AMC and UCL as our dataset and we segment them using 5-fold cross validation as described previously. The average DM values were 87.5 ± 3.4 which is lower than that in Table IV ($p < 0.03$). From the above three experiments we draw the following conclusions: 1) more training data improves segmentation performance as they capture the variations in real world cases; 2) comparatively more training data is required for test data from a different scanning protocol to achieve performance similar to data from the same scanner; and 3) mixing the data from both hospitals leads to better generalization in testing.

**F. Robustness to Noise**

Zero mean Gaussian noise of different variance ($\sigma$) was added to the original images after intensity normalization, and the diseased regions were segmented using the previously described steps of VOI detection, probability map generation and segmentation. The trained classifiers were the same as used before, i.e., new classifiers were not trained on images with simulated noise. The motivation was to investigate whether the current set of features, although trained on one set of images, are robust for increased noise levels on a new test set of images. The segmentation performance of Our under different noise levels is summarized in Table VII and visual results shown in Fig. 9. With increasing noise levels the performance is degraded up to the point when DM goes below 70%. The classifiers learned from training data are robust up to noise of $\sigma = 0.05$. Beyond that the performance drops significantly due to indistinguishable regions. Consequently the feature extraction is inaccurate which leads to poor segmentation.

Fig. 8. DM values for different number of training data from the same hospital.
Fig. 7. Segmentation results for Patient 20 (UCL). The manual annotations are shown in red with other colours showing the following annotations: green—Our (Column 1), yellow—Our$_{UC}$ (Column 2), magenta—Our$_{nV}$ (Column 3), cyan—Our$_{nVC}$ (Column 4). Each row shows results for different slices of the same volume.

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>DM</th>
<th>HD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>87.2</td>
<td>9.3</td>
</tr>
<tr>
<td>0.005</td>
<td>81.2</td>
<td>14.7</td>
</tr>
<tr>
<td>0.01</td>
<td>79.1</td>
<td>13.6</td>
</tr>
<tr>
<td>0.05</td>
<td>76.3</td>
<td>16.9</td>
</tr>
<tr>
<td>0.1</td>
<td>73.2</td>
<td>18.1</td>
</tr>
</tbody>
</table>

**TABLE VII**

Quantitative measures for segmentation accuracy of Our under different degrees of noise. DM: Dice metric in % and HD: Hausdorff distance in mm.

G. Algorithm Limitations

Accurate VOI localization is an important part of our method. If the identified VOI is not accurate then the final segmentation is poor. Figure 10 (a),(c) shows two consecutive slices for one such volume (AMC dataset 5) where the identified VOI does not include large parts of the diseased area. The identified supervoxels that are part of the VOIs are shown in yellow. The final segmentation results are shown in Figs. 10 (b),(d). Since the VOI does not include the entire diseased colon section, it negatively affects the final segmentation. Complex tissue structure, noise levels that hamper accurate feature extraction and low resolution contribute to supervoxel misclassification in some images. But inclusion of neighboring supervoxels overcomes these effects to some extent. However in this particular case, many diseased supervoxels are not neighbors of the initial classified diseased supervoxels. Thus the overall segmentation accuracy is low.

H. Computational Cost

Our method basically consists of the following steps: supervoxel segmentation and classification to get the VOI, analyzing every voxel within the VOI for disease activity by generating probability maps and graphcut segmentation. Supervoxel segmentation of a single volume is quite fast and takes about 30 seconds. However the subsequent classification of each supervoxel takes a lot of time because of the large number of samples. The VOI identification requires between 21 – 28 minutes. The subsequent step to generate probability maps is more time consuming as features have to be extracted for each voxel. Depending upon the size of the VOI, a further 24 – 27 minutes are required to get the probability maps. Thus on an average the entire method from start to finish takes 47 – 57 minutes because of being coded in MATLAB. In future we aim to explore other possibilities of reducing analysis time through more efficient coding or deploying parallel processing.

For each of the 45 datasets we have available the manual annotations to extract features (supervoxel and individual voxel). They are stored separately and depending upon the training set, the features of the appropriate patient datasets are accessed for training the RF classifier. There are a total of 2266 supervoxels for each class which is an average of 51 supervoxels per patient. To train the supervoxel classifier for 36 patients (4/5 of the patient data is used for training) using the code in [53] we require 8 minutes on an average (approximately 1836 samples of 121 dimensions). The RF classifier for generating probability maps requires 33 minutes to be trained due to the significantly high number of voxel wise samples. Care is taken to ensure that there are approximately equal number of samples from each class for both classifiers. Both the classifiers are trained offline, and their training time does not include the test times reported previously.
We arrived at this training. Table VIII shows the training time ($T$) (c) magenta-Our. With increasing $N$ in Table VIII in terms of DM values. When $N_T > 50$ there is no significant increase in DM ($p > 0.1$) but the training time increases significantly. The best trade-off between $N_T$ and DM is achieved for 50 trees and is the reason to choose such a size of our RF ensemble.

### IX. Conclusion and Future Work

In this work we have proposed a pipeline to segment tissues affected by Crohn’s disease from a test MRI volume. Given a test volume we first oversegment it using supervoxel segmentation, and classify each supervoxel with random forests classifiers and intensity, texture and curvature features. These steps yield an approximate VOI encompassing the diseased tissues. Each voxel within the VOI is further analyzed for disease activity. Probability maps are first generated for each VOI voxel using intensity, texture, curvature and context features. These maps give the probability of a voxel being diseased, normal or background tissue, and are obtained by a second set of random forest classifiers. The negative log-likelihood of the probability values is the penalty cost for a graph cut segmentation framework.

For spatial smoothness constraints we make use of semantic information from the second set of trained RF classifiers. The RF classifiers give a measure of the importance of each feature in the classification task. The importance measures are aggregated for different feature types and normalized to get a set of weights for intensity, texture and curvature. The smoothness cost is designed to incorporate this semantic information where intensity, texture and curvature differences of neighboring pixels are weighted by the values obtained from semantic information.

Experimental results on 45 patient datasets affected with CD show that our method is able to achieve a high level of segmentation accuracy. We also analyze the importance of individual features, segmentation performance under different degrees of added noise, and different parameters. Results highlight the robustness of our method. One shortcoming of our method is the large computation time. In future work we aim to address this issue with the help of more efficient code as well as possible parallel implementation of different stages of our pipeline.

### Appendix

The supervoxel segmentation algorithm consists of three steps: cluster centres initialisation, iterative clustering and enforce connectivity and is summarized in Algorithm 2. Details of the whole algorithm can be found in [49].

### Acknowledgment

The authors thank the European Community’s Seventh Framework Programme for funding the VIGOR++ Project, and also thank the anonymous reviewers for their very helpful and constructive feedback. The authors also thank ETH Zurich, AMC Amsterdam, TU Delft and UC London for their support in carrying out all research related to this work.
Algorithm 2 SLIC Supervoxels

Input Image with \(N\) voxels, desired number of supervoxels \(K\)

Output Supervoxel labels of all voxels \(m\)

1. Initialised cluster centres \(C_k = [x_k, y_k, z_k]^T\) sampled at \(\sqrt{N/K}\) regular lattice steps \(S = \sqrt{N/K}\).
2. Move cluster centres to the lowest gradient position in a \(3 \times 3 \times 3\) neighbourhood.
3. Set label \(l(v) = -1\) for each voxel \(v\).
4. Set distance \(d(v) = \text{inf}\) for each voxel \(v\).

/* Iterative clustering */

repeat
  for each cluster centre \(C_k\) do
    for each voxel \(v\) in a \(2S \times 2S \times 2S\) neighbourhood centred at \([x_k, y_k, z_k]^T\) do
      Compute distance \(d_{vk}\) between \(C_k\) and \(v\):
      \[d_{vk} = \sqrt{\sum_{k=1}^{3}(x_k-y_k)^2} + \sqrt{\sum_{k=1}^{3}(x_k-z_k)^2}.\]
      \((T)\) denotes texture values; \(x_k, y_k\) and \(z_k\) are the spatial coordinates of the cluster centre; \(m, S\) are constants.
      if \(d_{vk} < d(v)\) then
        Set \(l(v) = k\)
        Set \(d(v) = d_{vk}\)
      end if
    end for
  Compute new cluster centres.
  Compute residual error \(E\).
until \(E \leq \text{threshold}\) or maximum number of iterations is achieved

/* Enforce connectivity */

for all voxels \(v\) which have not been assigned a new label do
  \(m \leftarrow \text{Flood Fill}(v, l)\)
  if the connected component containing \(v\) has less than \(N/K\) voxels then
    Merge it into one of its neighbour connected components
  else
    Assign a new label to the connected component
  end if
end for
return \(m\)

REFERENCES


