Detection of ulcerative colitis lesions from weakly annotated colonoscopy videos using bounding boxes

Safaa Al-Ali, John Chaussard, Sébastien Li-Thiao-Té, Éric Ogier-Denis, Alice Percy-Du-Sert, Xavier Treton, Hatem Zaag

Detection of ulcerative colitis lesions from weakly annotated colonoscopy videos using bounding boxes

SUBMISSION DATE / POSTED DATE

15-11-2023 / 22-11-2023

CITATION

Al-Ali, Safaa; Chaussard, John; Li-Thiao-Té, Sébastien; Ogier-Denis, Éric; Percy-du-Sert, Alice; Treton, Xavier; et al. (2023). Detection of ulcerative colitis lesions from weakly annotated colonoscopy videos using bounding boxes. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.24566263.v1

DOI

10.36227/techrxiv.24566263.v1

Detection of ulcerative colitis lesions from weakly annotated colonoscopy videos using bounding boxes

Safaa Al-Ali, John Chaussard, Sébastien Li-Thiao-Té, Éric Ogier-Denis, Alice Percy-du-sert, Xavier Treton, and Hatem Zaag

***Abstract*— Ulcerative colitis is a chronic disease characterized by bleeding and ulcers in the colon. Currently, the gastroenterologist reviews the colonoscopy video to assess the disease severity using an endoscopic score. This task is time-consuming and does not consider the size and the number of lesions. Consequently, automatic detection methods were proposed enabling fine-grained assessment of lesion severity. However, they depend on the quality of the training set, and its specificity to the application context. To suit the local clinical setup, we opted for an internal training dataset containing only rough bounding box annotations around lesions. Color information is the primary indicator used by specialists to recognize the lesions. Thus, we propose to use linear models in suitable color spaces to detect lesions. We introduce an efficient sampling scheme for exploring the set of linear classifiers and removing trivial models i.e. those showing zero false negative or positive ratios. Using bounding boxes leads to exaggerated false negative/positive ratios due to mislabeled pixels, especially in the corners, resulting in decreased models' accuracy. Therefore, we propose to evaluate the model sensitivity on the annotation level instead of the pixel level. Our sampling strategy can eliminate up to 25% of trivial models. Despite the limited annotations' quality, the detectors achieved good performance (93% specificity/89% sensitivity for bleeding and 57% specificity/83% sensitivity for ulcers). The best models exhibit low variability when tested on a small subset of endoscopic images. However, the inter-patient model performance was variable suggesting that appearance normalization is critical in this context.**

***Index Terms*— Bleeding, bounding box annotation, lesions detection, model selection, sensitivity, ulcer, ulcerative colitis.**

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment.

S. A.A. was with Université Sorbonne Paris Nord (USPN), Laboratoire Analyse, Géométrie et Applications, LAGA, CNRS, UMR 7539, F-93430, Villetaneuse, France. She is now with Centre Inria d'université Côté d'Azur (Epione Team), 2004 Rte des Lucioles, 06902, Valbonne, France (e-mail: safaa_alali@hotmail.com).

J. C. is with USPN (e-mail:Chaussard@math.univ-paris13.fr)

S. L.T.T is with USPN (e-mail:lithiao@math.univ-paris13.fr)

E. O.D. is with institut national de la santé et de la recherche médicale-INSERM, Paris, France (e-mail:eric.ogier-denis@inserm.fr)

X. T. is with Hôpital Beaujon Gastrentérologie et assistance nutritive, Clichy, France (e-mail:xtreton@gmail.com)

A. P.S. is with Hôpital Beaujon Gastrentérologie et assistance nutritive, Clichy, France (e-mail:alicepercie dusert@hotmail.fr)

H.Z. is with USPN (e-mail: hatem.zaag@math.cnrs.fr).

I. Introduction

INFLAMMATORY bowel diseases (IBDs) are chronic inflammatory illnesses, in which the lining of the bowel becomes inflamed and presents lesions such as bleeding and ulcers [1]. The two main forms are Crohn's disease (CD) and Ulcerative colitis (UC) both of which may cause serious discomfort and long-term complications for the affected patients.

Proper diagnosis and management of these conditions are essential to improve the patient's quality of life and prevent further complications. Colonoscopy [2] and Wireless Capsule Endoscopy (WCE) [3] are the methods of reference for evaluating and monitoring IBDs severity, and hence making treatment decisions and assessing treatment response. These techniques allow direct visualization of the inner lining of the gastrointestinal tract. More precisely, a colonoscopy uses a flexible thin hose equipped with a mini-video camera and is performed by an experienced clinician for UC. On the other hand, WCE uses an embedded, pill-sized, camera that can be swallowed, and is more suited to the diagnosis of CD.

Bleeding and ulcers are common lesions associated with both diseases, UC and CD. Color information is the primary indicator used by specialists to distinguish between mucosal lesions and the surrounding normal or healthy mucosa. In particular, bleeding lesions usually show dark red areas whereas ulcers appear as white spots on the gut wall, both distributed with diverse shapes and sizes. Currently, experts review manually colonoscopy or WCE videos, which can represent around 10,000 frames. This process is a hard and time-consuming task leading to only considering the characteristics of the most severe lesions.

Therefore, automated lesion detection can bring significant benefits by improving the reproducibility of severity assessment and decreasing the physicians' burden as well. Many methods have been proposed to automatically detect endoscopic bleeding and ulcer lesions. In this light, computer analysis techniques try to solve two kinds of problems. First, given a set of pixels, a region of interest (ROI) or a complete frame, binary classification algorithms are used to select a label between "lesion" (also called abnormal) or "not lesion" (also called normal). Second, segmentation algorithms are used to find regions with the same label such as bleeding or ulcers regions. Numerous automatic detection methods

used color features to train their models [4]–[8], while others combine it with texture information to enhance detection performance [9]–[11].

The presented work focuses on bleeding and ulcer detection from colonoscopy videos obtained in the context of UC disease during an ongoing collaboration with the Bichat-Beaujon Hospital in Paris, France. Although the lesions' appearance is similar between UC and CD, we anticipate that the currently available methods are biased to their training set, i.e. to well-delineated video-capsule images ([12]–[16]) as opposed to complete colonoscopy videos obtained on the local instruments. Consequently, we built a custom dataset of colonoscopy videos of real patients, called *Vatic* specific to this collaboration. To minimize the burden of the annotation process, we propose that doctors use an interface inspired by the Vatic software [17] that allows the delineation of the lesions by bounding boxes instead of a precise delineation of lesion boundaries. Classical machine learning or Convolutional Neural Network (CNN) approaches for lesion detection usually depend on the quality of the training dataset, which naturally affects the accuracy of the detector. Consequently, using complex machine learning algorithms such as CNNs in our case may generate unsatisfactory detection results. Therefore, we decided to alternatively employ linear models. Due to their simplicity, these models can provide valuable insights and interpretable results. Following [7], [8], [12], [18], [19], we decided to use linear models in convenient color spaces for bleeding and ulcer detection. We also propose an efficient sampling scheme to explore the set of linear models that rejects trivial classifiers that classify all the pixels into the same class. Since the bleeding and ulcer lesions are of variable and complex geometric shapes, their delimitation by bounding boxes is quite imprecise. Indeed, the pixels surrounding the lesion were included in the annotation although they are healthy pixels. To deal with this problem, we propose to take into account annotation errors to compute the sensitivity of the detector. Specifically, we consider the mislabeled pixels within the bounding box annotations as correctly identified abnormal pixels rather than considering them as false negatives. Finally, we analyze the performance of the best models across the initial training and additional testing datasets by considering only small subsets of endoscopic images.

The contributions of this paper are threefold:

- Proposition of a sampling strategy to effectively explore the set of linear models by only considering nontrivial models.
- Introduction of a sensitivity performance that can deal with bounding box annotations imprecision enhancing detection models accuracies.
- Study of the variability of the detectors across the patients, even inside the training set. Our study shows that the models used are not universal and personalized models should be developed for each patient.

The rest of the paper is organized as follows: in Section II, we present an overview of the current state-of-the-art methods proposed for UC lesions detection. Then, in Section III, we propose a classification method based on linear models computed using pixel color features. Additionally, we propose an efficient sampling scheme to explore the set of linear models that rejects trivial classifiers. Next, we introduce performance criteria that can deal with bounding box annotation problems. In Section IV, we demonstrate that our proposed method achieves good-quality detection of bleeding and ulcers in the ROC space. To prepare for clinical validation, we evaluate the accuracy of our performance estimates on a set of small subset of endoscopic images. We show that the proposed classifiers exhibit good performance, and yield reliable results, but that most of the variability is indeed related to patients' variability (see Section V).

II. Related work

A. Automatic detection of bleeding

Most of the current methods perform classification in a color space with maximum contrast between bleeding and nonbleeding regions. As bleeding pixels are red, it is natural to consider detection and classification in the RGB colorspace, or direct transformations of RGB ([6], [7], [18], [20]–[22]).

In 2011, Fu *et al.* [20] trained a 3-layer perceptron on the ratios (R/G,R/B, R/G+B+R) for each pixel and applied morphological erosion. Later in 2014, [6], the authors extended their approach by working with superpixel regions, and a Support-Vector Machine (SVM) classifier trained on 60,000 pixels. In the same year, Ghosh *et al.* [18] applied a K-Nearest Neighbors (KNN) classifier to statistical parameters extracted from the R/G histogram. The authors reported that the combination of only three parameters, namely {median, variance, kurtosis} was sufficient to identify bleeding frames with an accuracy of 98.5%. This work was later extended in [7] by working on 7 pixels $\times$7 pixels blocks.

Some bleeding detection algorithms work on the histogram bin levels instead of the pixel values [18], [21]. Kundu *et al.* [21] computed Regions of Interest (ROIs) defined by the color ratios $r/b \geq m$ and $r/g \geq n$ computed in the normalized RGB color space, denoted by rgb and applied a KNN classifier to 64 histogram bins in the green channel. The parameters $m = 2.8$ and $n = 2$ are chosen according to the maximal accuracy of pixel detection compared to the ground truth provided for 65 endoscopic images. In [22], the authors combined the RGB values into a single number with bit concatenation and applied an SVM classifier on the bins of the resulting histogram. In [12], the authors used a similar technique before PCA dimension reduction and classification with KNN.

Other color spaces were also considered in [5], [23], [24]. In [23], the authors used an SVM classifier with statistical features computed in Luma In-phase Quadrature (YIQ) color space. Deeba *et al.* [5] merged two SVM classifiers built from statistical features extracted respectively from RGB and Hue-Saturation-Value (HSV) color histograms. In [24], the authors trained a three-layer probabilistic neural network on statistical features from RGB and Hue-Saturation-Intensity (HSI) pixel intensities. Recently, Pogorelov *et al.* [9] proposed to consider image texture besides color. They used RGB color features

and 22 texture parameters extracted from the grey-level co-occurrence matrix. The authors tested many classification methods and found that the SVM classifier performed best.

B. Automatic detection of ulcers

Ulcers show as pinkish white, which explains why most methods focus on detecting bright pixels [8], [25]. In [25], the authors trained an SVM classifier on statistical features in RGB and CIElab (Lab) spaces and concluded that (L, a, G) channels give the best detection performance. The authors later extended their work in [8] with more colorspaces (RGB, HSV, YCbCr, CMYK, YUV, CIElab, XYZ) and found that (Cr,Y,B) is the best feature combination.

Ulcers also appear as rough surfaces which can be detected based on texture features [10], [11], [19]. In [11], the authors proposed to combine color (S from HSV and M from CMYK) and Leung-Malik filters [26] with an SVM classifier. In [19], the authors applied an SVM classifier to statistical moments of the Contourlet transform and Log Gabor filter in HSV and YCbCr color spaces. In Yeh *et al.* [10], the textural features were obtained from the Grey Level Co-occurrence Matrix. Different combinations of the number of features, feature selection algorithm, and classification algorithm were compared, and the best combination was obtained with decision trees, with 40 features selected by the ReliefF method.

III. Materials and Method

A. Colonoscopy videos dataset

From *Vatic* database, we used 5 videos (768 pixels×576 pixels) containing both bleeding (1629 frames) and ulcer (1760 frames) annotations for training, for a total of 4349 frames (see Table I). Each video was annotated by gastroenterologists with the help of the Vatic software [17].

TABLE I: Number of frames used for training: number of frames with bleeding annotations, number of frames with ulcer annotations, and total number in the video.

	Bleeding frames	Ulcer frames	Total number of frames
video 1	671	554	812
video 2	224	378	378
video 3	254	86	1116
video 4	140	204	910
video 5	340	538	1133
Total	**1629**	**1760**	**4349**

B. Proposed method

Our proposed method involves several steps outlined in Fig. 1. First of all, we remove all black pixels surrounding the informative pixels. Next, we compute the color histograms of healthy pixels. We thus propose an effective sampling method to explore the linear models. We also adjust the computation of the sensitivity criteria to encounter mislabeled pixels occurring during the annotation process using bounding boxes. Finally, we optimize the performance of the detectors utilizing the Youden index [27]. In what follows, we detail the process by showing some examples.

1) Image preprocessing: Due to the camera's field of view, only an octagonal portion of the image is actually recorded in the endoscopic video, and the outer portions are set to black (see Fig. 2). Additionally, some embedded textual information should be removed prior to bleeding or ulcer detection. Consequently, we detect pixels with small grey-level variance and grow the detected region with morphological dilation (5x5 square structuring element). Additionally, some unannotated areas are bright because of light shining on wet spots (specular reflection), so we remove the pixels $\mathbb{1}_{\{Y>c\}}$, with $c = 150$ chosen by visual inspection.

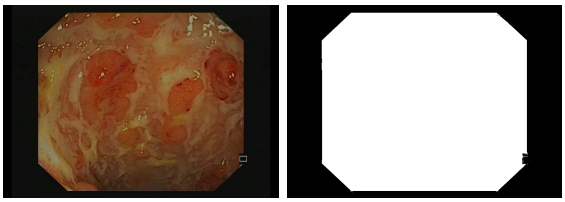

Fig. 2: Example of an endoscopic frame (on the left) and corresponding binary mask (on the right) used later to remove pixels that do not correspond to the colon wall during the training stage of the detectors.

2) Definition of bleeding and ulcer detectors: We previously pointed out that in colonoscopy videos, bleeding show as red patches and ulcers as pinkish-white patches on the gut wall (Fig. 3). As previous authors [7], [12], [18] have shown that the R/G ratio is relevant (it leads to 11% overlap in [7]) to detect bleeding, we consider linear classifiers in the (R,G) subspace, i.e. $\{aR + b \geq G \text{ for } (a, b) \in \mathbb{R}^2\}$. Similarly, following [8], [19], we consider linear classifiers in the (Cr,Y) subspace obtained using Cr and Y channels from YCbCr and CMYK color spaces respectively, i.e. $\{aCr + b \leq Y \text{ for } (a, b) \in \mathbb{R}^2\}$ to detect ulcer lesions. This corresponds to finding a straight separation line between the histograms of normal and lesions' pixels. Let's take the example of endoscopic figures given in Fig. 3. The best bleeding detector should lead to a minimum overlap ratio between normal (Fig. 3c) and bleeding pixels (Fig. 3e) in the (R, G) color space. On the other hand, the best ulcer detector should lead to a minimum overlap ratio between normal (Fig. 3d) and ulcer pixels (Fig. 3f) within the (Cr, Y) color space.

3) Proposed sampling strategy: The model search process consists of exploring all the linear models of the color spaces (R, G) and (Cr, Y) for bleeding and ulcer detection respectively. In Fig 4, we give the histograms of normal pixels of the training dataset (Table I). For each histogram, we plot a set of 100 random linear models. We can remark that classifiers that do not "cross" the histograms, herein highlighted in orange color, are trivial because they give the same label to all pixels. In particular, no normal pixel will be correctly identified by the detector and consequently, the true negative rate of this detector will be zero. To study the amount of these trivial models, we ran a series of 100 trials, each involving 100 randomly generated lines. The results show that when sampling linear models in (R,G), also denoted by RG, color space, an average of 9% of these models is "trivial" with a standard deviation

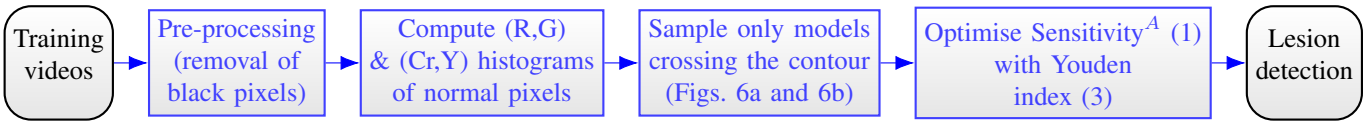

Fig. 1: Flowchart of the proposed method.

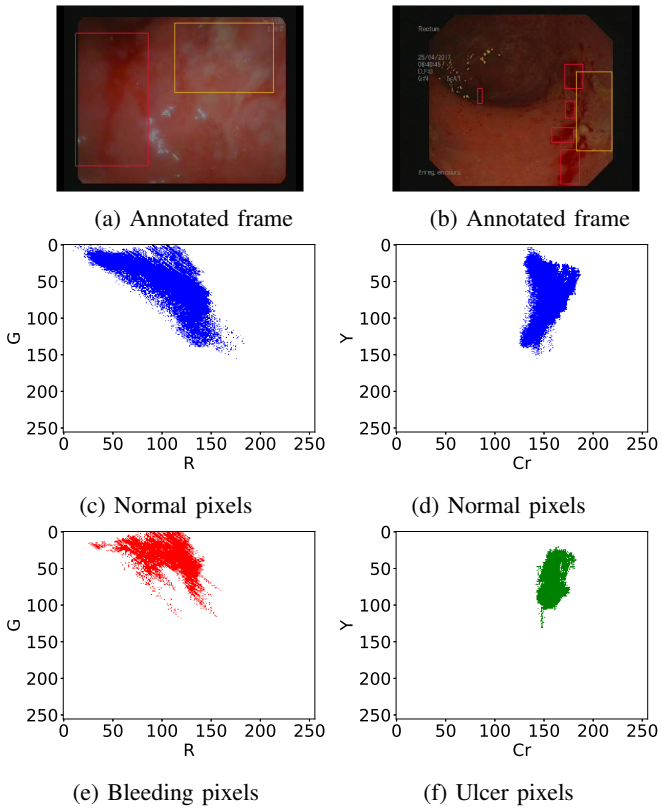

Fig. 3: (a,b) Annotated frames with bleeding (red bounding box) and ulcers (orange bounding box). Corresponding histograms of the normal pixels i.e. all pixels out of the bounding boxes (c-d), bleeding pixels (e), and ulcer pixels (f).

(std) of around 3%. In contrast, for the (Cr,Y), also denoted by CrY, space this amount increases significantly to achieve an average of 25% with a std of about 4%. When sampling the RG space using 10,000 random models, among them 9.41% are trivial whereas this number increases to 25.62% in the case of sampling CrY space. Therefore, we decided to eliminate these models and restrict the optimization space to the set of random linear classifiers that go through the interior of the histogram. Since the number of trivial models remains almost the same by testing more than one hundred models, we decided to restrict the search for lesion detectors by testing only one hundred random linear models.

Additionally, if a line goes through the interior, it must cross the boundary of the set. We can avoid sampling redundant linear classifiers by focusing on the contour of the histogram instead of its interior. To sample the set of lines, we will thus draw two points in the contour of the RG and CrY histograms and consider the associated linear classifiers (see Fig. 6).

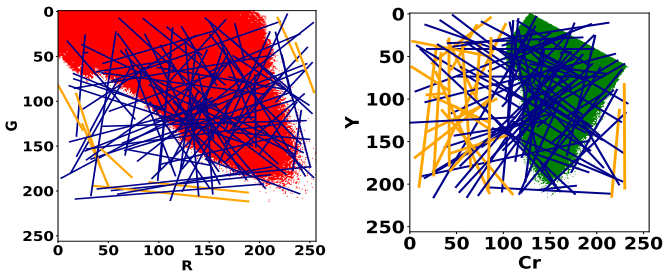

Fig. 4: Histograms of normal pixels for the training dataset (cf Table I) in the RG space (on the left) and CrY space (on the right). Among the set of 100 random linear models, 8% do not cross the RG histogram in opposite to 23% for the CrY histogram. Trivial models are represented in orange color.

4) Proposed performance metric of the detectors: In textbook statistics, the specificity $\frac{\sum TN}{\sum TN+\sum FP}$ or true negative (TN) rate measures the proportion of normal pixels correctly identified as such, and sensitivity $\frac{\sum TP}{\sum TP+\sum FN}$ or true positive (TP) rate measures the correctness of abnormal pixels detection. However, evaluating specificity and sensitivity hinges on reliable pixel annotations by gastroenterologists. Unfortunately, as previously discussed, the gastroenterologists' annotations in our database contain many errors because the regions of interest are provided as bounding boxes, whereas bleeding and ulcers have more complex shapes (see Fig. 5). Direct observation also suggests that many dark red pixels were not labeled as bleeding, and white pixels were not labeled as ulcers. Consequently, we expect over-inflated levels of FP and FN based on the database annotations. This will hide the correct classifier, and decrease the confidence in our results.

In Fig. 5, we illustrate the results of bleeding detection (in red) using a chosen random linear model, $G \leq 0.3R + 1$ and ulcer detection (in orange) using the linear model $Y \geq 0.5Cr + 8$. We report the performance metrics of the models in terms of TP, TN, FP, and FN computed on the pixel level in Table II. It can be seen that the model is able to correctly identify most of the annotated pixels (see the last row). However, as gastroenterologist annotations are usually wider than the actual lesion, some annotated pixels are not detected by our algorithm, then the false negative ratio is very high resulting in decreased sensitivity values (cf Table II).

To overcome these problems, we modify the definition of sensitivity to take the labeling problems into account. The pixels inside an annotation and not detected as such should not count as false negatives when assessing the algorithm's performance. Consequently, we will count all pixels belonging to an annotation as TP, as soon as one pixel is detected inside. As pixels inside an annotation are either true positives or false negatives, this corresponds to counting "detected annotations"

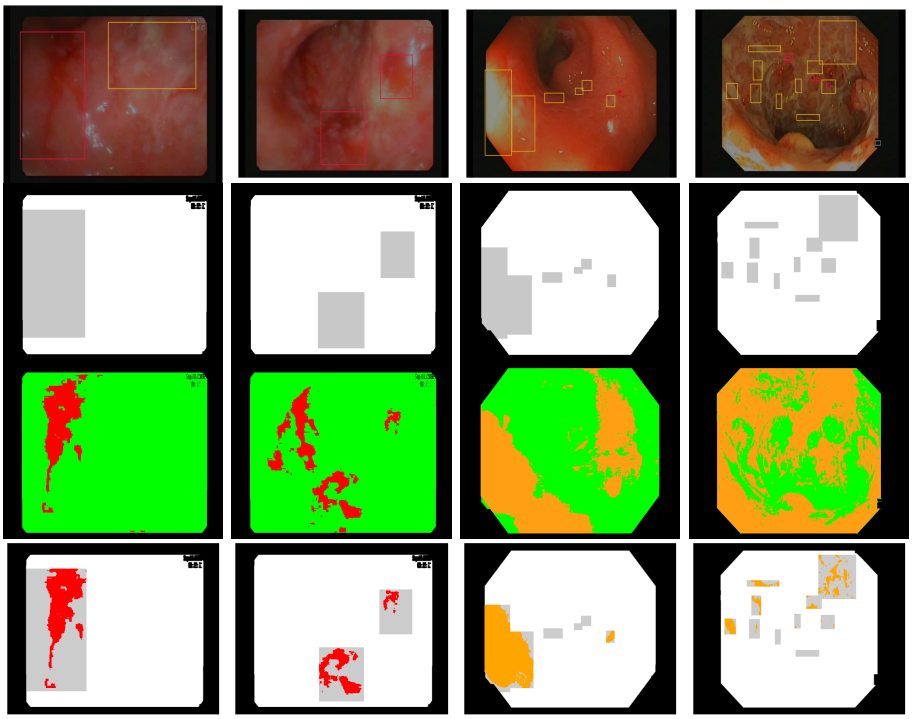

Image A Image B Image C Image D

Fig. 5: Endoscopic images extracted from *Vatic* dataset (first row). The second row represents the mask highlighting the ground truth obtained by the bounding boxes annotations (in gray). The third row shows the results of bleeding detection (in red) using the linear model $G \leq 0.3R + 1$ and ulcer detection (in orange) using the linear model $Y \geq 0.5Cr + 8$. The last row shows the intersection between the models' detection and the ground truth.

TABLE II: Performance results for endoscopic images given in Fig. 5. TPA represents the number of pixels within the detected annotations and PA denotes the total number of pixels of all the annotations presented in the frame.

Image identity	TP	TN	FP	FN	TPA	PA	Specificity	Sensitivity	Sensitivity^A
Image A	23229	234006	2161	66942	93936	93936	99.08%	25.76%	100 %
Image B	11181	263133	16183	35841	47022	47022	94.81%	23.78%	100 %
Image C	36692	238390	18724	10619	46318	50616	92.72%	77.55%	91.51%
Image D	8556	192292	54952	30270	36041	38982	77.77%	22.04%	92.46%

instead of "detected pixels". More precisely, we count in terms of "area", and define the sensitivity criteria as follows:

$$\text{Sensitivity}^A = \frac{\text{Area of detected annotations}}{\text{Total area of annotations}}. \quad (1)$$

In comparison with the standard sensitivity criteria, Sensitivity^A may provide a compromise between bounding box annotations and the detector's ability to correctly identify them (cf Table II). Specificity was not modified, as we expect missing annotations to represent a small number of pixels relative to nonannotated pixels $\sum TN + \sum FP$.

Finally, the detector performance is measured in a sensitivity vs (1-specificity) plot or Receiver Operating Characteristic (ROC) space. As we are only interested in single detectors, each detector's performance is represented by a point. The ideal classifier corresponds to the upper left corner. Other good models are a compromise between sensitivity and specificity and are close to (0,1). We select the classifier that maximizes the Youden index [27]:

$$\hat{m} = \operatorname*{argmax}_m d_{\text{ROC}}(\{y = x\}, m), \quad (2)$$

$$= \operatorname*{argmax}_m \left(\text{Sensitivity}^A + \text{Specificity}(m) - 1\right). \quad (3)$$

IV. Results

A. Best lesions detectors

As explained in Section III-B.3, we take a random sample of size 100 from the set of linear models that cross the contour of the histogram of normal pixels. Fig. 6 shows the sampled models in histogram space and in ROC space.

Table III shows the performance of the three best linear models in terms of specificity, Sensitivity^A, and standard sensitivity. As shown in Fig. 6, the models achieve good performance results in ROC space, i.e. specificity and Sensitivity^A. Fig. 7 (in the 2nd and 4th row) shows that there is a good visual agreement between the colors of detected lesions and the expert annotations. The best linear models can focus on the relevant areas rather than the total annotation, and select candidate ROIs that were not annotated. As expected, the detected areas do not overlap "fully" with the annotations, which is the reason for the low standard sensitivity levels. Based on the 3 best models, we estimate that around 90% of bleeding annotations are incorrect, and 80% of the ulcer annotations (see Table III). As a result, training with the standard sensitivity would provide nonsensical models, whereas we can achieve good performance with Sensitivity^A.

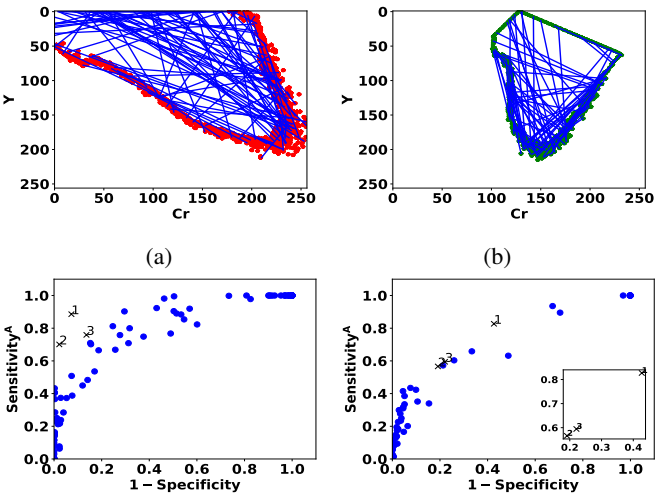

Fig. 6: 100 linear classifiers are sampled by drawing two points from the contour of the (R,G) histogram of normal pixels (left) and (Cr,Y) histogram (right). The performance of the models for bleeding (left) and ulcer (right) is shown in ROC space. The three best models are shown in black.

TABLE III: Performance of the best linear models for bleeding and ulcer detection. Good performance is obtained based on Sensitivity^A, but standard sensitivity is low due to annotation errors.

Best models for bleeding	Specificity	Sensitivity^A	Sensitivity
$G \leq 0.298R - 1.03$	92.71%	88.58%	9.56%
$G \leq 0.264R - 4.837$	97.79%	70.08%	4.16%
$G \leq -0.066R + 31.58$	86.39%	75.92%	13.79%
Best models for ulcers	Specificity	Sensitivity^A	Sensitivity
$Y \geq 0.698Cr - 42.799$	57.33%	82.71%	38.85%
$Y \geq 0.505Cr + 8.816$	80.88%	56.62%	14.27%
$Y \geq 0.499Cr + 6.318$	77.91%	59.46%	17.17%

In Table IV, we summarize the results of our models compared to two methods found in the literature. As we are interested in detecting annotations, we apply two simultaneous color ratios as done in [21] and find the optimal parameters $\hat{m} = 5.95$ and $\hat{n} = 3.75$ to detect bleeding ROIs. The KNN algorithm has not been further employed as the authors did. On the other hand, for ulcer detection, an SVM model with an RBF kernel and 10-fold cross-validation was trained on our dataset using two color bands Cr and Y as done in [8]. Based on a grid search within the values range $(-8, 7, 6, \ldots, 6, 7, 8)$, we find that the optimal parameters are $C = 0.79$ in terms of regularization constant and $\gamma = 3.03$ in terms of kernel hyper-parameter. We then computed the detection performance for both resulted models on our training dataset (cf Table I) using the standard specificity and the proposed Sensitivity^A (cf section III-B.4). Reported results show that linear models exhibit better compromise between specificity and Sensitivity^A compared with [8] and [21]. SVM model fails on abnormality detection, here the ulcers found in *Vatic*. We thus tried to make data augmentation on the ulcer pixels to maintain a balance in the training dataset, but Sensitivity^A remained low.

In Fig. 7, we present some annotated frames with the corresponding detection using our models as well as the models computed based on [8] and [21]. We find that our best linear models show better compromise between the detection of healthy pixels and lesions pixels than the other methods.

TABLE IV: Performance of the best lesions detectors compared to the literature.

Models	Specificity	Sensitivity^A
Proposed bleeding detector	**92.71%**	88.58%
Linear model-Kundu [21]	16.68%	**99.84%**
Proposed ulcer detector	57.33%	**82.71%**
SVM algorithm-Suman [8]	**99.84%**	21.14%

V. Discussion

As discussed in Section II-A, the RGB color space, and especially the Red and Green channels, has previously been used successfully for bleeding detection, whereas the YCbCr color space was used for ulcer detection. The information present in the pixel color is not altered by a change of color space, but a suitable color space presents this information more straightforwardly, and dimension reduction methods such as PCA can automatically perform this. In this manuscript, choosing the right colorspace based on the previous literature (see [5], [9], [13], [22] for bleeding and [8], [11], [19] for ulcers) enables us to work with 2D linear models instead of 3D models.

The use of bounding box annotations in our database (see Fig. 3) entails a considerable quantity of ground truth errors because annotations do not match the arbitrary and complicated shapes of the lesions. This is a major difficulty in our context, regardless of the type of model or machine-learning approach. To ease the annotation burden, semi-automatic region selection algorithms have been proposed. In the work of Sainju *et al.* [28], the authors use the growing region algorithm [29] to create homogeneous bleeding regions from consecutive capsule endoscopy frames. A seed is manually selected by the user and then enlarged by adding 8-connected neighbors, and the new centroid is taken as the seed for the following frame. This method extracts only one region per lesion, which can unbalance the normal and bleeding regions in the training database. In addition, it does not perform well in the absence of lesions due to forward and backward camera movements or in patients with mild forms of UC. In [5], the authors use a similar method to extract the bleeding regions but keep only a single frame rather than the complete sequence.

In this paper, we propose to adjust the performance criterion of lesion detection rather than automatically annotate the dataset. We chose to work with linear models instead of more sophisticated approaches in order to provide results that are easy to interpret and use in clinical practice. In addition, the good performance obtained in this and previous studies ([5], [7], [8], [11], [18], [19]) suggests that clinical validation of the approach is the critical step, as opposed to more sophisticated approaches such as SVM or neural networks.

To evaluate the validity of our results, we did not perform cross-validation, but show the results of computing specificity and Sensitivity^A on a random subset of frames in each video in Fig. 8. Cross-validation selects random subsets and finds the

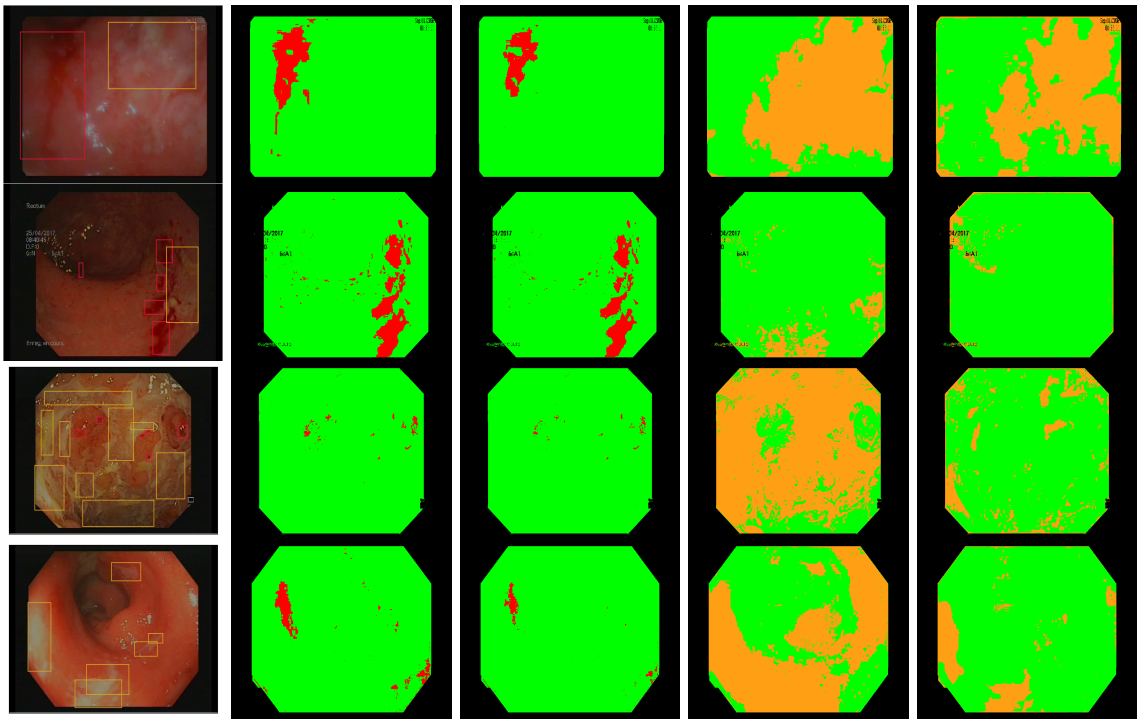

Fig. 7: Annotated frames from *Vatic* database (first column), bleeding detection with our best linear model (second column), bleeding detection by two simultaneous linear models [21] (third column), ulcer detection with our best linear model (fourth column), ulcer detection using SVM [8] (last column).

best model for each subset. Consequently, it selects different models at each run and evaluates the performance of the optimization algorithm. For clinical practice, we are interested in the performance of specific models, their reliability, and their generalization to new patients. Fig. 8 shows the performance of the 3 best models for the patients in the training dataset (left) and 5 new patients (right). For each patient, we estimated specificity and Sensitivity^A on 20 random subsets of a video, each containing 10% of the frames. Only three points are drawn, but the size of the ellipses is computed from the standard deviations of the 20 subsets. Fig. 8 shows that specificity and Sensitivity^A are estimated precisely, even on a fraction of the frames. This suggests that computational time can be reduced by using only a small subset of the video. However, the performance varies a lot between patients, even inside the training set. This means that the selected models are not universal and that specific models should be trained for each patient. This observation was not reported in previous works because the datasets used contain frames that are not organized "by patient". Consequently, methodological advances are necessary to make colonoscopy videos comparable, in order to apply trained models to new patients.

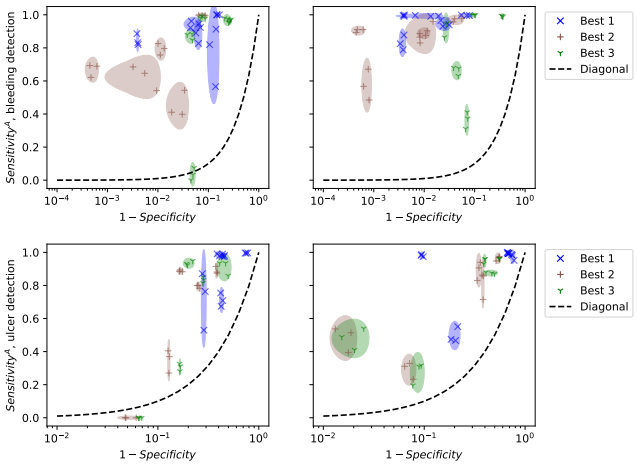

Fig. 8: Performance of the 3 best linear models depending on the patient, 5 training videos (left), and 5 test patients (right).

VI. Conclusion

This paper studies the automatic detection of bleeding and ulcers in colonoscopy videos for UC severity assessment based on a training dataset containing many annotation errors. We decided to deal with the annotations problem rather than proposing a sophisticated machine learning algorithm to improve detection performance as done by current studies. As in previous studies, we explore the set of linear classifiers and propose an efficient optimization method based on sampling the contour of the color histogram. This allows us to eliminate around 25% of trivial models which leads to focusing only on interesting models i.e. those giving nonzero true negative and true positive ratios. By adjusting the definition of sensitivity, we can circumvent the effect of the annotation errors using bounding boxes, and select good pixel-level lesion detectors. The best linear models obtain 93% specificity / 89% sensitivity for bleeding detection and 57% specificity / 83% sensitivity for ulcer detection, and reliable performance estimates can be

obtained from random subsets of the dataset. These show that the best detectors achieve good performance, but that patient-to-patient variability is dominant in this problem, and that further procedures to normalize the appearance of the videos are needed for clinical applications.

Patients' Consent

The patients' videos were anonymous, and analyzed after obtaining their consent. The study was approved by the local research study committee.

ACKNOWLEDGMENT

This work was performed in the context of the Investissements d'Avenir programme ANR-11-IDEX-0005-02 and 10-LABX-0017, Sorbonne Paris Cité, Laboratoire d'excellence INFLAMEX. Safaa Al-Ali received funding from the Paris Region Fellowship Programme, attributed by the DIM Math-Innov. We express our sincere thanks to Eric Ogier-Denis and Xavier Treton from CRI, Inserm 1149 and APHP Beaujon Hospital for their insight during this collaboration.

References

[1] S. Jäger, E. F. Stange, and J. Wehkamp, "Inflammatory bowel disease: an impaired barrier disease," *Langenbeck's archives of surgery*, vol. 398, no. 1, pp. 1–12, 2013.

[2] F. Probert, A. Walsh, M. Jagielowicz, T. Yeo, T. D. Claridge, A. Simmons, S. Travis, and D. C. Anthony, "Plasma nuclear magnetic resonance metabolomics discriminates between high and low endoscopic activity and predicts progression in a prospective cohort of patients with ulcerative colitis," *Journal of Crohn's and Colitis*, vol. 12, no. 11, pp. 1326–1337, 2018.

[3] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, no. 6785, pp. 417–417, 2000.

[4] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[5] F. Deeba, M. Islam, F. M. Bui, and K. A. Wahid, "Performance assessment of a bleeding detection algorithm for endoscopic video based on classifier fusion method and exhaustive feature selection," *Biomedical Signal Processing and Control*, vol. 40, pp. 415–424, 2018.

[6] Y. Fu, W. Zhang, M. Mandal, and M. Q. . Meng, "Computer-aided bleeding detection in wce video," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 2, pp. 636–642, 2014.

[7] T. Ghosh, S. A. Fattah, K. A. Wahid, W.-P. Zhu, and M. O. Ahmad, "Cluster based statistical feature extraction method for automatic bleeding detection in wireless capsule endoscopy video," *Computers in biology and medicine*, vol. 94, pp. 41–54, 2018.

[8] S. Suman, F. A. Hussin, A. S. Malik, S. H. Ho, I. Hilmi, A. H.-R. Leow, and K.-L. Goh, "Feature selection and classification of ulcerated lesions using statistical analysis for wce images," *Applied Sciences*, vol. 7, no. 10, p. 1097, 2017.

[9] K. Pogorelov, S. Suman, F. Azmadi Hussin, A. Saeed Malik, O. Ostroukhova, M. Riegler, P. Halvorsen, S. Hooi Ho, and K.-L. Goh, "Bleeding detection in wireless capsule endoscopy videos—color versus texture features," *Journal of applied clinical medical physics*, vol. 20, no. 8, pp. 141–154, 2019.

[10] J.-Y. Yeh, T.-H. Wu, W.-J. Tsai, *et al.*, "Bleeding and ulcer detection using wireless capsule endoscopy images," *Journal of Software Engineering and Applications*, vol. 7, no. 05, p. 422, 2014.

[11] Y. Yuan, J. Wang, B. Li, and M. Q.-H. Meng, "Saliency based ulcer detection for wireless capsule endoscopy diagnosis," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 2046–2057, 2015.

[12] T. Ghosh, S. A. Fattah, and K. A. Wahid, "Chobs: color histogram of block statistics for automatic bleeding detection in wireless capsule endoscopy video," *IEEE journal of translational engineering in health and medicine*, vol. 6, pp. 1–12, 2018.

[13] A. R. Hassan and M. A. Haque, "Computer-aided gastrointestinal hemorrhage detection in wireless capsule endoscopy videos," *Computer methods and programs in biomedicine*, vol. 122, no. 3, pp. 341–353, 2015.

[14] D.-Y. Liu, T. Gan, N.-N. Rao, Y.-W. Xing, J. Zheng, S. Li, C.-S. Luo, Z.-J. Zhou, and Y.-L. Wan, "Identification of lesion images from gastrointestinal endoscope based on feature extraction of combinational methods with and without learning process," *Medical image analysis*, vol. 32, pp. 281–294, 2016.

[15] R. Nawarathna, J. Oh, J. Muthukudage, W. Tavanapong, J. Wong, P. C. De Groen, and S. J. Tang, "Abnormal image detection in endoscopy videos using a filter bank and local binary patterns," *Neurocomputing*, vol. 144, pp. 70–91, 2014.

[16] M. D. Vasilakakis, D. K. Iakovidis, E. Spyrou, and A. Koulaouzidis, "Dinosarc: Color features based on selective aggregation of chromatic image components for wireless capsule endoscopy," *Computational and mathematical methods in medicine*, vol. 2018, 2018.

[17] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *International journal of computer vision*, vol. 101, no. 1, pp. 184–204, 2013.

[18] T. Ghosh, S. K. Bashar, M. S. Alam, K. Wahid, and S. A. Fattah, "A statistical feature based novel method to detect bleeding in wireless capsule endoscopy images," in *2014 International Conference on Informatics, Electronics & Vision (ICIEV)*, pp. 1–4, IEEE, 2014.

[19] N. E. Koshy and V. P. Gopi, "A new method for ulcer detection in endoscopic images," in *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, pp. 1725–1729, IEEE, 2015.

[20] Y. Fu, M. Mandal, and G. Guo, "Bleeding region detection in wce images based on color features and neural network," in *2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 1–4, IEEE, 2011.

[21] A. K. Kundu, S. A. Fattah, and M. N. Rizve, "An automatic bleeding frame and region detection scheme for wireless capsule endoscopy videos based on interplane intensity variation profile in normalized rgb color space," *Journal of healthcare engineering*, vol. 2018, 2018.

[22] T. Ghosh, S. A. Fattah, C. Shahnaz, and K. A. Wahid, "An automatic bleeding detection scheme in wireless capsule endoscopy based on histogram of an rgb-indexed image," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4683–4686, IEEE, 2014.

[23] T. Ghosh, S. A. Fattah, S. Bashar, C. Shahnaz, K. A. Wahid, W.-P. Zhu, and M. O. Ahmad, "An automatic bleeding detection technique in wireless capsule endoscopy from region of interest," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*, pp. 1293–1297, IEEE, 2015.

[24] G. Pan, G. Yan, X. Qiu, and J. Cui, "Bleeding detection in wireless capsule endoscopy based on probabilistic neural network," *Journal of medical systems*, vol. 35, no. 6, pp. 1477–1484, 2011.

[25] S. Suman, N. Walter, F. A. Hussin, A. S. Malik, S. H. Ho, K. L. Goh, and I. Hilmi, "Optimum colour space selection for ulcerated regions using statistical analysis and classification of ulcerated frames from wce video footage," in *International Conference on Neural Information Processing*, pp. 373–381, Springer, 2015.

[26] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *International journal of computer vision*, vol. 43, no. 1, pp. 29–44, 2001.

[27] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[28] S. Sainju, F. M. Bui, and K. A. Wahid, "Automated bleeding detection in capsule endoscopy videos using statistical features and region growing," *Journal of medical systems*, vol. 38, no. 4, p. 25, 2014.

[29] D.-C. Tseng and C.-H. Chang, "Color segmentation using perceptual attributes," in *11th IAPR International Conference on Pattern Recognition. Vol. III. Conference C: Image, Speech and Signal Analysis,*, vol. 1, pp. 228–231, IEEE Computer Society, 1992.