Thieme

Assessment of bowel cleansing quality in colon capsule endoscopy using machine learning: a pilot study

OPEN ACCESS

Authors

Maria Magdalena Buijs[1,2], Mohammed Hossain Ramezani[3], Jürgen Herp[4], Rasmus Kroijer[1,2], Morten Kobaek-Larsen[2], Gunnar Baatrup[1,2], Esmaeil S. Nadimi[4]

Institutions

1 Department of Surgery, Odense University Hospital, Svendborg, Denmark
2 Institute of Clinical Research, University of Southern Denmark, Odense, Denmark
3 Mads Clausen Institute, University of Southern Denmark, Sønderborg, Denmark
4 Applied Statistical Signal Processing Group, Embodied Systems for Robotics and Learning, Faculty of Engineering, University of Southern Denmark, Denmark

Bibliography

DOI https://doi.org/10.1055/a-0627-7136 |
Endoscopy International Open 2018; 06: E1044–E1050

ISSN 2364-3722

Corresponding author

Maria Magdalena Buijs, Department of Surgery, Odense University Hospital, Baagøes Allé 15, 5700 Svendborg, Denmark
Phone: +4565415190
maria.magdalena.buijs@rsyd.dk

ABSTRACT

Background and study aims The aim of this study was to develop a machine learning-based model to classify bowel cleansing quality and to test this model in comparison to a pixel analysis model and assessments by four colon capsule endoscopy (CCE) readers.

Methods A pixel analysis and a machine learning-based model with four cleanliness classes (unacceptable, poor, fair and good) were developed to classify CCE videos. Cleansing assessments by four CCE readers in 41 videos from a previous study were compared to the results both models yielded in this pilot study.

Results The machine learning-based model classified 47 % of the videos in agreement with the averaged classification by CCE readers, as compared to 32 % by the pixel analysis model. A difference of more than one class was detected in 12 % of the videos by the machine learning-based model and in 32 % by the pixel analysis model, as the latter tended to overestimate cleansing quality. A specific analysis of unacceptable videos found that the pixel analysis model classified almost all of them as fair or good, whereas the machine learning-based model identified five out of 11 videos in agreement with at least one CCE reader as unacceptable.

Conclusions The machine learning-based model was superior to the pixel analysis in classifying bowel cleansing quality, due to a higher sensitivity to unacceptable and poor cleansing quality. The machine learning-based model can be further improved by coming to a consensus on how to classify cleanliness of a complete CCE video, by means of an expert panel.

Introduction

Colon capsule endoscopy (CCE) is a promising technique for evaluation of the colonic and rectal mucosa [1]. However, due to its novelty, the quality and repeatability of CCE investigations still remain largely unknown. A recent study by Buijs et al. (article in preparation) showed a lower interobserver and intraobserver agreement among experts on assessing bowel cleansing quality than on detection of polyps. Detection of polyps seemed to benefit from experience, whereas interobserver and intraobserver agreement on cleansing quality among expert readers was similar to CCE readers with a short formal training.

The rationale for assessing bowel cleansing quality is to determine if the investigation is sufficient to exclude prevalence of polyps larger than 5 mm. This threshold is commonly used in bowel cleansing assessments in bowel cleansing scales in colonoscopy [2]. Because unacceptable bowel cleansing in CCE warrants a diagnostic colonoscopy, reliably assessing cleansing quality is important to reduce the number of unnecessary diagnostic colonoscopies without missing relevant pathology.

The only published scale for bowel cleansing quality in CCE is the Leighton-Rex scale, which classifies five bowel segments in

four different cleansing levels and two bubble effect scales [3]. The two lower scales are considered as inadequate cleansing and the two higher levels as adequate cleansing. There are no guidelines on how to assess bowel quality of the whole video, as this is based on the clinical judgement of the reader. Moreover, the Leighton-Rex scale has not been validated. In colonoscopy there is a large variety of cleansing scales. The best known are the Aronchick Scale, Ottawa Scale, Harfield Cleansing Scale, the Chicago Scale and the Boston Bowel Preparation Scale (BBPS) [4–8]. A recent review evaluated the validity and reliability of all available scales and concluded that the BBPS was the best-validated scale with the strongest correlation to clinical outcomes like polyp detection rate [9]. The problem with all these classifications is that they are subjective and therefore observer-dependent.

Consistent assessment of bowel cleansing quality is important to reliably compare the quality of bowel preparation in both routine clinical use and clinical trials. Only one study described a computer-assisted method of assessing bowel cleansing quality, based on a pixel analysis [10]. In that study the Clean Colon Software Program (CCSP) assessed 50 colonoscopy videos with very good interobserver agreement to cleansing assessments by four experienced endoscopists.

The objective of our pilot study was to assess if a method based on pixel analysis used in the CCSP could also be applied in CCE or if applying an adapted version based on machine learning techniques would result in better agreement with CCE readers' assessments of bowel cleansing quality.

Methods

Study design

In this pilot study, the nonlinear index based on the pixel analysis model from the CCSP [10] and a machine-learning algorithm based on the support vector machines were applied on CCE videos of 41 screening participants. The results of both models were consequently been compared to cleanliness evaluations by four CCE readers.

Data collection

An interobserver study on CCE videos was conducted at our center (article submitted). The 42 videos that were included in this study were selected from a study evaluating diagnostic accuracy of CCE in comparison to colonoscopy [11]. The sample size calculation was based on an interclass correlation coefficient of 0.85 (95 % CI: 0.75–0.95) and warranted 31 videos if assessed by two independent observers. Thirty-two videos with acceptable (good or fair) cleansing and 10 videos with unacceptable (or poor) cleansing as assessed by trained staff (Corporate Health, Hamburg, Germany), were selected from 136 complete videos by an independent research nurse. Unfortunately, data from one of the videos with acceptable cleansing were not retrievable and therefore unavailable for analysis by the algorithms. Bowel cleansing quality in all videos was assessed by two internationally renowned experts (each evaluated over 1500 videos) and two medical doctors who had only short formal training. Cleanliness was assessed in three sections (right, transverse and left colon) and overall, in four classes: unacceptable, poor, fair and good.

The models analyzed the CCE videos from the first cecal image to the last rectal image, as selected by the CCE readers.

Developing the models

Bowel cleansing quality in the CCE videos was analyzed in three steps. First, individual pixels were qualified as either clean or dirty. Subsequently, cleanliness of each image was determined based on the number of clean and dirty pixels it contained. All frames were scaled in four cleansing levels: unacceptable, poor, fair or good. The bowel cleansing quality of the whole video was determined by the median cleansing level of all frames.

Nonlinear index

In the nonlinear index the pixel identification (J) is performed based on the red, green and blue (RGB) elements of the pixel, as in the CCSP.

$$J = \frac{R-G}{G-B} \quad (1)$$

Distribution of different colors in the pixel determines if the pixel is clean or dirty as compared to threshold T_0 in the following equation.

$$\begin{cases} \text{clean} & \text{if} \quad J \le T_0 \\ \text{dirty} & \text{if} \quad J > T_0 \end{cases} \quad (2)$$

A threshold of 0.7 was chosen in this study because it agreed best with the pixel classification.

Support vector machine

Support vector machines (SVM) are based on machine-learning concepts, in which an "expert's input" is used to train the model to classify a variable. We used SVM to determine the cleanliness of a pixel. A medical doctor (RK) classified pixels to be either clean or dirty in a random selection of CCE frames.

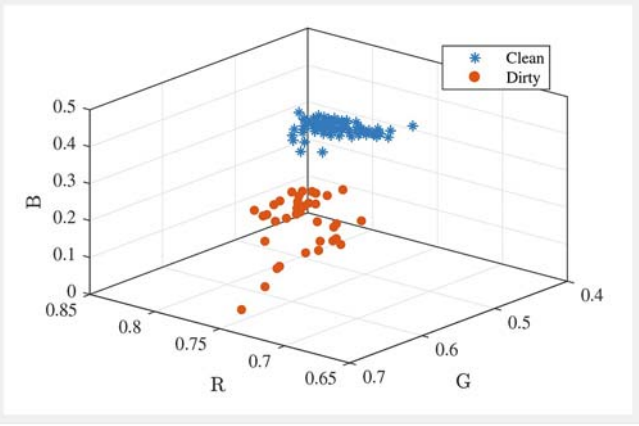
The model based on this data made a clear distinction between the dirty and clean pixels, as is visualized in ▶ **Fig. 1**.

The next step was to determine the cleanliness of a video frame, based on the number of clean and dirty pixels. Pixels that were overexposed or underexposed in the frame were excluded from this analysis. Cleanliness was subsequently assessed by the following equation, in which N_A is the number of classified pixels per frame and $f(d_i)$ represents the cleanliness of one pixel using SVM algorithms described in the **Appendix**.

$$I_k = \frac{1}{N_A} \sum_{i=1}^{N_A} f(d_i) \quad (3)$$

The value of I_k was used to determine the level of cleanliness of the k^{th} frame based on different thresholds as unacceptable, poor, fair or good. Thresholds in this study were predicted and corrected using learning techniques embedded in the SVM algorithm, based on assessments of images by a medical doctor (MMB). Examples of the cleansing classes are presented in ▶ **Fig. 2**.

The complete video was thereafter assessed by determining median cleanliness of all the separate frames and weighted based on the number of classified pixels (N_A) in the frames. A



▶ **Fig. 1** SVM classifier for clean and dirty pixels. SVM classifier for clean and dirty pixels in the RGB space based on extensive expert valuations of colon capsule endoscopy frames.

more detailed description of the SVM model can be found in the appendix of this article.

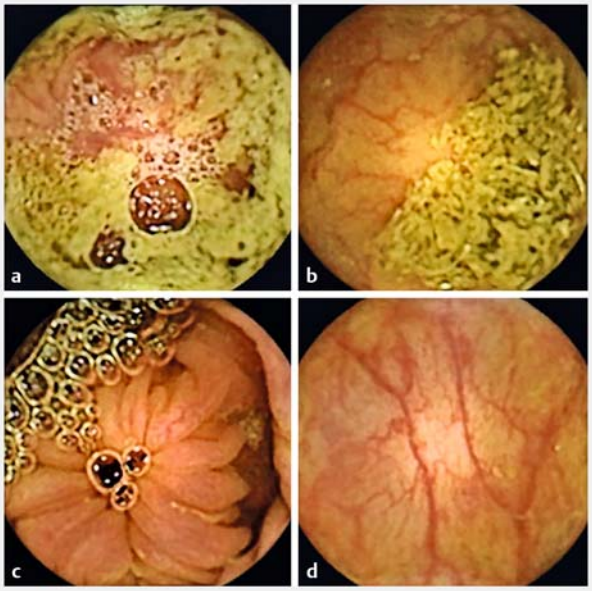
Statistical analysis

Descriptive statistics were obtained from the nonlinear index and SVM model evaluating the underlying empirical data. Inter-observer variability between different CCE readers was determined by calculating means and standard deviations of the cleansing quality classifications. Validity of both models was evaluated by performing a sensitivity analysis to determine the confidence interval of cleansing classifications within one video based on data from both video heads. We chose to perform a separate analysis for the videos with unacceptable cleansing quality according to at least one observer to visualize agreement between readers and algorithms on unacceptable cleansing quality. All calculations were performed in MATLAB® R2017a.

Results

Bowel cleansing quality in all videos was analyzed with the nonlinear index and SVM model and compared to cleansing assessments by four CCE readers (▶ **Fig. 3**). Mean assessment of cleansing quality by the CCE readers is shown with red lines in ▶ **Fig. 3** and demonstrates a discrepancy between different observers. One of the CCE observers, expert 2, qualified none of the videos as unacceptable. However, the graph clearly shows higher agreement of the SVM model with different observers, as compared to the nonlinear index.

To facilitate comparison of the models to all CCE readers, their averaged cleansing classification of the videos was used in the comparisons below. The nonlinear index classified 32 % of the videos in agreement with the CCE readers, as compared to 47 % in the SVM model. The nonlinear index classified the cleansing quality in none of the videos as unacceptable and only in a few videos as poor. The relative error shows that the nonlinear index had a tendency to overestimate the bowel cleansing quality as compared to CCE readers.



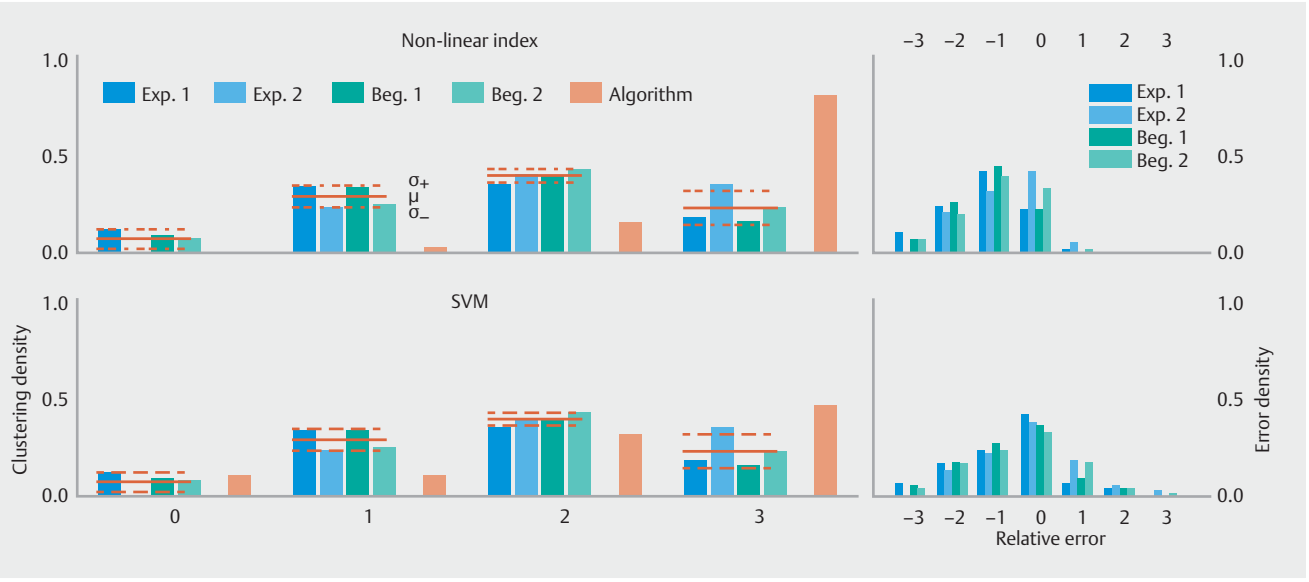
▶ **Fig. 2** Examples of SVM class labels in individual frames. From left to right: Unacceptable (0), Poor (1), Fair (2) and Good (3).

The nonlinear index model overestimated 43 % of the videos by one class if compared to the CCE readers and 32 % of the videos differed more than one class from the CCE readers' assessments.

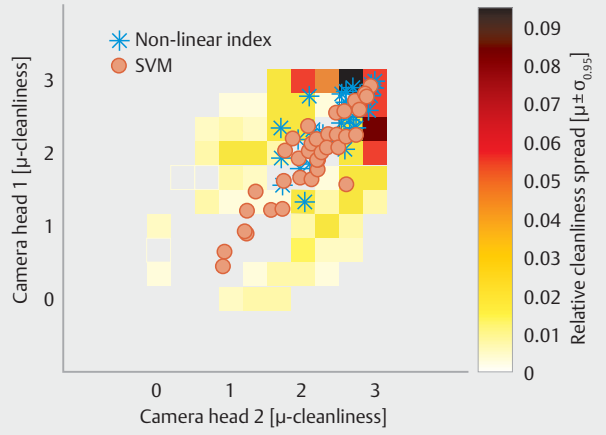
In the SVM model, however, 41 % of the videos were classified either one class above or below the CCE readers' assessment of cleansing quality and only 12 % of the assessments varied more than one class from the classification by the CCE readers.

Sensitivity analysis of the models showed agreement between the videos from both video heads, as well as how the classification of each frame within a video compares to the final classification of that video (▶ **Fig. 4**). The small confidence interval in the fair and good classes indicates that 95 % of the frames in those classes have a similar classification, whereas variability in the lower cleansing classes was larger. Variability between both camera heads was larger in the nonlinear index than in the SVM model.

A separate analysis of agreement on unacceptable classifications between both algorithms and CCE readers was performed by selecting all 11 videos that were classified as unacceptable by at least one observer (▶ **Fig. 5**). The nonlinear index classified 10 of the "unacceptable" videos as either fair or good and one as poor, whereas the SVM classified seven videos as unacceptable of which five were in agreement with at least 1 CCE reader. In the two other videos, all CCE readers agreed on a fair classification in one and disagreed on the classification in the other. In four videos, SVM did not detect unacceptable cleansing, where 1 to 3 out of 4 CCE readers classified the video as unacceptable.



▶ **Fig. 3** Bowel cleanliness classification by Non-linear Index and SVM models compared to CCE readers. The graphs on the left side show classification of the videos by the four CCE readers and respectively non-linear index and the SVM model. All videos are classified as unacceptable (0), poor (1), fair (2) and good (3) by all observers. The algorithm bar represents classification by the different models. Mean assessment of the CCE readers is visualized with horizontal lines and the standard deviations with dotted lines. The graphs on the right side show the relative error of the models compared to the individual CCE readers. Exp. 1: expert 1; Exp. 2: expert 2; Beg. 1: beginner 1; Beg. 2: beginner 2.



▶ **Fig. 4** Sensitivity of the Non-linear Index and SVM models. Sensitivity of the models is weighted with respect to the number of

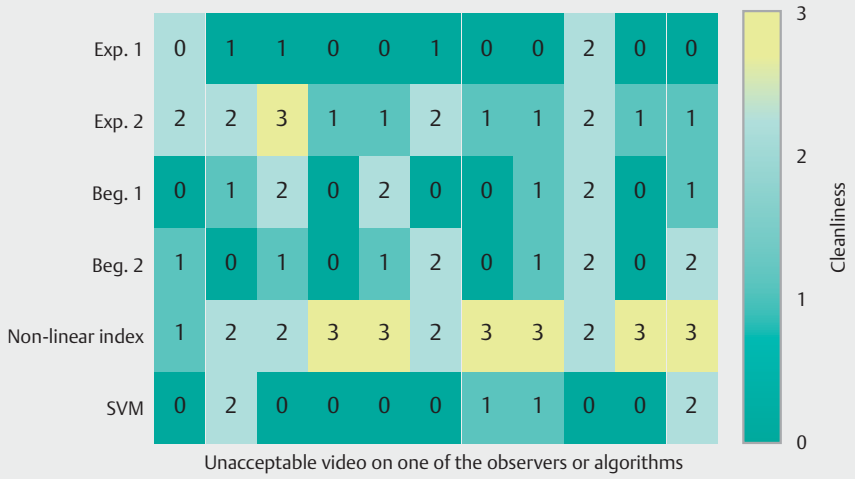
Discussion

This is the first study describing objective evaluation of bowel cleansing quality in CCE, which shows that machine learning-based models are able to distinguish between acceptable and unacceptable cleansing. Objective evaluation of bowel cleansing quality is more consistent than subjective assessment in determining which patients would require a diagnostic colonoscopy in routine clinical use as well as comparing bowel cleansing in clinical studies. The current evaluation is subjective as is clearly demonstrated by the differences in assessment by the CCE readers. This also shows the relevance of an objective measure of cleanliness. A previous study using colonoscopy images showed high variability in assessment of cleanliness among clinicians [12]. Online training for the BBPS improved consistency of assessments and led to high interobserver agreement [4].

The SVM model in this pilot study is in agreement with the averaged classification by CCE readers, with a disagreement of more than one class in only 12 % of the videos. In comparison to the nonlinear index, the SVM model is more sensitive to detection of unacceptable and poor cleansing and less prone to overestimation of cleansing level. The validity of the models as portrayed by the sensitivity analysis shows very consistent evaluation of bowel cleansing quality in the fair and good videos and lower consistency in poor and unacceptable videos, which is likely due to the fact that poor and unacceptable videos usually also have some “clean” sections.

Rosa-Rizotto et al. described the interclass correlation coefficient (ICC) between the observers and their model. We did not calculate the ICC because of the uneven distribution of observers, namely four CCE readers and one SVM model. In our opinion, the current presentation with the relative errors and separate assessment of “unacceptable” videos contributes to a greater understanding of the sensitivity and quality of the SVM model than the ICC would.

An important limitation in our study is the absence of a robust gold standard to assess cleanliness in CCE videos. The Leighton-Rex scale leaves this assessment up to the reader, even though it is the most important assessment of cleansing quality of CCE for both routine clinical and research purposes. Moreover, detection of landmarks within the colon is difficult due to the capsule moving back and forth within the colon and individual differences in anatomy, therefore, selecting five dif-



▶ **Fig. 5** Inter-observer comparison for all "unacceptable" videos. For each observer and both models, classification of videos that were classified as "unacceptable" by at least one observer are displayed. Classification: unacceptable (0), poor (1), fair (2) and good (3).

ferent segments might impair use of the Leighton-Rex scale. Use of three instead of five segments might improve usability of the Leighton-Rex scale. Another issue is that the Leighton-Rex scale has not been studied in comparison to detected polyps. Even though fecal matter might obscure the mucosa, it is possible that polyps are still detected, due to the 344° view and back and forth movement of the capsule. The primary reason to assess cleanliness is to evaluate risk of missed polyps, therefore, a cleansing scale should be studied in comparison to polyp detection rate as with the BBPS. To develop a reliable scale in CCE, assessments of cleansing should be evaluated by an expert panel to create consensus on bowel cleansing quality.

A limitation in the methods of this study is that we only assessed bowel cleansing quality of the whole video and did not provide a separate analysis for different segments as proposed by the Leighton-Rex scale. The purpose of this study was limited to investigating the possibilities for machine learning in cleanliness assessments, therefore, assessments of the complete video sufficed, but for future use in routine clinical settings, the model will need to be adjusted to assess cleansing per segment.

The SVM model was trained by two CCE readers with limited experience, which might have attributed to a bias in the learning process of the model, however, classifications of the model were not congruent with the two beginners. Further training of the SVM model by adding images and videos that have been assessed by other and more experienced CCE readers, ideally by an expert panel, will improve the quality of the model.

The SVM model has not been tested in cases with melanosis coli, therefore it is unknown if the model can accurately assess cleansing in those patients.

Interestingly enough the results from the sensitivity analysis in ▶ **Fig. 4** show that videos with fair or good cleansing quality according to the SVM model generally have mostly video frames with a similar classification. The variation is larger in poor and unacceptable videos, which might be solved by assessing the videos per segment instead of as a complete video. The SVM model is still in development, exposure to more images and videos with poor or unacceptable cleanliness will make it better in identifying "poor" and "unacceptable" videos. We are planning to organize the aforementioned expert panel to assess short videos with different levels of cleanliness to determine a standard for CCE that can also be used to train the SVM model.

A newer version of the SVM model can be used as a standard for bowel cleansing quality and therefore be used in future studies to enable comparisons of bowel cleansing classification between different clinicians, centers and countries. Another possibility would be to compare the bowel cleansing quality of CCE and colonoscopy, to investigate if medical doctors assess those investigations similarly.

Conclusion

In this preliminary study, the learning-based SVM model was superior to the pixel analysis model in classifying bowel cleansing quality. The SVM model is capable of distinguishing between different cleanliness classes in CCE, however, specificity for detecting unacceptable cleansing is not in perfect agreement with CCE readers. This is partly due to inconsistency in assessment of cleanliness among the readers. A more consistent evaluation of cleanliness, as determined by an expert panel, could be used to improve the SVM model.

Acknowledgements

Funding for this study was provided through research funds from Odense University Hospital and the Region of Southern Denmark, as well as the Danish Cancer Society.

Competing interests

None

References

[1] Spada C, Pasha SF, Gross SA et al. Accuracy of first- and second-generation colon capsules in endoscopic detection of colorectal polyps: a systematic review and meta-analysis. Clin Gastroenterol Hepatol 2016; 14: 1533 – 1543.e8

[2] Johnson DA, Barkun AN, Cohen LB et al. Optimizing adequacy of bowel cleansing for colonoscopy: recommendations from the US Multi-Society Task Force on Colorectal Cancer. Am J Gastroenterol 2014; 109: 1528 – 1545

[3] Leighton JA, Rex DK. A grading scale to evaluate colon cleansing for the PillCam COLON capsule: a reliability study. Endoscopy 2011; 43: 123 – 127

[4] Calderwood AH, Jacobson BC. Comprehensive validation of the Boston Bowel Preparation Scale. Gastrointest Endosc 2010; 72: 686 – 692

[5] Aronchick CA LW, Wright SH, DuFrayne F et al. Validation of an instrument to assess colon cleansing [abstract]. Am J Gastroenterol 1999; 94: 2667

[6] Rostom A, Jolicoeur E. Validation of a new scale for the assessment of bowel preparation quality. Gastrointest Endosc 2004; 59: 482 – 486

[7] Halphen M, Heresbach D, Gruss HJ et al. Validation of the Harefield Cleansing Scale: a tool for the evaluation of bowel cleansing quality in both research and clinical practice. Gastrointest Endosc 2013; 78: 121 – 131

[8] Gerard DP, Foster DB, Raiser MW et al. Validation of a new bowel preparation scale for measuring colon cleansing for colonoscopy: the chicago bowel preparation scale. Clin Transl Gastroenterol 2013; 4: e43

[9] Parmar R, Martel M, Rostom A et al. Validated Scales for colon cleansing: a systematic review. Am J Gastroenterol 2016; 111: 197 – 204

[10] Rosa-Rizzotto E, Dupuis A, Guido E et al. Clean colon software program (CCSP), proposal of a standardized method to quantify colon cleansing during colonoscopy: preliminary results. Endosc Int Open 2015; 3: E501 – E507

[11] Kobaek-Larsen M, Kroijer R, Dyrvig AK et al. Back-to-back colon capsule endoscopy and optical colonoscopy in colorectal cancer screening individuals. Colorectal Dis 2018; 20: 479 – 485

[12] Ben-Horin S, Bar-Meir S, Avidan B. The impact of colon cleanliness assessment on endoscopists' recommendations for follow-up colonoscopy. Am J Gastroenterol 2007; 102: 2680 – 2685

Appendix

Soft switch

In this study, a nonlinear classification is applied with a Gaussian radial basis function as kernel. Further, the pixel classification function $f(d_i)$ in Eq. (3) is modified to resemble a soft transition between dirty and clean pixels. By defining d_i as the distance between the i^{th} pixel to the hyperplane in the RGB feature space, where the sign represents whether the pixel is clean or dirty, the function $f(d_i)$ determines the cleanliness of the pixel as follows:

$$f(d_i) = \begin{cases} -1 & \text{if} \quad d_i \le -\alpha \\ \frac{1}{\alpha} d_i & \text{if} \quad -\alpha < d_i < \alpha. \\ 1 & \text{if} \quad \alpha \le d_i \end{cases} \tag{4}$$

Applying Eq. (4) to the pixels in the vicinity of the hyperplane with radius α, the cleanliness is determined based on the distance of the point to the hypersphere. The points outside α have binary cleanliness.

Class variables

As Eq. (3) returns a continuous variable, the cleanliness of a frame is evaluated by comparing I_k to a series of thresholds given by:

$$L^j = \begin{cases} 0\ (\text{Unacceptable}) & \text{if} \quad T_1 < I_k \le 1 \\ 1\ (\text{Poor}) & \text{if} \quad T_2 < I_k \le T_1. \\ 2\ (\text{Fair}) & \text{if} \quad T_3 < I_k \le T_2 \\ 3\ (\text{Good}) & \text{if} \quad -1 \le I_k \le T_3 \end{cases} \tag{5}$$

Parameter estimation

As the SVM classifier is trained on domain knowledge, there are no free parameters in the cleanliness assessment. However, the parameters in Eqs. (4) and (5), namely α, and T_k , $k = 1,2,3$, are estimated through an optimization process.

In order to define and solve the optimization problem at hand, a gold standard is necessary. For this purpose, a variety of colon frames from different patients and different cleanliness levels are selected. The cleanliness of frames is assessed by a medical doctor. Finally, a set of N_g assessed frames is selected:

$$L_g^j; j = 1, 2, \ldots, N_g. \tag{6}$$

The parameters can be estimated by solving

$$\min \sum_{j=1}^{N_g} |E^j| = \min \sum_{j=1}^{N_g} |L_g^j - L^j| \tag{7}$$

where E^j is the estimated error of the frame's cleanliness.