Towards Automatic Direct Observation of Procedure and Skill (DOPS) in Colonoscopy

Mirko Arnold[1], Anarta Ghosh[1], Glen Doherty[2], Hugh Mulcahy[2], Christopher Steele[3], Stephen Patchett[4] and Gerard Lacey[1]

[1]*School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland*

[2]*Centre for Colorectal Disease, St. Vincent's University Hospital, University College Dublin, Dublin, Ireland*

[3]*FRCP, Letterkenny General Hospital Donegal, Letterkenny, Ireland*

[4]*Beaumont Hospital, Royal College of Surgeons in Ireland, Dublin, Ireland*

Keywords: Medical Computer Vision, Medical Image Applications, Endoscopic Imaging, Vision-based Quality Assessment, Colonoscopy, Machine Learning.

Abstract: The quality of individual colonoscopy procedures is currently assessed by the performing endoscopist. In light of the recently reported quality issues in colonoscopy screening, there may be significant benefits in augmenting this form of self-assessment by automatic assistance systems. In this paper, we propose a system for the assessment of individual colonoscopy procedures, based on image analysis and machine learning. The system rates the procedures according to criteria of the validated Direct Observation of Procedure and Skill (DOPS) assessment, developed by the Joint Advisory Group on GI Endoscopy (JAG) in the UK, a system involving expert assessment of procedures based on an assessment form.

1 INTRODUCTION

Colonoscopy is considered the gold standard for colorectal cancer screening. In addition to a thorough visualisation of the large intestine, it is possible to directly take tissue samples or remove polyps. Detection and removal of such polyps can prevent them from developing into cancer, which is why many organisations recommend regular colorectal screening for people over a certain age (U.S. Preventive Services Task Force, 2008; WGO, 2007). With the commencement of screening programs for asymptomatic patients in more and more countries (Benson et al., 2008), the number of colonoscopy procedures is constantly growing.

While colonoscopy can reduce the incidence of colorectal cancer, it does not eliminate the risk completely. In fact, studies have shown that in practice, the percentage of patients developing colorectal cancer shortly after having undergone colonoscopy screening is between 2 % and 6 % (Bressler et al., 2007). Since the development of small polyps into cancer is known to be a gradual process that evolves slowly over years, this means that a significant number of polyps remain undetected despite colonoscopy screening.

The cause for these miss rates is not yet well understood. It is likely to be a combination of a number of factors. Screening technique and bowel preparation play an important role, as they determine the amount of colonic mucosa that is visualised. Perceptual and cognitive aspects have to be taken into account where visualised polyps are not correctly identified by the endoscopist.

In practice, the quality of colonoscopy procedures is self-assessed by the performing endoscopist. Additionally, a number of measures are recorded for statistical evaluation. Examples of such a measures are the average withdrawal time or the adenoma detection rate. These measures, however, can only assess the average performance of the endoscopist.

For *individual* procedures there exist subjective quality measures such as the direct observation and assessment of the procedures by one or more experts. Due to the cost of trained experts it is impracticable to use this form of quality assessment more than for occasional audits or as part of the examination of trainees. The routine assessment of individual procedures is the task of the performing endoscopist alone.

In this paper, we propose an image analysis and machine learning based system for automatic assessment of individual procedures according to criteria

Arnold M., Ghosh A., Doherty G., Mulcahy H., Steele C., Patchett S. and Lacey G..
Towards Automatic Direct Observation of Procedure and Skill (DOPS) in Colonoscopy.
DOI: 10.5220/0004300800480053
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 48-53
ISBN: 978-989-8565-48-8

of the *Direct Observation of Procedure and Skill* (DOPS) assessment method developed by the Joint Advisory Group on GI Endoscopy (JAG) in the UK. We consider JAG DOPS to be particularly relevant, as it is the most mature among the available assessment systems and in active use within the NHS Bowel Cancer Screening Programme in the UK, while others have yet to be implemented. A standard JAG DOPS assessment involves the observation of colonoscopy procedures by two trained experts, who rate the procedures independent of each other by filling out a predefined assessment form.

This assessment form is divided into four groups of criteria: 1) Assessment, Consent, Communication, 2) Safety and Sedation, 3) Endoscopic Skills During Insertion and Procedure and 4) Diagnostic and Therapeutic Ability. We concentrate on the criteria in the third group with the objective to measure automatically, to what degree the technical prerequisites are in place for a high quality colonoscopy screening.

2 BACKGROUND

Automatic computation of quality indicators from colonoscopy video data has only been marginally addressed in the literature, despite the interest of the medical research community.

Hwang, et. al (Hwang et al., 2005), combined indistinct frame detection, camera motion estimation and lumen recognition to obtain a number of quality measures that are mostly related to durations of semantic segments of the video. Examples are the duration of the insertion phase and the withdrawal phase, or the duration of the withdrawal phase disregarding all indistinct frames (the *clear withdrawal time*). An interesting measure is also the ratio of *wall views* and *lumen views*, which should be properly balanced, according to the authors. In a more recent article (Oh et al., 2009), the same authors included also their intervention detection method to determine the clear and intervention-free withdrawal time.

Liu, et al. (Liu et al., 2007), addressed a different aspect of procedure quality by evaluating, whether the camera was pointed at all sides of a colon segment. They defined the location of the lumen as the centre, subdivided the view deviations from the lumen direction into four quadrants and computed a histogram of the number of images, in which the camera was pointed in the direction of the different quadrants. The authors argued that examination of all four quadrants is desirable. This can be seen as a first attempt to measure the amount of mucosal surface that was visualised during a procedure. Liu, et al., recently proposed an amended version of their approach (Liu et al., 2010), mainly with an improved method for detecting the lumen position. Apart from this, the histogram measure was replaced by counting the number of spirals (the coverage of all 4 quadrants) in a procedure. Both approaches do not take into account the forward and backward motion of the camera. A fast movement through a 20 cm long segment of the colon would get a similar score as a careful slow inspection of a very short segment.

In summary, the literature offers only few suggestions for quality measures, that can be determined automatically. Measuring the insertion and withdrawal time is only valid when looking at an average over many procedures. Measures for individual cases, such as the wall-view/lumen-view ratio or the quadrant coverage histogram, have yet to be evaluated for validity. We consider it beneficial to look into measures that represent generally accepted insertion and examination techniques and best practices, which we do in this paper.

3 APPROACH

There are two different types of data we consider in our approach. One is video data from the endoscopic camera. The other are measurements of the longitudinal and circular motion of the shaft of the endoscope outside the anus using a sensor device. We use a number of algorithms to measure patterns of image features and endoscope motion for modelling the underlying characteristics of the JAG DOPS assessment criteria.

The individual measures are organised into two levels. The first contains all measures characterising single images, while the second level of measures describes characteristics of the complete procedures. The measures of single images are included in the second level by summarising their behaviour over the course of the procedure.

For the procedure measures and their mapping to DOPS criteria we use data obtained from an experiment we conducted, in which endoscopists performed screening procedures on a colonoscopy training model. The data comprises videos from the endoscope camera together with motion sensor readings, profile data of the endoscopists and ratings of the procedures by two trained experts according to JAG DOPS criteria. The motion sensor readings are combined with the image based characteristics to measure a number of endoscope handling patterns. Furthermore, we use the recorded longitudinal motion of the shaft of the endoscope to estimate the depth of inser-

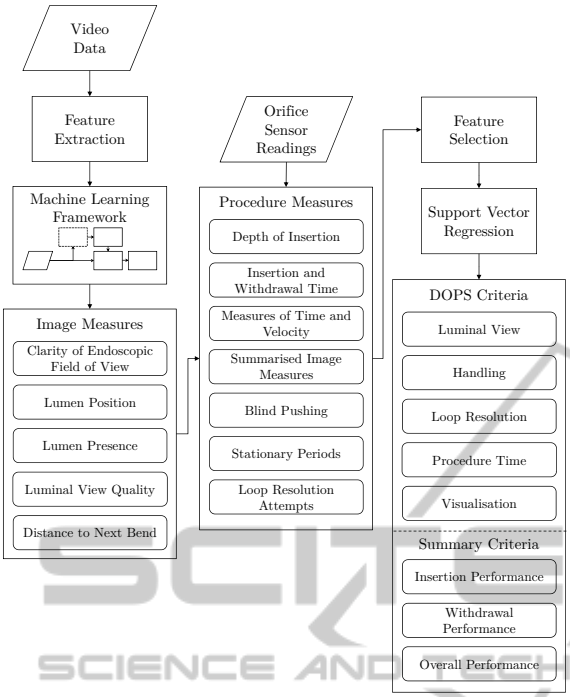

Figure 1: Layout of the complete quality assessment system.

tion of the endoscope. All image based characteristics can therefore be analysed for their behaviour over time and depth of insertion.

This combination of image and endoscope motion characteristics results in a large set of measures, describing colonoscopy procedures in great detail. We use subsets of these measures as features for the training of regression models for each of the chosen JAG DOPS criteria. We evaluate the proposed method by comparing the model predictions to the ratings of the trained experts. The complete system is shown in Figure 1.

For brevity, we keep the description of the individual building blocks to a minimum, concentrating on the system as a whole. Future publications will contain details on each of the novel algorithms involved in the system.

3.1 Image Measures and Machine Learning Framework

All the involved measures of characteristics of single endoscopic images are based on image features, which are used to train predictive models in a universal machine learning framework. This framework involves automatic forward feature selection, parameter optimisation (grid search with iterative refinement) and training of support vector machines. Depending on the properties of the particular measure, we use different kinds of error measures for optimisation and different types of support vector machines.

Among the image measures, we use a novel measure for the clarity of the field of view, extending the current state of the art by introducing multiple grades of clarity as opposed to the previously proposed binary classification into informative and indistinct images (see, e.g., (Arnold et al., 2009; Oh et al., 2007)). The clarity measure is based on different representations of the amount of structure in the image. These structure features are obtained from a wavelet decomposition of the image and differences between intensity histograms of horizontal, vertical and diagonal lines in the image. For regression, we use a ν support vector regression model.

By introducing measures for different characteristics of luminal views in single images, i.e., luminal view quality, lumen presence, position of the lumen and distance to the closest bend, we achieve a detailed description of colonoscopic images with direct implications to visualisation quality and endoscope handling skills. The lumen characteristics are inferred from intensity, shape and colour features obtained from maximally stable extremal regions (MSER, (Matas et al., 2004)) in the image, together with intensity and colour features of the whole image. In a first step, one of the MSER is chosen as the most likely representation of the lumen region. Given this region and the associated features, we use a C support vector machine for classifying the lumen as either *present* or *absent*. If the lumen is present, its position is computed as the centroid of the lumen region. ν support vector regression models are used for measuring luminal view quality and distance to the closest bend.

All image based measures benefit from the methods for detecting and inpainting specular highlights in endoscopic images that we have proposed earlier in (Arnold et al., 2010). By either omitting or inpainting specular pixels in the image we reduce the influence of the strong gradients and intensity saturation in these areas.

3.2 Endoscope Motion and Measures of Procedure Characteristics

For the characterisation of complete procedures we incorporate measurements of a motion sensor, located outside the anus, which measures longitudinal and circular displacement of the endoscope. Due to the motion sensor being optimised for a specific colonoscopy training model, it was necessary to design an experiment for data collection, in which video and

sensor data could be recorded simultaneously. The obtained data was complemented by information on the experience of the participating endoscopists and assessments of the procedures by two trained experts according to JAG DOPS criteria. This way we were able to collect video and motion sensor data, for the development of measures for procedure characteristics, and associated DOPS ratings to train and evaluate models for automatic DOPS assessment.

Given the obtained motion sensor data, we use a moving average filter to estimate the depth of insertion of the endoscope. The speed of the endoscope can be directly obtained. We use these measurements in methods to automatically infer the insertion and withdrawal times of procedures and to detect stationary periods during insertion, attempts of loop resolution and occurrences of pushing without a clear view. All these measures are based on handling patterns detected in the motion sensor data and, in the case of pushing without clear view, combining these with the measure for the clarity of the field of view in images.

The estimated depth of insertion also allows us to divide the colon into a number of segments, opening up new possibilities for summarising the behaviour of certain characteristics over the course of a procedure. The proposed image measures can therefore be mapped to meaningful procedure characteristics by applying various statistics for their summarisation. Previously it was only possible to analyse the behaviour over time, lacking any form of spatial information. Combining the spatial segmentation with the time scale, handling patterns and the set of proposed image measures, we obtain a set of 92 procedure measures.

3.3 Mapping Procedure Measures to DOPS Criteria

We consider the following criteria from the JAG DOPS assessment (we use the short terms in parentheses in the following):

- Maintains luminal view / inserts in luminal direction. (*Lumen*)
- Uses torque steering and control knobs appropriately. (*Handling*)
- Recognises and logically resolves loop formation. (*Looping*)
- Completes procedure in reasonable time. (*Time*)
- Adequate mucosal visualisation. (*Visualisation*)

In addition, we included criteria summarising the performance during the insertion phase (*Insertion*), the withdrawal phase (*Withdrawal*) and the whole procedure (*Overall*).

Table 1: Groups of features and their relevance for the different DOPS criteria.

	Lumen	Handling	Looping	Time	Visualisation	Insertion	Withdrawal	Overall
INSERTION								
Backward Speed	-	×	-	×	-	×	-	×
Forward speed	-	×	-	×	-	×	-	×
Circular Speed	-	×	-	-	-	×	-	×
Blind Pushing	-	×	×	-	-	×	-	×
Clarity	×	×	-	-	-	×	-	×
Loop Resolution	-	×	×	-	-	×	-	×
Lumen Dist. To Centre	×	×	-	-	-	×	-	×
Dist. To Bend	×	-	-	-	-	×	-	×
Lumen Pos. X	×	×	-	-	-	×	-	×
Lumen Pos. X Stdev	×	×	-	-	-	×	-	×
Lumen Pos. Y	×	×	-	-	-	×	-	×
Lumen Pos. Y Stdev	×	×	-	-	-	×	-	×
Lumen Presence	×	×	-	-	-	×	-	×
Lumen View Quality	×	×	-	-	-	×	-	×
Stationary Time	-	×	×	-	-	×	-	×
Time	-	×	×	×	-	×	-	×
WITHDRAWAL								
Backward Speed	-	-	-	×	×	-	×	×
Forward speed	-	-	-	×	×	-	×	×
Circular Speed	-	×	-	-	×	-	×	×
Blind Pushing	-	×	×	-	-	-	×	×
Clarity	×	×	-	-	×	-	×	×
Lumen Dist. To Centre	×	×	-	-	×	-	×	×
Dist. To Bend	×	-	-	-	×	-	×	×
Lumen Pos. X	×	×	-	-	×	-	×	×
Lumen Pos. X Stdev	×	×	-	-	×	-	×	×
Lumen Pos. Y	×	×	-	-	×	-	×	×
Lumen Pos. Y Stdev	×	×	-	-	×	-	×	×
Lumen Presence	×	×	-	-	×	-	×	×
Lumen View Quality	×	×	-	-	×	-	×	×
Time	-	-	-	×	×	-	×	×

The major challenge in training models for mapping procedure measures to DOPS criteria is the size of the data set. It consists of 25 procedure videos with associated motion sensor measurements and DOPS ratings produced by two trained experts. The JAG DOPS assessment system is a subjective method of assessment, which means that full agreement is unlikely, even between trained assessors. In such a small data set, these deviations can have a strong impact on measures of agreement between assessors. We therefore report results separately, taking each assessor on its own as a reference for our method. 25 examples are also insufficient for performing automatic feature selection, as we did earlier within the machine learning framework to measure image characteristics. Having 92 features available this method would in-

Table 2: Agreement between predictions of the automatic system and the ratings of assessor 1, measured using Kendall's τ, Pearson's r and Krippendorff's α. For τ and r the values in parentheses are the corresponding p-values. For α the table shows the lower and upper limits of the 95% confidence intervals in parentheses.

	τ (p-value)	α (95% CI lower lim., upper lim.)	r (p-value)
Lumen	0.442 (0.006)	0.169 (-0.111, 0.426)	0.648 (<0.001)
Handling	0.516 (0.001)	0.642 (0.412, 0.815)	0.705 (<0.001)
Looping	0.294 (0.077)	0.276 (-0.056, 0.564)	0.465 (0.019)
Time	0.290 (0.088)	0.229 (-0.154, 0.538)	0.344 (0.092)
Visualisation	0.460 (0.005)	0.607 (0.499, 0.711)	0.541 (0.005)
Insertion	0.404 (0.011)	0.434 (0.239, 0.618)	0.525 (0.007)
Withdrawal	0.488 (0.003)	0.566 (0.325, 0.756)	0.490 (0.013)
Overall	0.586 (<0.001)	0.400 (0.208, 0.575)	0.783 (<0.001)

Table 3: Agreement between predictions of the automatic system and the ratings of assessor 2, measured using Kendall's τ, Pearson's r and Krippendorff's α.

	τ (p-value)	α (95% CI lower lim., upper lim.)	r (p-value)
Lumen	-0.459 (0.007)	-0.530 (-1.000, -0.052)	-0.573 (0.003)
Handling	0.241 (0.138)	0.213 (-0.054, 0.445)	0.335 (0.102)
Looping	-0.016 (0.940)	0.024 (-0.316, 0.328)	0.093 (0.660)
Time	-0.005 (1.000)	-0.202 (-0.660, 0.209)	-0.491 (0.013)
Visualisation	0.081 (0.639)	0.131 (-0.333, 0.525)	0.120 (0.567)
Insertion	-0.263 (0.109)	-0.216 (-0.664, 0.137)	-0.431 (0.031)
Withdrawal	-0.041 (0.818)	0.026 (-0.309, 0.322)	-0.171 (0.414)
Overall	0.081 (0.629)	0.095 (-0.239, 0.377)	-0.105 (0.618)

evitably lead to overfitting and, therefore, poor performance on unseen data. Choosing the relevant features manually based on domain knowledge is not trivial, since there is seldom an intuitive advantage of any summarisation operation over the others for the different features.

We therefore take a hybrid approach to feature selection. We organise the features into groups, each of which is made up of a set of features representing the same underlying procedure characteristic. We then use our domain knowledge to select the relevant groups (characteristics) for each of the target DOPS measures. Within each group we perform correlation analysis to chose the most relevant feature of the group. Only the feature with the highest correlation with the measure in the training examples is retained in each group. This approach allows us to reduce the dimensionality of the feature space significantly. This, in turn, reduces the tendency of overfitting. Table 1 lists the target measures and feature groups we identified and shows which group we consider relevant for each of the target measures. By using this method, depending on the DOPS criterion in question, the number of features is reduced from 92 to between 5 and 30.

We use ν support vector regression models for prediction of the DOPS criteria. To make best use of the small data set, the SVMs are trained in a nested cross-validation scheme. In the outer loop, in each cross-validation fold, we leave out a single example for testing and hand the rest to the inner loop. In the inner loop, parameters are optimised with a grid search approach, again using leave-one-out cross-validation.

4 EVALUATION

For the performance evaluation of the proposed system, we compute the strength of association between the trained SVMs and each of the two assessors. We are using Kendall's τ coefficient for this analysis, providing Pearson correlation r for comparison. The results are shown in Tables 2 and 3. For comparison, Table 4 shows the agreement between the two assessors.

We can see a moderate association for most of the measures when comparing to assessor 1. Interestingly, the method fails to achieve any significant agreement with assessor 2. This may be due to rating outliers, which can have a strong effect given such a small sample size. There appears to be a stronger association between our method and assessor 1 than between the ratings of the two assessors in all except the *looping* criterion. Association between the predictions of our method and the ratings of assessor 1 is always statistically significant at the 0.1 significance

Table 4: Agreement between the ratings of assessor 1 and the ratings of assessor 2, measured using Kendall's τ, Pearson's r and Krippendorff's α.

	τ (p-value)	α (95% CI lower lim., upper lim.)	r (p-value)
Lumen	0.217 (0.262)	0.230 (-0.360, 0.680)	0.314 (0.126)
Handling	0.311 (0.084)	0.310 (-0.070, 0.640)	0.366 (0.072)
Looping	0.494 (0.008)	0.470 (0.180, 0.720)	0.571 (0.003)
Time	0.041 (0.864)	0.050 (-0.390, 0.450)	0.051 (0.810)
Visualisation	0.185 (0.342)	0.210 (-0.360, 0.640)	0.181 (0.387)
Insertion	0.347 (0.055)	0.370 (0.070, 0.610)	0.462 (0.020)
Withdrawal	0.323 (0.082)	0.340 (-0.130, 0.730)	0.381 (0.060)
Overall	0.402 (0.025)	0.410 (0.070, 0.710)	0.518 (0.008)

level, whereas between the two assessors, the association is insignificant for 3 criteria.

4.1 Discussion

The results our method achieves when compared to assessor 1 are promising. A significant degree of association for all criteria indicates that certain criteria can indeed be measured automatically from video and motion sensor data. The poor performance when compared to assessor 2, however, suggests that the data we collected may contain outliers. Given the size of the data set and the resulting uncertainty of the results, the problem will have to be addressed with a larger scale experiment. However, the findings of this preliminary study suggest that the proposed system has potential to accurately assess JAG DOPS quality criteria.

REFERENCES

Arnold, M., Ghosh, A., Ameling, S., and Lacey, G. (2010). Automatic segmentation and inpainting of specular highlights for endoscopic imaging. *EURASIP Journal on Image and Video Processing*, 2010:12.

Arnold, M., Ghosh, A., Lacey, G., Patchett, S., and Mulcahy, H. (2009). Indistinct frame detection in colonoscopy videos. In *Proc. 13th Int. Machine Vision and Image Processing Conf. IMVIP '09*, pages 47–52.

Benson, V. S., Patnick, J., Davies, A. K., Nadel, M. R., Smith, R. A., and Atkin, W. S. (2008). Colorectal cancer screening: A comparison of 35 initiatives in 17 countries. *International Journal of Cancer*, 122(6):1357–1367.

Bressler, B., Paszat, L., Chen, Z., Rothwell, D., Vinden, C., and Rabeneck, L. (2007). Rates of new or missed colorectal cancers after colonoscopy and their risk factors: a population-based analysis. *Gastroenterology*, 132(1):96–102.

Hwang, S., Oh, J., Lee, J., Cao, Y., Tavanapong, W., Liu, D., Wong, J., and de Groen, P. (2005). Automatic measurement of quality metrics for colonoscopy videos. In *Proc. 13th annual ACM international conference on Multimedia*, pages 912–921. ACM New York, NY, USA.

Liu, D., Cao, Y., Tavanapong, W., Wong, J., Oh, J., and de Groen, P. (2007). Quadrant coverage histogram: a new method for measuring quality of colonoscopic procedures. In *Proc. 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3470–3473.

Liu, X., Tavanapong, W., Wong, J., Oh, J., and de Groen, P. C. (2010). Automated measurement of quality of mucosa inspection for colonoscopy. *Procedia Computer Science*, 1(1):951 – 960. ICCS 2010.

Matas, J., Chum, O., Urban, M., and Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761 – 767. British Machine Vision Computing 2002.

Oh, J., Hwang, S., Cao, Y., Tavanapong, W., Liu, D., Wong, J., and de Groen, P. (2009). Measuring objective quality of colonoscopy. *IEEE Transactions on Biomedical Engineering*, 56(9):2190–2196.

Oh, J., Hwang, S., Lee, J., Tavanapong, W., Wong, J., and de Groen, P. (2007). Informative frame classification for endoscopy video. *Medical Image Analysis*, 11(2):110–127.

U.S. Preventive Services Task Force (2008). Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. *Annals of Internal Medicine*, 149(9):627–637.

WGO (2007). World gastroenterology organisation / international digestive cancer alliance practice guidelines: Colorectal cancer screening. http://www.worldgastroenterology.org/colorectal-cancer-screening.html: [Last accessed: 6 Dec 2012].