Evaluation of Cross-validation Protocols for the Classification of Endoscopic Images of Colonic Polyps

M. Häfner [1], M. Liedlgruber [2,*], S. Maimone [2], A. Uhl [2], A. Vécsei [3], and F. Wrba [4]

[1] Department for Internal Medicine, St. Elisabeth Hospital, Vienna
[2] Department of Computer Sciences, University of Salzburg, Austria
[3] St. Anna Children's Hospital, Vienna, Austria
[4] Department of Clinical Pathology, Medical University of Vienna, Austria

[*] *Corresponding author e-mail: mliedl@cosy.sbg.ac.at*

Abstract

We evaluate different cross-validation (CV) protocols for an automated classification of colonic polyps. For this purpose we select six previously developed methods which achieved promising results already in the past. We then evaluate the methods using the cross-validation protocols leave-one-image-out (LOO-CV), leave-one-parent-image-out (LOPIO-CV), leave-one-lesion-out (LOLO-CV), and leave-one-patient-out (LOPO-CV).

We show that, in general, the more restrictive cross-validation protocols lead to high results drops. While in case of LOO-CV the accuracies are rather high across all methods evaluated, the picture changes the more strictness a cross-validation mode imposes on the set of training images.

1 Introduction

Colonic polyps have a rather high prevalence and are known to either develop into cancer or to be precursors of colon cancer. Hence, an early detection of such polyps is important as this can lower the mortality rate drastically.

The current gold standard for the examination of the colon is colonoscopy, performed by using a colonoscope. Modern endoscopy devices are able to take pictures from inside the colon, allowing to obtain images for a computer-assisted analysis with the goal of detecting abnormalities. To be able to acquire highly detailed images a magnifying endoscope can be used [1]. Such an endoscope represents a significant advance by providing images which are up to 150-fold magnified, thus uncovering the fine surface structure of the mucosa as well as small lesions.

In the past, we developed various different methods for the classification of colonic polyps in high-magnification chromo-colonoscopy which delivered promising results already (e.g. [3–6, 10]). The image databases used in these approaches were rather limited in terms of the database size, hindering an accurate assessment of the system accuracies. Hence, we mainly used the leave-one-image-out CV (LOO-CV) protocol to estimate the accuracies of our methods. However, in a recent study Hegenbart et al. showed that LOO-CV is prone to biased results [7].

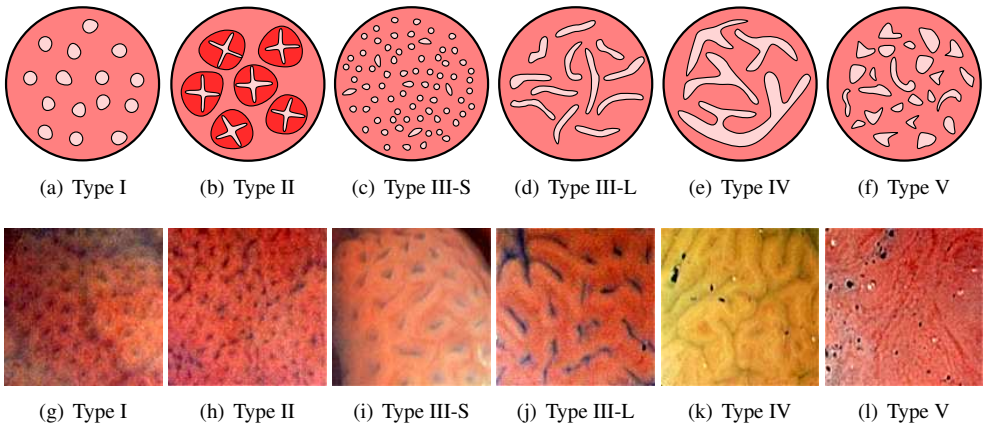
In this work we therefore aim at a comparison of different CV protocols. For this purpose we selected a set of six methods for polyp classification which delivered promising results already in the past. We evaluate each of these methods with the CV protocols leave-one-image-out, leave-one-parent-image-out (LOPIO-CV), leave-one-lesion-out (LOLO-CV), and leave-one-patient-out (LOPO-CV). We then compare the differences between the protocols in terms of the classification rates achieved.

The remaining part of this work is organized as follows: in Section 2 we provide the medical background of this work. This is followed by a brief summary of the previously developed methods and a description of the CV modes compared in Section 3. In Section 4 we give details about the experimental setup and discuss the results obtained. Section 5 concludes this paper.

2 Medical background

Due to the fact that colonic polyps have a rather high prevalence and are known to either develop into cancer or

This work is partially funded by the Austrian Science Fund (FWF) under Project No. L366-N15 and Project No. TRP-206 and by the Austrian National Bank "Jubiläumsfonds" Project No. 12514.



(a) Type I (b) Type II (c) Type III-S (d) Type III-L (e) Type IV (f) Type V

(g) Type I (h) Type II (i) Type III-S (j) Type III-L (k) Type IV (l) Type V

Figure 1. Illustration of (a)-(f) the different pit pattern types according to Kudo et al. and (g)-(l) example images for each pit type.

to be precursors of colon cancer, an early detection of such pathologies can lower the mortality rate drastically. Hence, automated classification systems targeted at the assessment of the malignant potential of colonic polyps aim at avoiding random and, probably, unnecessary biopsies. As a consequence such systems could potentially help to save time, lower the cost for colonoscopy procedures, and reduce the risk of complications during such procedures.

One classification scheme, commonly used to to distinguish between the different types of polyps, is the pit pattern classification scheme, originally introduced by Kudo et al. [9]. Based on the visual pattern of the mucosal surface this system allows to differentiate between normal mucosa, hyperplastic lesions (non-neoplastic), adenomas (a pre-malignant condition), and malignant cancer. A schematic illustration of the pit pattern classification along with example images for each pit type are given in Fig. 1.

In this work we focus on a classification between non-neoplastic (types I and II) and neoplastic lesions (types III-S to V) and a 3-class classification according to [8] in this work. This classification groups the six different pit pattern types into normal lesions (types I and II), non-invasive lesions (types III-S, III-L, and IV), and invasive lesions (type V). This scheme is of particular interest since normal mucosa needs not to be removed, non-invasive lesions must be removed endoscopically, and invasive lesions must not be removed endoscopically.

3 Methodology

3.1 Evaluated Methods

In the past we developed various different methods for the classification of colonic polyps. For the experiments in this work we selected a subset of six methods, which have already shown to yield very promising classification results. In the following we briefly describe these methods (the classifiers and color spaces used to evaluate the methods along with a rough indicator for the dimensionality of the features used are given in brackets):

- **WT-BBC** (Bayes classifier, RGB, Moderate)
  The Best Basis Centroids method employs the Best-basis algorithm [2] to find an optimal basis for each training image and computes a centroid over all resulting bases. After transforming all images into this basis, the most informative subbands (with respect to a cost function) are used to compute statistical features based on the respective coefficients [10].

- **WT-DWT** (Bayes classifier, RGB, Moderate)
  This method transforms an image to the wavelet domain using the discrete wavelet transform. From the most informative subbands (according to a cost function) statistical features are extracted from the respective coefficients [10].

- **JC-MB-LBP** (k-NN classifier, CIELAB, High)
  This method is based on a noise-insensitive extension of the LBP operator [12]. This operator is applied to two color channels of an image and, based on the transformed channels, a 2D joint-color histogram is created for each image [3].

- **LCVP MR A**$^{(2)}$ (k-NN classifier, CIELAB, High)
  The LCVP method is based on a color-extension to the original LBP operator. By treating each pixel as a 3D color vector, each image is treated as a color vector field. Then, similar to LBP, neighboring color vectors are compared, which results in an LCVP-transformed image. Based on the transformation result at different scales, a 1D histogram is created for each scale. These histograms are then concatenated [6].

- **Weibull** (k-NN classifier, RGB, High)
  The dual-tree complex wavelet transform (DTCWT) [13] is used to decompose an image. Based on the resulting detail subband coefficient magnitudes the empirical histogram is modeled by two-parameter Weibull distributions. The Weibull parameters are then arranged into a feature vector [5].

- **Edgefeatures** (k-NN classifier, RGB, Low)
  This method is based on the detection of pit candidates. Once found, various shape and texture features are extracted for the pit candidates within an image. This method also employs a feature selection strategy to reduce the dimensionality of the final feature vectors [4].

The combination of the classifier and color space used is the one which yielded the highest overall classification results among different combinations. Since the optimal parameter configurations for the methods vary between the different CV protocols and the different classification cases considered (2-classes and 3-classes), we only provide rough indicators for the dimensionality of the underlying feature vectors for an image for each method ("low" corresponds to less than 50 features, "moderate" corresponds to a dimensionality between 50 and 100 features, and "high" corresponds to more than 100 features).

3.2 Cross-validation Protocols

A particular problem when classifying endoscopic images is the image database at hand which is quite often limited in terms of the number of images available [11], making an accurate estimation of system accuracies problematic. One common way to solve this problem are CV protocols, which allow to evaluate a classification system in a meaningful way. Depending on the image database used, different ways for cross-validation are possible.

The images used in our experiments originate from colonoscopies of different patients. Hence, one or more lesions per patient are present in our image database. Furthermore, one or more images per lesion are present (i.e. parent images), showing the respective lesion from different viewing angles. This results in an implicit hierarchy inherent to our database, which is depicted schematically in Fig. 2. To generate more images for the experiments we manually cut out one or more small patches (256×256 pixels) from each parent image, allowing us to use the following CV modes:

- **Leave-One-Image-Out (LOO-CV)**
  In this protocol one patch is used as a validation sample while the remaining patches are used to train a classifier (repeated for each patch in the image database). While LOO-CV is quite frequently used it is also prone to biased results, since, if there is more than one patch present for a parent image, these patches usually exhibit a high similarity (for example, patches 1 and 2 in Fig. 2). Hence, the classifier is trained with patches being very similar to the patch currently classified.

- **Leave-One-Parent-Image-Out (LOPIO-CV)**
  To overcome the problem of similar patches extracted from the same parent image, LOPIO-CV has the restriction that the classifier must not be trained with patches extracted from the parent image the patch under classification has been extracted from. But as we notice from Fig. 2, if there are multiple parent images showing the same lesion, we again run into the problem of training a classifier with patches similar to the patch under classification (for example, patches 2 and 4 in Fig. 2).

- **Leave-One-Lesion-Out (LOLO-CV)**
  To avoid training a classifier with patches of the lesion currently classified, LOLO-CV is even more restrictive than LOPIO-CV. Here the classifier must not be trained with patches which belong to the same lesion as the patch currently classified. Since different lesions within a patient are less likely to show high similarities LOLO-CV is sufficient for an accuracy estimation without any bias from similar images.

- **Leave-One-Patient-Out (LOPO-CV)**
  LOPO-CV is even more restrictive as it prohibits classifier training with patches which belong to the same patient as the patch currently classified (even if the patches belong to different image classes). But this protocol is also the most realistic one since in clinical practice there are usually no images available for a patient who undergoes colonoscopy (except for follow-up examinations).

4 Experiments

4.1 Experimental Setup

The image database used is based on 327 endoscopic color images (either of size 624×533 pixels or 586×502 pixels) acquired between the years 2005 and 2009 at the Department of Gastroenterology and Hepatology (Medical University of Vienna) using a zoom-colonoscope (Olympus Evis Exera CF-Q160ZI/L) with a magnification factor of 150. To acquire the images 40 patients underwent colonoscopy. Extracting patches from the original images resulted in an extended image set containing 716 images.

Lesions found during colonoscopy have been examined after application of dye-spraying with indigocarmine. Biopsies or mucosal resection have been performed in order to get a histopathological diagnosis. Biopsies have been taken

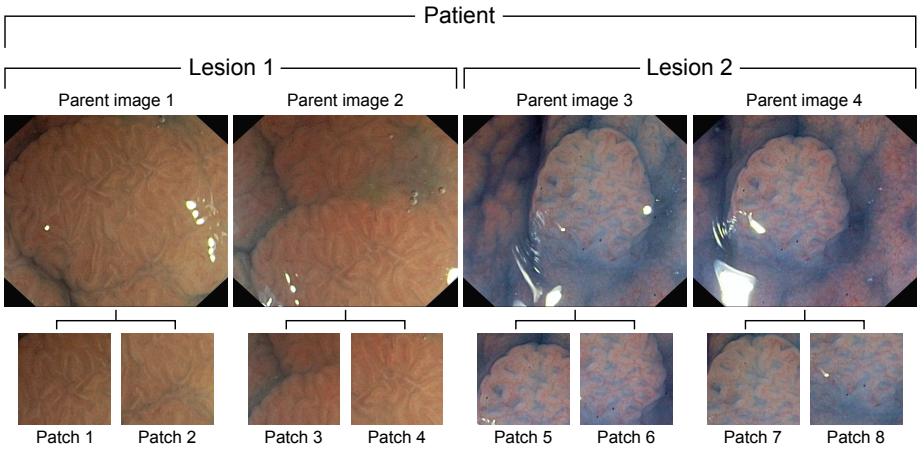


Figure 2. Illustration of the hierarchy in our image database (based on some sample images).

Histology	Pit Pattern	3 classes				2 classes			
		$\mathbf{N_O}$	$\mathbf{N_E}$	$\mathbf{N_P}$	$\mathbf{N_L}$	$\mathbf{N_O}$	$\mathbf{N_E}$	$\mathbf{N_P}$	$\mathbf{N_L}$
Normal	I	72	198	14	55	72	198	14	55
Hyperplasia	II								
Tubular adenoma Tubulovillous adenoma	III-L	212	420	27	100	255	518	32	129
Serrated adenoma Tubular adenoma	III-S								
Adenoma Tubular adenoma Tubulovillous adenoma	IV								
Adenocarcinoma Carcinoma Lymphoma	V	43	98	6	29				
	Total	**327**	**716**	**47**	**184**	**327**	**716**	**46**	**184**

Table 1. Detailed ground truth information for the image database used throughout our experiments.

from type I, II, and type V lesions, while type III and IV lesions have been removed endoscopically.

Table 1 shows the ground truth used, where $\mathbf{N_O}$, $\mathbf{N_E}$, $\mathbf{N_P}$, and $\mathbf{N_L}$ denote the number of original images, the number of images in the extended image set, the number of patients, and the number of different lesions, respectively. Since different types of lesions may develop in a patient, some patients appear in more than one class. Thus, the number of patients in Table 1 is slightly higher than the number of patients who underwent colonoscopy.

4.2 Results

Figures 3 and 4 show the overall rates achieved by our methods for different CV protocols. The overall picture is as expected: the rates drop more in case of more restrictive CV protocols. Hence, the highest classification rates have been achieved with LOO-CV, followed by LOPIO-CV, LOLO-CV, and LOPO-CV. Only in case of the Edgefeatures method we see a slight increase in the 2-classes case when using LOLO-CV instead of LOPIO-CV and in the 3-classes case when using LOPIO-CV instead of LOO-CV. But this can be attributed to the feature selection used in case of this method, which results in moderate result drops only when using more restrictive CV protocols. We also notice that the highest result drops between LOO-CV and LOPO-CV occur when using rather high-dimensional features (JC-MB-LBP, LCVP MR $A^{(2)}$, and Weibull). This indicates that those methods are also prone to overfitting.

The detailed classification results of our experiments can be found in tables 2 and 3. From Table 2 we notice that the sensitivity values are rather stable for each method across the different CV protocols (with result drops up to about 3%). The specificities, on the other hand, are subject to high fluctuations with result drops up to about 33%.

In the 3-classes case we notice a similar behavior (see Table 3). While results for the non-invasive images are rather stable when using LOPO-CV instead of LOO-CV (results – also across image classes – drop up to about 14%), the results drop significantly by up to about 32% for the normal images. The highest result drops can be observed for the invasive images (up to about 80%).

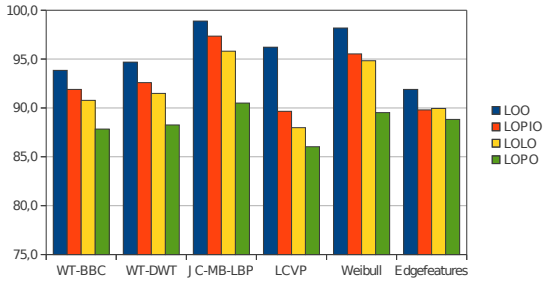


Figure 3. Overall accuracies yielded by the methods used and the different CV modes evaluated (2-classes case).

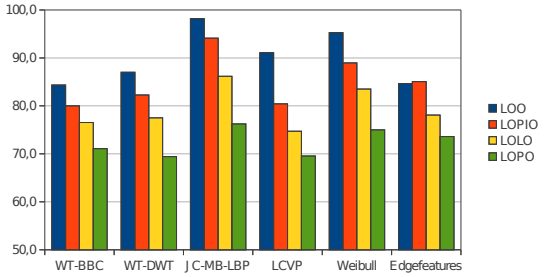


Figure 4. Overall accuracies yielded by the methods used and the different CV modes evaluated (3-classes case).

CV mode	Specificity	Sensitivity	Overall
WT-BBC [10]			
LOO	90.9	95.0	93.9
LOPIO	83.8	95.0	91.9
LOLO	79.3	95.2	90.8
LOPO	72.7	93.6	87.8
WT-DWT [10]			
LOO	91.9	95.8	94.7
LOPIO	87.9	94.4	92.6
LOLO	86.4	93.4	91.5
LOPO	79.3	91.7	88.3
JC-MB-LBP [3]			
LOO	98.0	99.2	98.9
LOPIO	95.5	98.1	97.3
LOLO	93.9	96.5	95.8
LOPO	75.3	96.3	90.5
LCVP MR $A^{(2)}$ [6]			
LOO	89.9	98.6	96.2
LOPIO	65.7	98.8	89.7
LOLO	60.6	98.5	88.0
LOPO	57.6	96.9	86.0
Weibull [5]			
LOO	96.0	99.0	98.2
LOPIO	87.9	98.5	95.5
LOLO	84.3	98.8	94.8
LOPO	66.2	98.5	89.5
Edgefeatures [4]			
LOO	76.8	97.7	91.9
LOPIO	74.8	95.6	89.8
LOLO	74.2	96.0	89.9
LOPO	71.7	95.4	88.8

Table 2. Detailed classification results for our methods when evaluated with the different CV modes (2-classes case).

One explanation for the observed result discrepancies of the single class results is the unbalanced image set. In addition, depending on the CV protocol, the training set gets reduced more or less. This is especially noticeable in case of LOPO-CV, where excluding one patient from the training set results in a decrease of up to 49 images (in case of LOLO-CV and LOO-CV only up to 23 and 1 images are removed from the training set, respectively). Hence, the reduced training set has a more severe impact on the classification in case of classes containing fewer images. In case of LOO-CV it is also very likely that training patches exist which are very similar to the patch classified (i.e. high chance of biased results). As a consequence, we see rather high overall results as well as high single class accuracies. But the more restrictions we put on the training set the smaller is the chance of such similarities, which is again most striking when comparing LOO-CV and LOPO-CV.

But, as already pointed out in Section 3, using LOLO-CV is usually sufficient for realistic classification accuracy estimates. Hence, it is not necessary to use the more strict LOPO-CV. In addition, the number of training samples is unnecessarily low in case of LOPO-CV, which is a reasonable explanation for the LOPO-CV result drops observed, considering the limited number of images in our image database. This is also supported by the fact that in case of LOPO-CV we loose up to about 36% of the training samples while for all other CV modes the respective fractions are considerably lower (see Table 4). Since in case of LOPO-CV all images of a patient across all classes are left out from training the total loss of training samples may be even higher in case of some patients.

5 Conclusion

In this work we compared six previously developed methods for an automated classification of colonic polyps. Due to the limited number of images in our image database we have put the main emphasis on the comparison of different CV protocols to overcome this problem.

We showed that, while in case of LOO-CV most methods are able to deliver rather high classification accuracies, the picture changes rapidly when using CV protocols which impose limits on the training samples available. The loss in terms of the classification results is especially noticeable for the classification results for the single classes.

We identified the following three main reasons for the observed behavior: first, especially in case of very restrictive CV protocols (i.e. LOPO-CV) the imbalance across the classes in our image database leads to an insufficient train-

CV mode	Normal	Non-invasive	Invasive	Overall
WT-BBC [10]				
LOO	89.9	87.9	58.2	84.4
LOPIO	83.3	87.6	40.8	80.0
LOLO	82.3	84.8	29.6	76.5
LOPO	67.7	89.3	0.0	71.1
WT-DWT [10]				
LOO	93.4	91.7	54.1	87.0
LOPIO	90.4	87.9	41.8	82.3
LOLO	88.4	85.0	23.5	77.5
LOPO	79.8	80.7	0.0	69.4
JC-MB-LBP [3]				
LOO	98.0	98.8	95.9	98.2
LOPIO	95.5	96.9	79.6	94.1
LOLO	93.9	89.8	55.1	86.2
LOPO	75.8	88.8	23.5	76.3
LCVP MR $A^{(2)}$ [6]				
LOO	89.9	95.0	76.5	91.1
LOPIO	67.7	91.2	60.2	80.4
LOLO	63.6	86.0	49.0	74.7
LOPO	59.6	81.0	40.8	69.6
Weibull [5]				
LOO	96.0	96.2	89.8	95.3
LOPIO	88.9	95.5	61.2	89.0
LOLO	84.3	95.2	31.6	83.5
LOPO	63.6	95.5	10.2	75.0
Edgefeatures [4]				
LOO	84.9	93.8	44.9	84.6
LOPIO	83.8	94.5	46.9	85.1
LOLO	78.3	93.6	11.2	78.1
LOPO	72.7	91.2	0.0	73.6

Table 3. Detailed classification results for our methods when evaluated with the different CV modes (3-classes case).

CV mode	Non-neoplastic	Neoplastic
LOO	1 (<1%)	1 (<1%)
LOPIO	1-6 (3%)	1-15 (3%)
LOLO	1-21 (11%)	1-23 (4%)
LOPO	2-46 (23%)	2-49 (9%)

CV mode	Normal	Non-invasive	Invasive
LOO	1 (<1%)	1 (<1%)	1 (1%)
LOPIO	1-6 (3%)	1-15 (4%)	1-7 (7%)
LOLO	1-21 (11%)	1-23 (5%)	1-13 (13%)
LOPO	2-46 (23%)	2-39 (9%)	5-35 (36%)

Table 4. Number of training samples lost per class and CV mode (maximum values are given as fractions in brackets).

ing of the respective classes. Second, in our case the patch-to-patient and patch-to-lesion ratios are rather high, leading to biased results in case of LOO-CV. When using more strict protocols the chance of classifying an image and having a very similar image in the training set vanishes. Third, especially methods, which are based on rather high-dimensional features, suffer from more restrictive protocols. This indicates that those methods are also prone to overfitting.

But we must point out that the results presented are just rough estimates since the low number of images available leads to inaccurate accuracy estimates because restrictive CV modes in some cases greatly reduce the training set.

References

[1] M. J. Bruno. Magnification endoscopy, high resolution endoscopy, and chromoscopy; towards a better optical diagnosis. *Gut*, 52(4):7–11, 2003.

[2] R. Coifman and M. Wickerhauser. Entropy based methods for best basis selection. *IEEE Transactions on Information Theory*, 38(2):719–746, 1992.

[3] M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba. Pit pattern classification using extended local binary patterns. In *Proceedings of the 9th International Conference on Information Technology and Applications in Biomedicine (ITAB'09)*, pages 1–4, 2009.

[4] M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba. Endoscopic image classification using edge-based features. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR'10)*, pages 2724–2727, 2010.

[5] M. Häfner, R. Kwitt, F. Wrba, A. Gangl, A. Vécsei, and A. Uhl. One-against-one classification for zoom-endoscopy images. In *Proceedings of the 4th International Conference on Advances in Medical Signal and Information Processing (MEDSIP'08)*, pages 1–4, 2008.

[6] M. Häfner, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba. Color treatment in endoscopic image classification using multi-scale local color vector patterns. *Medical Image Analysis*, 16(1):75–86, 2012.

[7] S. Hegenbart, A. Uhl, and A. Vécsei. Systematic assessment of performance prediction techniques in medical image classification - a case study on celiac disease. In *Proceedings of the 22nd International Conference on Information Processing in Medical Imaging (IPMI'11)*, pages 498–508, 2011.

[8] S. Kato, K.-I. Fu, Y. Sano, T. Fujii, Y. Saito, T. Matsuda, I. Koba, S. Yoshida, and T. Fujimori. Magnifying colonoscopy as a non-biopsy technique for differential diagnosis of non-neoplastic and neoplastic lesions. *World Journal of Gastroenterology*, 12(9):1416–1420, 2006.

[9] S. Kudo, S. Hirota, T. Nakajima, S. Hosobe, H. Kusaka, T. Kobayashi, M. Himori, and A. Yagyuu. Colorectal tumours and pit pattern. *Journal of Clinical Pathology*, 47(10):880–885, 1994.

[10] M. Liedlgruber and A. Uhl. Statistical and structural wavelet packet features for Pit pattern classification in zoom-endoscopic colon images. In *Proceedings of the 7th WSEAS International Conference on Wavelet Analysis & Multirate Systems (WAMUS'07)*, pages 147–152, 2007.

[11] M. Liedlgruber and A. Uhl. Computer-aided decision support systems for endoscopy in the gastrointestinal tract: A review. *IEEE Reviews in Biomedical Engineering*, 4:73–88, 2012.

[12] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[13] I. W. Selesnick, R. G. Baraniuk, and N. G. Kingsbury. The dual-tree complex wavelet transform - a coherent framework for multiscale signal and image processing. *IEEE Signal Processing Magazine*, 22(6):123–151, 2005.