Improving Pit–Pattern Classification of Endoscopy Images by a Combination of Experts

Michael Häfner[1], Alfred Gangl[1], Roland Kwitt[2], Andreas Uhl[2], Andreas Vécsei[4], and Friedrich Wrba[3]

[1] Dept. of Gastroenterology & Hepatology, Medical University of Vienna, Austria
[2] Dept. of Computer Science, University of Salzburg, Austria
{rkwitt,uhl}@cosy.sbg.ac.at
[3] Dept. of Clinical Pathology, Medical University of Vienna, Austria
[4] St. Anna Children's Hospital, Vienna, Austria

Abstract. The diagnosis of colorectal cancer is usually supported by a staging system, such as the Duke or TNM system. In this work we discuss computer–aided pit–pattern classification of surface structures observed during high–magnification colonoscopy in order to support dignity assessment of colonic polyps. This is considered a quite promising approach because it allows in vivo staging of colorectal lesions. Since recent research work has shown that the characteristic surface structures of the colon mucosa exhibit texture characteristics, we employ a set of texture image features in the wavelet-domain and propose a novel classifier combination approach which is similar to a combination of experts. The experimental results of our work show superior classification performance compared to previous approaches on both a two-class (non-neoplastic vs. neoplastic) and a more complicated six-class (pit–pattern) classification problem.

1 Motivation

Recent statistics of the American Cancer Society reveal that colorectal cancer is the third most common cancer in men and women and the second most common cause of US cancer deaths. Since most colorectal cancers develop from polyps, a regular inspection of the colon is recommended in order to detect lesions with a malignant potential or early cancer. A common medical procedure to examine the inside of the colon is colonoscopy, which is usually carried out with a conventional video–endoscope. A diagnostic benefit can be achieved by employing so called high–magnification endoscopes (aka zoom–endoscopes), which achieve a magnification factor of up to 150 by means of an individually adjustable lens. In combination with dye–spraying to enhance the visual appearance (chromo–endoscopy) of the colon mucosa, high–magnification endoscopy can reveal characteristic surface patterns, which can be interpreted by experienced physicians. Commonly used dyes are either methylene-blue, or indigo–carmine, which both lead to a plastic effect. In the research work of Kudo et al. [1], the macroscopic appearance of colorectal polyps is systematically described and results in the so called *pit–pattern* classification scheme.

G.-Z. Yang et al. (Eds.): MICCAI 2009, Part I, LNCS 5761, pp. 247–254, 2009.

The contribution of this work is a novel way for classifier combination to enhance the accuracy of differential diagnosis. We propose a fusion of three approaches from classification research and show that by using several recently proposed texture image features for endoscopy image classification, we achieve a remarkable increase in overall classification accuracy.

The remainder of this paper is structured as follows: in Sect. 2, we review the medical background an introduce the pit–pattern classification scheme. Section 3 discusses the feature extraction step as well as our classification approach. In Sect. 4, we present the experimental results of our work and Sect. 5 concludes the paper with a short summary and an outlook on future research.

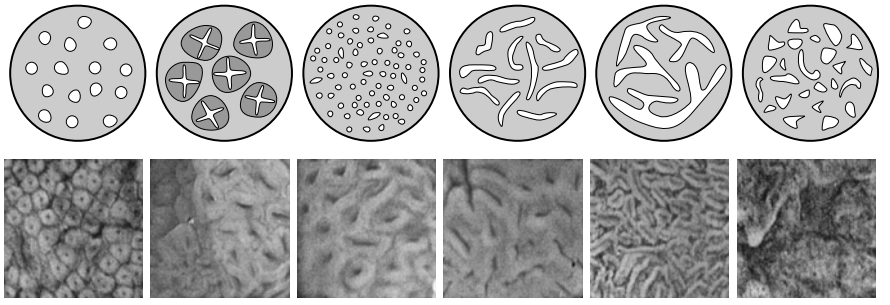
2 Pit-Pattern Classification

Polyps of the colon are a frequent finding and are usually divided into metaplastic, adenomatous and malignant. Since the resection of all polyps is rather time–consuming, it is imperative that those polyps which warrant resection can be distinguished. Furthermore, polypectomy[1] of metaplastic lesions is unnecessary and removal of invasive cancer may be hazardous. The classification scheme presented in [1] divides the mucosal crypt patterns into five types (pit–patterns I–V, see Fig. 1), which can be observed using a high–magnification endoscope.

While types I and II are characteristic of benign lesions and represent normal colon mucosa or hyperplastic polyps (non–neoplastic lesions), types III to V represent neoplastic, adenomatous and carcinomatous structures. Our classification problem can be stated as follows: the problem of differentiating pit–patterns I and II from III–L,III–S,IV and V will be denoted as the *two–class* problem (non–neoplastic vs. neoplastic), whereas the more complex and detailed discrimination of all pit–patterns I to V will be denoted as the *six–class* problem. At first sight, the pit–pattern classification scheme seems to be straightforward and easy to be applied. Nevertheless, it needs some experience and exercising to achieve good results. Correct diagnosis very much relies on the experience of the endoscopist as the interpretation of the pit–patterns may be challenging [2].

Our approach is motivated by the work of Kato et al. [3], where the authors state that assessing the type of mucosal crypt patterns can actually predict the histological findings to a very high accuracy. Regarding the correlation between the mucosal pit–patterns and the histological findings, several studies reported good results, although with quite different diagnostic accuracies. A comparative study by Kato et al. [4] shows that the classification accuracy in magnifying colonoscopy ranges from 80.6% to 99.1%. Another extensive study by Hurlstone et al. [5] report error rates of approximately 5%. In [6] the authors claim 95.6% for chromoendoscopy with magnification in contrast to diagnosis using conventional colonoscopy (84.0%) and diagnosis using chromoendoscopy without magnification (89.3%). In addition to that, inter–observer variability of magnification chromoendoscopy has been described at least for Barret's esophagus [7]. This inter–observer variability may to lesser degree be also present in

[1] The process of removing polyps.



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

Fig. 1. Schematic illustration of the pit–pattern characteristics (top row) together with exemplary pit–pattern images obtained during high–magnification colonoscopy (bottom row)

the interpretation of pit–patterns of colonic lesions. This work aims at allowing computer–assisted pit–pattern classification in order to enhance the quality of differential diagnosis.

3 Feature Extraction and Classification

We use a selection of texture feature extraction approaches in the wavelet–domain which have already been successfully applied in the context of endoscopy image classification. Our first feature set is computed using an approach presented in [8] where the authors decompose each image using the Dual–Tree Complex Wavelet Transform (DT–CWT) and model the absolute values of detail subband coefficients by two–parameter Weibull distributions. The Maximum–Likelihood estimates of the Weibull scale & shape parameter of each subband are then arranged into feature vectors for nearest–neighbor classification using the Euclidean distance. Although the work in [8] uses grayscale images, it can easily be extended for color images by simple feature vector concatenation of separately computed color–channel feature vectors. In another work [9], a set of image features is computed from both the classic pyramidal (DWT) and undecimated wavelet transform (SWT) by calculating so called color eigen–subband features (CES). The CES features are essentially the eigenvalues obtained during PCA on the stacked wavelet detail subbands of the color–channels. Since the study in [8] has shown that the DT–CWT produces highly discriminative features, we have extended the CES approach to work with the DT–CWT. The last feature extraction approach we take into account is presented in [10] and it based on the computation of so called color wavelet–energy correlation signatures (WECS) between wavelet–decomposed color–channels. The extension to the DT–CWT is again straightforward and the dimensionality of the feature vectors is doubled. In the following, feature vectors will be denoted by $\mathbf{v} \in \mathcal{F} \subset \mathbb{R}^d$, where $\mathcal{F}$ denotes the feature space and d denotes the feature space dimensionality.

3.1 Classification Setup

We propose a fusion of three main approaches from recent work in classification research. We combine nearest–neighbor (NN) classifiers in a One–Against–One [11] (aka round–robing or pairwise coupling) setup and optimize each classifier using sequential forward feature subset selection (SFFS) [12]. The different class predictions are then combined using a *voting–against* approach and class posterior probability estimation.

The One–Against–One Approach (OAO). The basic concept of the OAO classification strategy is to split a multi–class problem into smaller binary problems with the ulterior motive that the decision boundaries for binary problems are simpler and easier to learn than the decision boundaries in case of a multi–class problem. In OAO classification, one classifier is trained for each possible pair of classes. Given a c–class problem, we thus have to train a total of $c(c-1)/2$ classifiers. In the training stage, each binary classifier is trained using only the examples of the two classes it has to discriminate. For example, given that $V := \{\mathbf{v}_i\}_{1 \leq i \leq L}$ denotes the complete set of training vectors and $\gamma : \mathcal{F} \rightarrow \{1, \ldots, c\}$ denotes a function returning the true class membership of $\mathbf{v}$, then the binary classifier $C_{ij} : \mathcal{F} \rightarrow \{i, j\}$ – which discriminates between class i and j – is trained using the training subset $S_{ij} := \{\mathbf{v}_n | \gamma(\mathbf{v}_n) = i \vee \gamma(\mathbf{v}_n) = j\}$. Hence, we can view the OAO approach as some sort of expert system, where each classifier is an expert in discriminating only two particular classes. In such a classifier combining approach we require two important properties of the base classifiers [13]: *diversity* and *accuracy*. First, diversity signifies that the errors should be uncorrelated and second, accuracy refers to a classification accuracy of at least 50%. In pratical use, we also require efficiency, which refers to low computational cost. However, due to space limitations we will not deal with this issue here.

Increasing Diversity. The problem of combining NN classifiers is particularly interesting in the context of the diversity requirement, since the main approaches of classifier combination to increase diversity, such as bagging [14] or boosting [15] do not lead to the desired results. The root of the problem is the insensitivity of the NN classifier to changes in the training patterns, which is essentially the starting point for both bagging and boosting. In [16] this issue is discussed in great detail and a new random feature subset selection approach is proposed, where each classifier works on a random subset of all available features. Due to the sensitivity of the NN classifier w.r.t. changes in the feature set, this approach can increase diversity. Instead of using random feature subsets, we select each subset by means of SFFS, imposing no limit on the size of the resulting subsets. Starting with a subset of cardinality one, one feature is added in each iteration in case this feature improves leave–one–out crossvalidation accuracy. By using SFFS, we cover both requirements of accuracy and diversity at the same time. In combination with the OAO approach, we obtain $c(c-1)/2$ feature subsets after the training stage.

Combining Class Predictions. Since each classifier in our OAO ensemble will provide a class prediction, the question arises of how to combine the $c(c-1)/2$ predictions. Although it seems straightforward to employ a simple majority voting rule, this rule is logically incorrect w.r.t. OAO classification for one simple reason: given an arbitrary sample $\mathbf{v} \in \mathcal{F}$, a classifier C_{ij} will output either i or j as the predicted class label. However, this prediction is convenient, if and only if the sample $\mathbf{v}$ actually belongs to either class i or j. In that case the prediction is termed a *qualified* prediction. Otherwise, the prediction is termed an *unqualified* prediction. As a consequence, given that $C_{ij}(\mathbf{v}) = i$, we can at best conclude that $\mathbf{v}$ is not a member of class j. This interpretation is known as *voting against* [17] in contrast to *voting for*, which is correct only in case each classifier was trained to discriminate samples from all classes. The final prediction is obtained by counting the votes against each class and selecting the very one which received the smallest number of votes–against. Although the idea of voting–against seems to be pedantic at first sight, it allows a quite efficient way to compute the final prediction [11] and enables us to compute a closed–form estimation of the class posterior probabilities $P(i|\mathbf{v})$ [17]. Given that ϵ_{ji} is defined as the probability $P(C_{ij}(\mathbf{v}) = i|\gamma(\mathbf{v}) = j)$ (i.e. classifier C_{ij} outputs i though the sample belongs to class j) and w_i denotes the a–priori class probability of class i, the logarithm of the class posterior probability of class i can be calculated by

$$\begin{aligned} \log P(i|\mathbf{v}) = K + \log w_i + \\ \sum_{i \neq j} \log \left(\epsilon_{ji}, \text{ if } C_{ij}(\mathbf{v}) = j; \frac{1 - \epsilon_{ij} w_j}{1 - w_j} \text{ if } C_{ij}(\mathbf{v}) = i \right) + \\ \sum_{k, i \neq j} \log \left(\frac{1 - \epsilon_{kj} w_j}{1 - w_j} \text{ if } C_{kj}(\mathbf{v}) = k; \frac{1 - \epsilon_{jk} w_k}{1 - w_k} \text{ if } C_{kj}(\mathbf{v}) = j \right) \quad . \end{aligned} \tag{1}$$

The error terms ϵ_{ji} can be easily estimated in the training stage of the system from the outputs of classifier C_{ij} when presenting samples $\mathbf{v}$, $\gamma(\mathbf{v}) \neq i, j$. Further, the term K is simply a constant which is of no particular relevance for determining the final prediction. By using (1) we determine the predicted class label k (or equivalently the predicted in vivo staging of the endoscopy image) of a feature vector $\mathbf{v}$ by $k = \arg\max_i \log(P(i|\mathbf{v}))$.

4 Experimental Results

Our image database contains 484 RGB images of size 256×256, acquired in 2005/2006 at the Department of Gastroenterology and Hepatology (Medical

Table 1. Number of image samples per pit–pattern

I	II	III–L	III–S	IV	V
126	72	62	18	146	60

Table 2. Classification accuracy results for six different feature sets and the two combining approaches together with the McNemar–test results

	Voting–Against	**Cutzu [17]**		**Original Work**	
DWT & WECS [10]	90.08	91.53	+	79.96	++
DT–CWT & WECS [10]	95.04	95.25	−	86.57	++
DT–CWT & Classic [8]	96.69	97.11	−	93.18	++
DT–CWT & Weibull [8]	97.11	97.31	−	94.01	++
DT–CWT & CES [9]	97.31	97.73	−	88.84	++
DWT & CES [9]	93.18	94.43	+	84.09	++

University of Vienna) using a magnification endoscope (Olympus Evis Exera CF–Q160ZI/L) with a magnification factor of 150x. To enhance visual appearance, dye–spraying with indigo–carmine was applied and biopsies or mucosal resections were taken to obtain a histopathological diagnosis (*our ground truth*). For pit–patterns I,II and V, biopsies were taken, since these types need not be removed. Lesions of pit–pattern types III–S,III–L and IV have been removed endoscopically. Table 1 lists the number of image samples per class. We use exactly the same setup for feature extraction as presented in the original works [8,9,10], discussed in Sect. 3. The maximum decomposition depth of the wavelet transforms is set to $J = 6$. Regarding the dimensionality d of the resulting feature spaces $\mathcal{F}$ (using the DT–CWT), we obtain $d = 18J$ for [9], $d = 36J$ for [8] and $d = 18J$ for [10]. In case the DWT is used for the WECS approach, we obtain $d = 9J$. Table 2 lists the maximum leave–one–out crossvalidation accuracies for all feature extraction approaches and the two classifier combining schemes compared to the highest accuracies achieved in the original (color–extended) works. Since most of the results – especially between the combining schemes – are very similar, we conduct a McNemar–test [18] to test for statistically significant differences at the 5% significance level. The null–hypothesis H_0 is that there is no significant difference. A '+' indicates a rejection of H_0, while a '−' indicates that H_0 could not be rejected. Column four of Table 2 lists the McNemar–test results when comparing the combining schemes, column six lists the results when comparing the original work to the voting–against (first $\pm$ entry) and class posterior probability estimation (second $\pm$ entry) approach. As we can see, the best overall accuracy is obtained by the DT–CWT & CES features with 97.73%. We further notice that in the majority of cases, there is no significant difference between direct voting–against and class posterior probability estimation. However, compared to the original works, the OAO results are significantly superior with an average increase in leave–one–out crossvalidation accuracy of $\approx 8\%$. To get an impression of the misclassifications per class, Table 3 shows the confusion matrix of the DT–CWT & CES result. By breaking down the six–class problem to the two–class problem (see Sect. 2) we obtain an overall leave–one–out accuracy of 99.59%, which is considerably higher than in the original works. As a last note, we remind that although we use leave–one–out crossvalidation, all reported accuracies are actually *training* accuracies. Since high–magnification endoscopy is a rather new method for the diagnosis of colorectal cancer, there exists a

Table 3. Detailed confusion matrix results for the DT–CWT & CES features using Cutzu's class posterior probability estimation

	I	**II**	**III–S**	**III–L**	**IV**	**V**
I	**123**	3	0	0	0	0
II	4	**67**	0	0	1	0
III–S	0	0	**62**	0	0	0
III–L	0	0	0	**18**	0	0
IV	1	0	0	0	**145**	0
V	0	0	0	0	2	**58**

lack of data material which prevents to separate an independent set of test–images. As a result, it is highly probable that the accuracies are overestimated in a sense. Nevertheless, our results clearly indicate that computer–assisted pit–pattern classification based on the visual appearance of the colon mucosa can predict the histological results to a large extent.

5 Conclusion

In this paper[2], we have exploited the idea of combining a number of two–class classifiers to obtain an diagnostic prediction for high–magnification colonoscopy images. Our results show a remarkable improvement in leave–one–out crossvalidation accuracy compared to previous works. Since most of the computational effort (mainly feature selection) resides in the training stage there is no limiting factor which might prevent practical application. Depending upon the availability of a larger dataset, future research includes an evaluation of the approach using clearly separated training and test sets which is currently impossible.

References

1. Kudo, S.: Colorectal Tumours and Pit Pattern. Journal of Clinical Pathology 47(10), 880–885 (1994)
2. Hurlstone, D.: High-Resolution Magnification Chromoendoscopy: Common Problems Encountered in Pit-Pattern Interpretation and Correct Classification of Flat Colorectal Lesions. American Journal of Gastroenterology 97(4), 1069–1070 (2002)
3. Kato, S., Fujii, T., Koba, I., Sano, Y., Fu, K.I., Parra-Blanco, A., Tajiri, H., Yoshida, S., Rembacken, B.: Assessment of Colorectal Lesions Using Magnifying Colonoscopy and Mucosal Dye Spraying: Can Significant Lesions Be Distinguished? Endoscopy 33(3), 306–311 (2001)
4. Kato, S., Fu, K., Sano, Y., Fujii, T., Saito, Y., Matsuda, T., Koba, I., Yoshida, S., Fujimori, T.: Magnifying Colonoscopy as a Non-Biopsy Technique for Differential Diagnosis of Non-Neoplastic and Neoplastic Lesions. World of Gastroenterology 12(9), 1416–1420 (2006)

[2] This project was supported by the Austrian FWF Project, No.366-N15.

5. Hurlstone, D., Cross, S., Adam, I., Shorthouse, A., Brown, S., Sanders, D., Lobo, A.: Efficiacy of High Magnification Chromoscopic Colonoscopy for the Diagnosis of Neoplasia in Flat and Depressed Lesions of the Colorectum: a Prospective Analysis. Gut 53, 284–290 (2004)
6. Fu, K.: Chromoendoscopy using Indigo Carmine Dye Spraying with Magnifying Observation is the most Reliable Method for Differential Diagnosis between Non-Neoplastic and Neoplastic Colorectal Lesions. Endoscopy 36(12), 1089–1093 (2004)
7. Meining, A.: Inter- and Intra-Observer Variability of Magnification Chromoendoscopy for Detecting Specialized Intestinal Metaplasia at the Gastroesophageal Junction. Endoscopy 36(2), 160–164 (2004)
8. Kwitt, R., Uhl, A.: Modeling the marginal distributions of complex wavelet coefficient magnitudes for the classification of Zoom-Endoscopy images. In: Proceedings of the IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2007), Rio de Janeiro, Brazil, pp. 1–8 (2007)
9. Kwitt, R., Uhl, A.: Color Eigen-Subband Features for Endoscopy Image Classification. In: Proceedings of the 33rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008), Las Vegas, Nevada, USA, pp. 589–592 (2008)
10. Van de Wouwer, G., Livens, S., Scheunders, P., Van Dyck, D.: Color Texture Classification by Wavelet Energy Correlation Signatures. In: Proceedings of the 9th International Conference on Image Analysis and Processing (ICIAP 1997), pp. 327–334. Springer, Heidelberg (1997)
11. Park, S.-H., Fürnkranz, J.: Efficient pairwise classification. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 658–665. Springer, Heidelberg (2007)
12. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Morgan Kaufmann, San Francisco (1990)
13. Hansen, L., Salamon, P.: Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(10), 993–1001 (1990)
14. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996)
15. Freund, Y., Schapire, R.: Experiments with a new boosting algorithm. In: Proceedings of the 13th International Conference on Machine Learning (ICML 1996), Bari, Italy, pp. 148–156 (1996)
16. Bay, S.: Nearest Neighbor Classification from Multiple Feature Subsets. Intelligent Data Analysis 3(3), 191–209 (1999)
17. Cutzu, F.: How to do multi-way classification with two-way classifiers. In: 3rd Joint International Conference on Artificial Neural Networks and Neural Information Processing, Istanbul, Turkey, pp. 375–384 (2003)
18. Dietterich, T.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation 10(7), 1895–1923 (1998)