Automated Marsh-like Classification of Celiac Disease in Children using Local Texture Operators

A.Vécsei[a], G. Amann[b], S.Hegenbart[c], M.Liedlgruber[c], A.Uhl[c]

[a] *St. Anna Children's Hospital Vienna, Austria*
[b] *Department of Pathology, Vienna Medical University, Austria*
[c] *Department of Computer Sciences Salzburg University, Austria*

Abstract

Automated classification of duodenal texture patches with histological ground truth in case of pediatric celiac disease is proposed. The classical focus of classification in this context is a two-class problem: mucosa affected by celiac disease and unaffected duodenal tissue. We extend this focus and apply classification according to a modified Marsh scheme into four classes. In addition to other techniques used previously for classification of endoscopic imagery, we apply Local Binary Patterns (LBP) operators and propose two new operator types, one of which adapts to the different properties of Wavelet transform subbands. The achieved results are promising in that operators based on LBP turn out to achieve better results compared to many other texture classification techniques as used in earlier work. Specifically, the proposed wavelet-based LBP scheme achieved the best overall accuracy of all feature extraction techniques considered in the two-class case and was among the best in the four-class scheme. Results also show that a classification into four classes is feasible in principle, however, when compared to the two-class case we note that there is still room for improvement due to various reasons discussed.

Keywords: celiac disease, computer-aided classification, endoscopy, LBP, Marsh classification, children

1. Introduction

Celiac disease is a complex autoimmune disorder in genetically predisposed individuals of all age groups after introduction of gluten containing food. Commonly known as gluten intolerance, this disease has several other

names in literature, including cœliac disease, c(o)eliac sprue, non-tropical sprue, endemic sprue, gluten enteropathy or gluten-sensitive enteropathy. The gastrointestinal manifestations invariably comprise an inflammatory reaction within the mucosa of the small intestine caused by a dysregulated immune response triggered by ingested gluten proteins of certain cereals (wheat, rye, and barley), especially against gliadine. During the course of the disease, hyperplasia of the enteric crypts occurs and the mucosa eventually looses its absorptive villi thus leading to a diminished ability to absorb nutrients. The real prevalence of the disease has not been fully clarified yet. This is due to the fact that most patients with celiac disease suffer from no or atypical symptoms and only a minority develops the classical form of the disease. Since several years, prevalence data have been continuously adjusted upwards. Fasano et al. (2003) state that more than 2 million people in the United States, this is about one in 133, have the disease. People with untreated celiac disease, even if asymptomatic, are at risk for developing various complications like osteoporosis, infertility and other autoimmune diseases including type 1 diabetes, autoimmune thyroid disease and autoimmune liver disease.

Endoscopy with biopsy is currently considered the gold standard for the diagnosis of celiac disease. Besides standard upper endoscopy, several new endoscopic approaches for diagnosing celiac disease have been applied (Chand and Mihas, 2006). The modified immersion technique described in Cammarota et al. (2006) is based on the instillation of water into the duodenal lumen for better visualization of the villi. Furthermore, magnifying endoscopy (standard endoscopy with additional magnification) has been investigated (Cammarota et al., 2004). For conducting capsule endoscopy (see Petroniene et al. (2005)) the patient swallows a small capsule equipped with a camera that takes images of the duodenal mucosa during its passage through the intestine. All these techniques aim to detect total or partial villous atrophy and other specific markers that show a high specificity for celiac disease in patients. These markers include scalloping of the small bowel folds, reduction in the number or loss of Kerkring's folds, mosaic patterns and visualization of the underlying blood vessels (Niveloni et al., 1998). During endoscopy at least four duodenal biopsies are taken. Microscopic changes within these specimen are classified by a histological analysis according to a classification scheme by Oberhuber et al. (1999) which is based on Marsh (1992).

Automated classification as a support tool is an emerging option for endoscopic diagnosis and treatments (e.g. Karkanis (2003); Ameling et al. (2009);

Alexandre et al. (2008); Iakovidis et al. (2006); Liedlgruber and Uhl (2009)). Systems are being developed that support physicians during surgery or highlight malignant areas during endoscopy for further inspection. Such systems could also be used for training purposes. In the context of celiac disease, an automated system identifying duodenal areas affected by the disease would offer the following benefits (among other):

- Methods that help indicating specific areas for biopsy might improve the reliability of celiac disease diagnosis. As biopsying is invasive and the number of biopsy samples should be kept small, optimal targeting is desirable. This targeting can be supported by an automated system for identification of areas affected by celiac disease.

- The whole diagnostic work-up of celiac disease, including duodenoscopy with biopsies, is time-consuming and cost-intensive. To save costs, time, and manpower and simultaneously increase the safety of the procedure it would be desirable to develop a less invasive approach avoiding biopsies. Recent studies by Cammarota et al. (2006, 2007) investigating such endoscopic techniques report reliable results. These could be further improved by analysis of the acquired visual data (digital images and video sequences) with the assistance of computers.

- The (human) interpretation of the video material captured during capsule endoscopy (Petroniene et al., 2005) is an extremely time consuming process. Automated identification of suspicious areas in the video would significantly enhance the applicability and reduce the costs of this technique for the diagnosis of celiac disease.

In a prior study, Vécsei et al. (2008) suggest using histogram-based and Wavelet-based features for classification. Subsequent work (Vécsei et al., 2009) optimizes Fourier features used for classification by applying an evolutionary process already delivering competitive classification results. In recent work (Hegenbart et al., 2009), we have systematically compared the classification performance of two different image capturing techniques (i.e. conventional imaging vs. the modified immersion technique) and various pre-processing schemes using a set of different feature extraction and classification methods.

Ciaccio et al. (2010) measure the mean and standard deviation in brightness over 10×10 pixel subimages to identify areas affected by celiac disease

in capsule endoscopy, and also apply spectral analysis over sequential images to identify abnormal bowel motility.

Contributions In this work, we describe for the first time a system aimed at performing automated classification of duodenal texture patches according to a reduced 4-class Marsh-like classification system. Corresponding results are requested for a staging of the observed mucosa defects with impact on clinical practice regarding treatment. Local Binary Patterns (LBP) based feature extraction is applied to the problem of automated celiac disease diagnosis for the first time and turns out to outperform techniques previously applied. In particular, we propose two new operator types, one of which adapts to the different properties of Wavelet transform subbands and results in the best overall classification accuracy in the two-class scheme of all feature extraction schemes considered. In the four-class scheme the proposed method was still among the best methods. Moreover, we contribute in providing explicit strategies for threshold selection and quantization in operators proposed in earlier work.

Structure In section 2, we describe image acquisition and the establishment of ground truth information according to a modified Marsh classification. Section 3 covers LBP operators where we also propose two new operator types, among them a new Wavelet-based operator that combines two LBP-based operators to adapt to Wavelet subband properties. In section 4 we present experimental results where we compare the classification results of the proposed methods to techniques applied previously to classify endoscopic image material. Section 5 concludes the paper.

2. Image Acquisition and Marsh Classification

The image test set used, contains images taken during duodenoscopies at the St.Anna Children's Hospital using pediatric gastroscopes without magnification (GIF-Q165 and GIF-N180, Olympus, Hamburg). The main indications for endoscopy were the diagnostic evaluation of dyspeptic symptoms, positive celiac serology, anemia, malabsorption syndromes, inflammatory bowel disease, and gastrointestinal bleeding. Images were recorded by using the modified immersion technique, which is based on the instillation of water into the duodenal lumen for better visibility of the villi. The tip of the gastroscope is inserted into the water and images of interesting areas are taken. Gasbarrini et al. (2003) show that the visualization of villi with the immersion technique has a higher positive predictive value. Previous work by

Hegenbart et al. (2009) also found that the modified immersion technique is more suitable for automated classification purposes as compared to the classical image capturing technique. Images from a single patient were recorded during a single endoscopic session.

Our study population comprised only children suffering from signs and symptoms making upper endoscopy necessary. Therefore, the prevalence of celiac disease within this group was definitely higher than in the general population. Furthermore, there was a higher number of girls than boys (1.43:1) among the study group patients. Both findings, the higher prevalence of celiac disease and the female preponderance, should not bias the classification accuracy. Since endoscopy is an invasive procedure, a study like ours cannot be performed in a randomly selected sample from the general population due to ethical reasons since the medical indications for such an intervention are lacking. However, we consider our study group to be representative for the children needing endoscopic evaluation.

A fully automated system (as it is the final target of our project) would apply segmentation to decide which parts of an image are subject to feature extraction. However, as a first stage towards full automation we need to establish a database of image data, for which reliable texture classification can be developed and systematically optimized. For this purpose we have manually created a set of textured image patches with optimal quality to assess if the required classification is feasible under "idealistic" conditions and to establish reliable data. Thus, the captured data was inspected and filtered by several qualitative factors (sharpness, lack of distortions like specular reflections, visibility of features, etc.). To ensure the quality of extracted regions in terms of visibility of features the extraction was performed in accordance with a physician involved in this project.

There are two duodenal regions considered for extracting biopsy specimen. Those two regions (the duodenal Bulb and the Pars Descendens) have different geometrical properties (Hegenbart et al., 2009). There are no differences in the visual markers we use for classification among both regions however. In order to built an image database comprising enough images to be able to construct disjoint sets for training and evaluation of the specific classification methods, texture patches from both regions were combined. By restricting the images from the Pars Descendens to a frontal camera perspective (which make up the majority of images), inhomogeneities among the visual celiac markers are avoided.

	Characteristic Mucosal Changes
Marsh 0-2	No visible changes of villi structure
Marsh 3A	Mild villous atrophy
Marsh 3B	Marked villous atrophy
Marsh 3C	Absent villi

Table 1: Characteristic Changes of Mucosal Tissue caused by Celiac Disease.

In order to generate the ground truth for the texture patches used in experimentation, the condition of the mucosal areas covered by the images was determined by histological examination of biopsies from the corresponding regions. Severity of villous atrophy was classified according to the modified Marsh classification in Oberhuber et al. (1999). Two pathology residents and one senior pediatric pathologist examined the slides prepared from the submitted duodenal tissue samples. A final assessment of the grade of alteration of the mucosal architecture was performed by the supervising pathologist in every case. In cases of disagreement among the pathologists, only occurring in terms of subclassification of Marsh class 3 lesions, a final diagnosis was obtained from a consensual review of the slides on a multiheaded microscope.

This histological classification scheme identifies six classes of severity of celiac disease, ranging from class Marsh-0 (no visible change of villi structure) up to class Marsh 3C (absent villi). A visible change of the villous structure can be observed at Marsh 3A to Marsh 3C only.

We distinguish between Marsh classes Marsh-0 to Marsh-2 (not possible to diagnose mucosal damage via image analysis) and Marsh classes Marsh-3A to Marsh-3C. Therefore, images exhibiting underlying histological Marsh class Marsh-1 and Marsh-2 are not targeted by our system and were excluded from the analysis. In the following, we aim at two different classification problems: a four-class problem with classes Marsh-0, Marsh-3A, Marsh-3B, and Marsh-3C, and a two-class problem with the classes Marsh-0 and Marsh-3 (consisting of images of the latter three classes). Note that previous work has been entirely restricted to the two-class problem. Table 2 shows the number of texture patches and patients available per considered Marsh-class. As can be seen, for the two-class problem the number of images is well balanced, while for the four-class problem the Marsh-3 classes contain less images as compared with Marsh-0. Figures 1 shows examples for each considered class.

Please note that we manually enhanced the image contrast to improve the visibility of celiac markers for the reader.

(a) Marsh 0 (b) Marsh 3A (c) Marsh 3B (d) Marsh 3C

Figure 1: Celiac Images showing Examples of the considered Marsh Classes.

As can be seen, the visible differences between the specific Marsh-classes are rather small and can often be masked by either a bad image quality (blur or distortions) or a suboptimal perspective towards the mucosal plane. This could make accurate classification in the four-class case hard to achieve.

2.1. Image Database Construction

The constructed image database originates from 171 patients (131 control patients and 40 patients with diagnosed celiac disease). Texture patches with a fixed size of 128×128 pixels were extracted from the full sized frames (which are of size 768×576 pixels in case of the GIF-Q165 and 528×522 pixels in case of the GIF-N180 endoscope). In some cases multiple non-overlapping texture patches were extracted from a single full sized frame in order to build an image set of reasonable size. The patch size of 128×128 pixels turned out to be optimally suited in previous experiments (Hegenbart et al., 2009). The applied algorithms were not dependant on the specific endoscopic camera used.

In total 753 texture patches met the required qualitative properties. Based on this set of texture patches two distinct sets for training and evaluation were created. The construction was done in an automated way such that the number of images is balanced between the non-celiac class Marsh-0 and the celiac classes Marsh-3A to Marsh-3C. While creating the two distinct sets, care was taken that the number of patches per patient is as evenly balanced as possible. Also, no images from a single patient are within both image sets. The actual construction was done using a pseudo random number gen-

erator based on a Gaussian distribution to avoid any bias within the data sets. Table 2 shows the distribution of images and patients per class.

	0	**3A**	**3B**	**3C**	**Total**
Texture Patches					
Training Set	155	50	56	51	312
Evaluation Set	151	45	58	46	300
Patients					
Training Set	66	6	7	8	87
Evaluation Set	65	5	6	8	84

Table 2: Distribution of the Texture Patches and Patients in the Image Database.

3. Feature Extraction based on Local Binary Pattern Operators

The basic Local Binary Pattern (LBP) operator was introduced to the community by Ojala et al. (1996). This method belongs to the class of geometric parametrization algorithms. Šajn and Kononenko (2008) use multiresolution image parametrization for improving texture classification using association rules to extract a set of features. Malik et al. (1999) extended the Texton model to gray scale textures. Their method includes Gabor filtering and hence includes calculating the weighted mean of pixel values in a small neighborhood. The LBP operator considers each pixel in a neighborhood separately. Hence the LBP could be considered as a micro-texton. The operator is used to model a pixel neighborhood in terms of pixel intensity differences. This means that several common structures within a texture are represented by a binary label. The joint distributions of these binary labels are then used to characterize a texture. The operator is parametrized by a corresponding value for the used radius from the center (r) and the number of considered neighbors (p). The LBP operator is then defined as

$$LBP_{r,p}(x,y) = \sum_{k=0}^{p-1} 2^k \, s(I_k - I_c), \tag{1}$$

with I_k being the value of neighbor number k and I_c being the value of the corresponding center pixel. The neighbor pixels are positioned at equidistant positions on a circle around the center pixel with radius r using bilinear interpolation. The actual ordering of neighbor pixels is not relevant to the extracted information. The s function acts as sign function, mapping

to 1 if the difference is smaller or equal to 0 and mapping to 0 else. The LBP histogram with i intervals computed for an image I using p LBP neighbors is formally defined as

$$H_I(i) = \sum_{x,y} (LBP_{r,p}(x,y) = i) \qquad i = 0, \cdots, 2^p - 1. \tag{2}$$

The basic operator uses an eight-neighborhood with a 1-pixel radius. To overcome this limitation, the notion of scale is used as discussed by Ojala et al. (2002) by applying averaging filters to the image data before the operators are applied. Thus, information about neighboring pixels is implicitly encoded by the operator. The appropriate filter sizes for a certain scale is calculated as described by Mäenpää (2003).

To compute the distance (or similarity) of two different histograms we apply the histogram intersection metric. This metric is later interpreted as distance by a k-nearest neighbors (k-nn) classifier. For two histograms (H_1, H_2) with N intervals and interval number i being referenced to as $H(i)$, the similarity measure is defined as

$$H(H_1, H_2) = \sum_{i=1}^{N} \min(H_1(i), H_2(i)). \tag{3}$$

3.1. Extended Local Ternary Patterns with adaptive Threshold

As the LBP operator is sensitive to noise, the Local Ternary Pattern operator (LTP) was introduced by Tan and Triggs (2007). The modification is based on a thresholding mechanism which implicitly improves the robustness against noise. In our scenario endoscopic images are used which usually are noisy as a result of the endoscopic procedure. The bowel is illuminated by a point source located at the tip of the endoscope. The camera has a fixed focus, hence some areas that are either too close or too far away from the position of the camera are blurred. Additionally, the three dimensional nature of the bowel leads to uneven illumination leading to noisy regions within the captured images. The LTP operator is used to ensure that pixel regions that are influenced by these kind of distortions do not contribute to the computed histograms. The LTP approach is similar to the Peripheral Ternary Sign Correlation (PTESC) as used in Yokoi (2007). The PTESC operator however, was not used in the context of texture classification. The basic idea of LTP is to introduce a threshold for calculating the patterns:

$$s(x) = \begin{cases} 1, \text{ if } x \geq T_h \\ 0, \text{ if } |x| < T_h \\ -1, \text{ if } x \leq -T_h. \end{cases} \tag{4}$$

The ternary decision leads to two separate histograms, one representing the distribution of the patterns resulting in a -1, the other representing the distribution of the patterns resulting in a 1.

$$\begin{aligned} H_{I,lower}(i) &= \sum_{x,y}(LBP_{r,p}(x,y) = -i) \quad i = 0, \cdots, 2^p - 1 \\ H_{I,upper}(i) &= \sum_{x,y}(LBP_{r,p}(x,y) = i) \qquad i = 0, \cdots, 2^p - 1. \end{aligned} \tag{5}$$

The neighbor information of pixels that lie within the threshold is encoded implicitly by this splitting. A problem is that not the joint distribution of lower and upper patterns is considered but the marginal distributions. An alternative is to encode the patterns as trinary numbers. Nevertheless this approach creates rather huge and therefore sparse histograms (3^8-intervals instead of 2^8). This can result in instable results of the histogram similarity measures. All tests show inferior results of this trinary encoding, therefore the experiments were conducted using the concatenation of both histograms. The two computed histograms are concatenated and then treated like a single histogram.

The actual optimal values to use for thresholding are unknown a priori. Tan and Triggs (2007) use a fixed threshold that was found empirically and is beneficial for their input data. In case of endoscopic images however it is not safe to make assumptions about the average image quality. By applying an adaptive threshold based on the spatial image statistics we make sure that noisy regions do not contribute to the computed histograms while information present within high quality regions are not lost due to a threshold that was chosen too high. When calculating an adaptive threshold care has to be taken to avoid that visible texture-distortions (such as visible duodenal folds) affect the calculation of the threshold too heavily. The calculation is therefore based on an expected value for the standard deviation of the image (β). This value was found based on the specific training data used during experimentation and represents the average standard deviation of pixel intensity values within all texture patches in the training set. The value α is used as a weighting

factor combined with the actual pixel standard deviation of the considered image (σ) and is used to adapt the threshold to match the considered image characteristics. The value for α was found empirically in the context of this work and was set to 0.1.

$$T_h = \begin{cases} \beta^{\frac{1}{2}} + \alpha\sigma, & \text{if } \sigma > \beta^{\frac{1}{2}} \\ \beta^{\frac{1}{2}} - \alpha\sigma, & \text{if } \sigma \leq \beta^{\frac{1}{2}}. \end{cases} \tag{6}$$

Information extracted by the LBP-based operators from the intensity function of a digital image can only reflect first derivative information. This might not be optimal, therefore Huang et al. (2004) suggest using a gradient filtering before feature extraction. By doing this the velocity of local variation is described by the pixel neighborhoods. The naming conventions of this extension are not consistent within literature. We will therefore stick to the naming of Huang et al. (2004) (extended LBP, or ELBP). The extended LTP (ELTP) operator is consequently introduced in perfect analogy to the ELBP operator. ELTP is based on the LTP operator instead of the LBP operator to suppress unwanted noise in the gradient filtered data. Of course, the actual manner how to compute the gradient information has to be defined for a specific operator.

3.2. Local Binary Patterns with Contrast Measure and its Quantization

As the LBP operator is invariant in terms of monotonic grayscale changes, the strength of a pattern can not be represented. Texture however, can be seen as a combination of the spatial structures (patterns) and the strength of these structures (contrast). Therefore Ojala et al. (1996) introduce the LBP/C operator to combine both properties. The contrast and the local binary patterns supplement each other in a very useful way. The LBP are sensitive to rotational changes but invariant to monotonic grayscale variations where the contrast measure is rotation invariant but sensitive to grayscale changes. The rotation invariant local contrast measure for a pattern calculated at center (x, y) with a radius r considering p neighbors is calculated as

$$C_{r,p}(x, y) = \frac{1}{p}\left(\sum_{k=1}^{p}(I_k - \mu_{r,p}(x, y))^2\right), \tag{7}$$

with

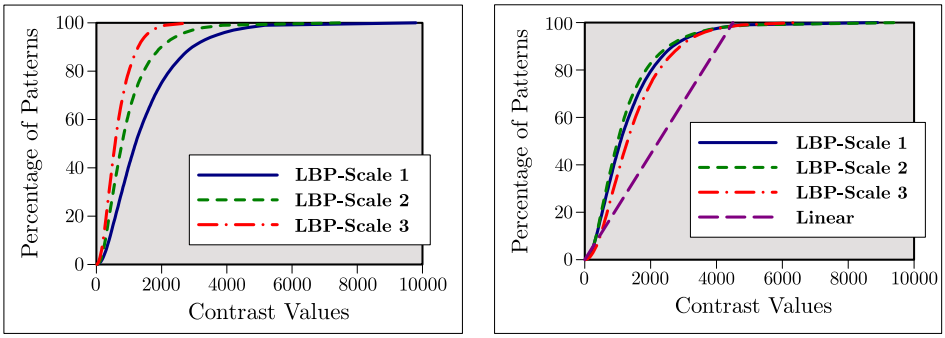
$$\mu_{r,p}(x, y) = \frac{1}{p}\left(\sum_{k=1}^{p} I_k\right). \tag{8}$$

$C_{r,p}$ is the variance within the support area of the operator (among all neighbors of a specific center pixel) and is interpreted as the strength of a pattern. The histogram is extended to two dimensions using the contrast measure as index in one dimension, modeling the joint distribution of both random variables. Usually the contrast values (c) are quantized to reduce the numbers of indices into the histogram. The best number of quantization intervals is unclear a priori. A small number leads to bad discrimination where a too large number leads to sparse histograms.

$$H_I(i,c) = \sum_{x,y} (LBP_{r,p}(x,y) = i \wedge C_{r,p}(x,y) = c) \qquad i = 0, \cdots, 2^p - 1 \quad (9)$$

The set of possible contrast values ranges from 0 to 16265.25 (the highest value results from a set of the neighboring pixels with half of the pixels having the minimum value (0) and half of the pixels with the maximum value (255)). Obviously it is highly unlikely to find a pixel neighborhood with these properties in a natural image. The distribution of contrast values is far from being uniform. Therefore a linear mapping of the contrast value to the corresponding interval index is inadequate as it would result in unevenly filled histograms. In this case a high percentage of patterns would be associated with only a few quantized contrast values and the discriminative power could not be improved. Ojala et al. do not suggest an explicit way how to quantize the contrast values however. Considering that the discriminative power in case of combined features is not determined by the number of patterns associated with a certain contrast range but determined by the actual patterns associated with a contrast value, we try to find a mapping that results in equally dense histograms. The mapping was found by estimating the empirical distribution function using the training data during each experiment. As the multiscale LBP-extension is used, the effects of low-pass filtering have to be considered. Obviously the distribution of the contrast values is affected by this filtering as shown in Figure 2 which displays the empirical cumulative distribution function that was found using the image data. The y-axis shows the percentage of patterns with a contrast value less or equal to the corresponding value on the x-axis. We therefore normalize the values by division, using the standard deviation of the contrast values. This is done for each image during feature extraction. The optimal number of contrast intervals was found empirically during the experiments by considering all values within a range from 2 to 22. The right side of the figure demonstrates the results

of the normalization of the contrast distribution and compares them with a linear distribution function.



(a) Deviation by Scale (b) Normalized

Figure 2: Cumulative Distributions of the Contrast Measures

3.3. The Wavelet based LBP Operator (WT-LBP)

All LBP-based operators can be categorized into two families by considering the underlying intensity function. Operators that reflect first derivative information (such as LBP, LBP/C and LTP) as well as operators that reflect second derivative information (ELBP and ELTP). The operators reflecting first derivative information are based on the unmodified intensity function of a digital image. The other operators are based on the first derivative of the underlying intensity function. This derivative describes the velocity of local variation. Therefore the extracted information reflects second derivative information. Taking into consideration that all LBP-based operators that were used successfully in the field of texture classification belong to one of the before mentioned categories, a combination of operators from either type seems promising. By using the Wavelet representation of the images a natural connection between both categories can be established.

In Wang et al. (2008) Haar Wavelets are used in combination with uniform LBP to improve the texture retrieval rate as compared to "pure" LBP. Liu and Ding (2009) use non-separable Wavelets with LBP to describe textures while Su et al. (2009) use Gabor-Wavelets in combination with LBP to represent texture in an active appearance model. The approaches of Su et al. and Liu et al. are based on the high frequency subbands while Wang et al. also use the approximation subbands for feature extraction. The subbands

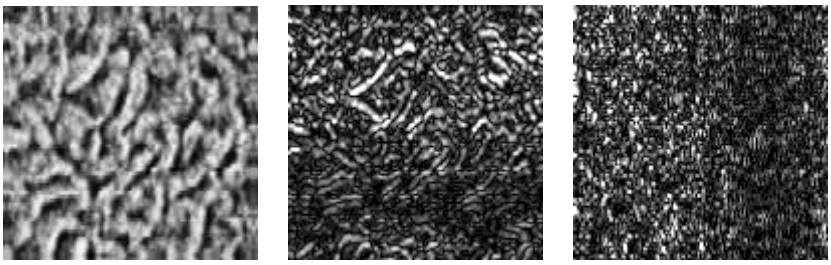
however have varying characteristics, therefore using a single operator (all referenced techniques use LBP) for extracting features from all subbands is not optimal. When considering the properties of the Wavelet transform, one can see that there is a natural relation to extensions suggested to the basic LBP operator:

- **Multiscale**
  The scaling function used within the Wavelet transform leads to a successive downscaling of the transformed signal. This corresponds to a decrease in resolution. When considering the LBP multiscale extension, pixel intensities are described as a weighted sum of the pixels within a neighborhood. As averaging filter are used for different scales, this corresponds to a decrease in resolution as well.

- **High Frequency Information**
  In Mallat's vertical and horizontal analysis (Mallat, 1989), the decomposition algorithm is based on two variables x and y leading to a priorization of each direction. The detail subbands contain high frequency information of the input signal. High frequency components in an image correspond to edge information. As the magnitude of each coefficient represents the strength of an edge we can interpret the detail subband coefficients as the speed of variation of pixel intensity differences. This is used within the operator based on using gradient filtering (ELBP and ELTP).

- **Supplemental Features**
  The coefficients of the detail subbands represent the information that is lost due to the downscaling of the approximation subband. Therefore the information present in the detail subbands complements the information present within the approximation subband in a natural way. Since both, high-frequency and low-frequency texture information have been promising in the context of classifying celiac disease in endoscopic images, we combine these features to improve the discriminative power. This in parallel to the LBP/C operator where supplemental features (the binary labels and the contrast values) are combined to improve the discriminative power.

As a consequence we propose a new Wavelet-based operator which is constructed by combining suitable variants of the basic LBP operator. The



(a) Approximation (b) Horizontal Details (c) Vertical Details

Figure 3: Coefficients of Wavelet Subbands.

properties of the specific operators and the Wavelet decomposition is taken into account when constructing this new WT-LBP operator. Both the approximation and detail subbands are used for feature extraction. By using all subbands, different components of textures can be described optimally.

- **Detail Subbands**
  The detail subbands contain high frequency components and are in a way similar to the information that is represented by gradient images. The set of Wavelet functions spans the differences between the spaces spanned by the various scales of the scaling function. In contrast to the ELBP and ELTP operators the detail subband coefficients contain the information that is lost due to the downscaling process of the Wavelet transform. By combining features from all subbands no important information is lost overall. This is in contrast to the Sobel filtering. Even more, the high frequency components can be used at different scales without losing information (although in our case only in a dyadic stepping). We are interested in the energy distribution of the coefficients, therefore the absolute values of the coefficients are used.

  Figure 3 shows the approximation subband as well as the absolute values of the coefficients of the horizontal and vertical detail subbands of a wavelet decomposed mucosa texture image. As can be seen, due to using a discrete signal, the detail coefficients contain some amount of noise. To avoid introducing this noise to the computed histograms the LTP operator is used to extract features from the detail subbands. Applying the LTP operator is similar to the quantization of coefficients. The LTP operator that is applied to the detail coefficients does not use the multiscale extension in order to avoid the low pass filtering of the

high frequency information since different scales are represented by the Wavelet transform coefficients anyway. The radius of the LTP that is used within the WT-LBP is set to 1.5 pixels. This is similar to a 3×3 pixel window, however since we use interpolation the actual values of the diagonal neighbors might be slightly different.

- **Approximation Subbands**
  The approximation subband represents the low frequency components of the image. By using dyadic sampling the bandwidth of the image is halved during each iteration. This is a problem, as we can not guarantee that the size of texture elements corresponds to this sampling. It is possible to miss texture components by applying the basic LBP operator to the approximation subband coefficients. Therefore the LBP multiscale extension is used to extract features from the approximation subband. As the LTP and LBP operator can not describe the strength of the patterns and the LBP/C operator proved to be very effective, the LBP/C operator is used to extract features from the approximation subbands. We use a maximum LBP-scale of 3 and a minimum LBP-scale of 1 since higher scales are obtained by the Wavelet decomposition anyway.

Figure 4 demonstrates the process of extracting features using two scales of the WT-LBP operator. The filter bank that is used in the experiments is the biorthogonal Cohen-Daubechies-Feauveau (CDF) 9/7 analysis filter also used within JPEG2000.

3.4. Operator Parameters

The performance of the LBP-based operators is determined by a significant set of parameters. The used neighborhood size of the operator controls how many neighboring samples are involved in building the pattern. A neighborhood size too small leads to poor discrimination while a neighborhood too large generates sparse histograms. Most authors suggest using an eight-neighborhood resulting in 256 patterns for the LBP operator. Mäenpää et al. (2000) suggest using only a subset of all possible patterns called the uniform patterns. This subset is characterized by the property that at maximum two transitions from 0 to 1 or vice versa are allowed within each pattern (58 patterns satisfy this condition). This constraint leads to a robust subset for classification. Additionally the dimensionality of the histogram is reduced

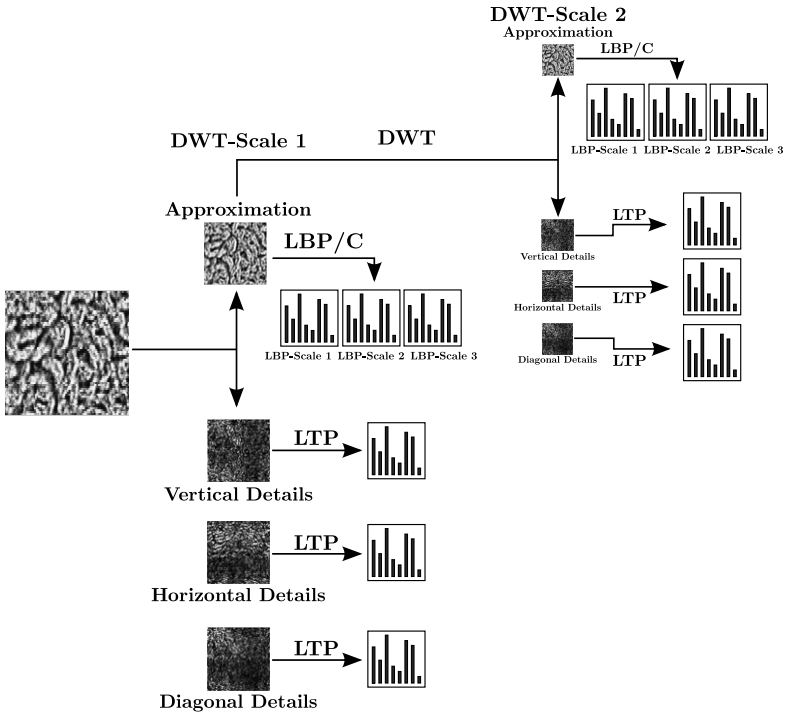


Figure 4: Two-Scale Wavelet Based LBP Operator (WT-LBP-Operator).

which is beneficial for our task. In all experiments the subset of uniform Local Binary Patterns is used for classification. In case of the LTP operator and those based upon the LTP operator, two histograms are concatenated to represent the pattern distributions. Therefore the size of the combined histogram is twice the size of the LBP based operator.

The LBP-scale parameter (with the meaning as in Mäenpää (2003)) of the operator describes how many pixels are actually involved for each neighboring sample. An increasing LBP-scale represents a lower image resolution and is used to describe large scale structural information that could otherwise not be represented. It is unclear a priori which LBP-scales are best suited to represent a given texture. Experiments show however, that features extracted using LBP-scales greater than 3 do not contribute useful information for classification. Therefore the extracted features are based on using a set of LBP-scales ranging from 1 to 3.

Huang et al. (2004) compute the gradient magnitudes to generate the ELBP histograms. In general however, mucosal images may have a dominant

orientation (this could be related to the physician's style however). Hence a filter orientation might be superior over the other. If one orientation is dominant within the image, the calculation of the gradient magnitude might introduce an error. Therefore both gradient images are directly used for computing the LBP histograms. We additionally use a so called diagonal orientation which represents the mean of both gradient orientations.

Obviously, not all filter orientations, Wavelet subbands, or LBP-scales are equally well suited for feature extraction. All combinations of these parameters are used to compute the histograms. During the classification process we optimize the best combination of histograms for each image set and classification problem by using a feature subset selection algorithm (SFS, Jain and Zongker (1997)). The absolute overall classification rate was used as criterion function for the optimization. Mäenpää et al. (2000) use feature subset selection methods to find an optimal subset of patterns for classification. A single histogram could however be interpreted as a single feature. In this work, the combination of used histograms was optimized but not the subset of patterns within histograms.

4. Experiments

To be able to assess the performance of the proposed extensions and the WT-LBP operator and to gain insight into the possible performance of the four-class modified Marsh classification scheme, we applied as a comparison a set of several different feature extraction methods that provided promising results in classification of endoscopic image data in earlier work. The abbreviations of the techniques used throughout this work are shown in bold. We used the following feature extraction methods (given in alphabetical order):

- **DT-CWT Correlation** Signatures: We have extended the Wavelet Correlation Signatures approach of de Wouwer et al. (1997) to work with the Dual-Tree Complex Wavelet Transform in Häfner et al. (2008). The correlation between subbands of different (and equal) color channels is computed based on the mean and standard deviation of coefficient magnitudes. The DT-CWT decomposition depth is set to four levels and the color space is RGB. The resulting features vectors have 144 elements.

- **DT-CWT-Weibull**: The Dual-Tree Complex Wavelet Transform is used to decompose the images into 6 scales and the empirical his-

togram of the detail subband coefficient magnitudes is modeled by two-parameter Weibull distributions. The Weibull parameters are then arranged into a feature vector (Kwitt and Uhl, 2007). In case of color images (which applies here) a feature vector has 216 elements.

- **ELBP**: Extended Local Binary Patterns (Huang et al., 2004) are used with an 8-neighborhood and LBP-scales ranging from 1 to 3. The image is gradient filtered by applying a Sobel filter using a horizontal, a vertical and a diagonal orientation. The optimal filter directions and LBP-scales are determined by using the SFS algorithm. The histogram dimensionality is 58.

- **ELTP**: The approach is applied as introduced in section 3.1. The ELTP operator is used with an 8-neighborhood and LBP-scales ranging from 1 to 3. The used α value was 0.1. In analogy to the ELBP operator, the filter orientations as well as color channels and LBP-scales are optimized using the SFS approach. The dimensionality of a each histogram is 116.

- **FFT-Evolved**: By using the FFT an image is transformed into the respective power spectrum. Multiple ring-shaped filters are then applied to the spectrum of each color channel of the RGB color model to concentrate on discriminative frequency subbands only. Since the number of possible ring filters is quite large, an evolutionary algorithm is used to find an optimal set of filters for each color channel (Vécsei et al., 2009) (sets are denoted by F_1, F_2, and F_3). For each of these ring filters the mean of the coefficient magnitudes within such a ring is used as a feature. This results in a feature vector for each color channel having a length equal to the number of rings used. By concatenating the feature vectors of all color channels of an image, the final feature vector is obtained having a length of $|F_1| + |F_2| + |F_3|$ (restricted to less then 20 elements).

- **Gabor, Classic**: The Gabor Wavelet Transform is used with 4 scales and 6 orientations, the mean and standard deviation of the coefficient magnitudes within a subband are used as features (Manjunath and Ma, 1996; Häfner et al., 2009c). The resulting feature vectors have 144 elements in case of color images.

- **LBP**: The Local Binary Pattern operator (Ojala et al., 1996) is used in an 8-neighborhood to compute histograms for each LBP-scale employed (in the range 1 - 3). The optimal combination of LBP-scales and color channels is found by the optimization routine (SFS) as described in section 3.4.

- **LBP/C**: The Local Binary Pattern operator combined with a contrast measure (Ojala et al., 1996) is used in an 8-neighborhood to compute histograms for each LBP-scale employed (in the range 1 - 3). The optimal combination of LBP-scales and color channels is found by the SFS algorithm. The optimal number of quantization intervals used for the contrast measure is optimized from 2 to 22 by an exhaustive search during each training. Let c_n be the number of used contrast values. The dimensionality of a single histogram is $58 \cdot c_n$.

- **LTP**: The Local Ternary Pattern operator (Tan and Triggs, 2007) is used in an 8-neighborhood to compute histograms for each LBP-scale employed (in the range 1 - 3). The adaptive thresholds are computed using an α value of 0.1. The optimal combination of LBP-scales, color channels and filter orientations is found by using the SFS algorithm. The dimensionality of a single histogram is 116.

- **WT-BBC**: The Best Basis Centroids method (Liedlgruber and Uhl, 2007) uses the Best-Basis algorithm to find an optimal basis for each image in a training set and computes a centroid over all resulting Wavelet packet decomposition structures (maximal decomposition depth 3). After transforming all images into this basis, the most informative subset of the resulting subbands (with respect to a cost function) is used to compute the energy over all coefficients within a subband. These values are concatenated to form the feature vector for an image. We use all color channels of the RGB color model and end up with a final feature vector length of $3 \cdot S$ with S being the number of selected subbands.

- **WT-LBP**: The approach is used as introduced in section 3.3 by applying a three stage dyadic Wavelet transform of the image data. The optimal combination of Wavelet-scales, color channels and LBP-scales is found by applying the SFS algorithm. In parallel to the LBP/C operator the quantized contrast values are found by an exhaustive search within the range of 2 to 22. The used α value was 0.1. For c_n contrast

values, the approximation subband based histograms have a dimensionality of $58 \cdot c_n$, while the detail subband based histograms have a dimensionality of 116.

- **WT-GMRF**: This method (Häfner et al., 2009a) first transforms an image to the Wavelet domain using the pyramidal discrete Wavelet transform (two stages) resulting in $3 \cdot 3 \cdot 2 = 18$ detail subbands since we use each color channel of the RGB color model. For each of these detail subbands the Markov parameters of a Gaussian Markov Random Field are estimated. The number of parameters resulting from one detail subband depends on the neighborhood order (neighborhoods used are of Geman type (Geman and Geman, 1984)). In addition to the Markov parameters we use the approximation error for each subband as a feature too. Hence, when assuming a neighborhood consisting of n neighbors, we have $\frac{n}{2} + 1$ features per subband (the neighborhoods are symmetric). Since we estimate these parameters for each subband in each color channel, the final feature vector length equals to $18(\frac{n}{2} + 1)$.

- **WT-LDB**: The Local Discriminant Basis algorithm is employed to find an optimal Wavelet packet decomposition basis (maximal number of stages is 3) with respect to discrimination between the image classes into which all images are transformed to. Based on the resulting decompositions we use the energy contained within a subband as feature, where only the S most discriminative subbands are used for feature extraction. The most discriminative subbands are found by computing the discriminative information for every respective subband following Saito and Coifman (1994). Since we use all color channels of the RGB color model we end up with feature vectors having a length of $3 \cdot S$ (Liedlgruber and Uhl, 2007).

In case of the methods FFT-Evolved, WT-BBC, WT-GMRF, and WT-LDB the images were always pre-processed by applying CLAHE (Zuiderveld, 1994) followed by a Laplace Sharpening with a kernel size of 9×9 (Gonzalez and Woods, 2002). For the other techniques no image pre-processing has been applied.

For classification we apply a k-nearest neighbors (k-nn) classifier to the extracted features. In the classifier, all methods except for the LBP-based ones use the Euclidean distance metric for the k-nn classification. The LBP-based methods use the histogram intersection as distance metric. While each

of the employed techniques has been published with a certain specific classifier (often leading to better results compared to k-nn classification), we want to give more emphasis to the features used by applying a common classifier. The optimal k-value was determined by an exhaustive search through the admissible corresponding parameter range. Based on previous experiments with the different techniques, the parameter range is specified as follows. For all methods k (the number of neighbors considered) is chosen from 1 to 15 except for the FFT-Evolved Method. The results of the FFT-Evolved methods are optimized by an evolutionary process, which either assigns k=1 or k=2 depending on the used chromosomes. On a tied decision among classes the a priori probability (class frequencies) is used for the final classification decision.

To evaluate how well the methods and estimated parameters perform on an independent dataset we constructed two disjoint sets of texture patches as explained in section 2. Parameter and feature optimization (including the k-value of the k-nn classifier) was based on using a leave-one-out cross validation (LOOCV, (Fukunaga, 1990)) on the training set. The evaluation of the methods accuracy was then performed by applying the trained classifiers to the evaluation set. No prior knowledge was used concerning the classification of the evaluation set.

To improve the results obtained by the k-nearest neighbor classifier, we use an Ensemble classifier as described in Häfner et al. (2009b); Vécsei et al. (2009). This classifier aims at achieving a higher overall classification accuracy and more stable results across different image classes by combining different methods. The performance of the Ensemble classifier is dependant on single methods with high accuracy and a high measure of diversity among each other. Therefore the selection algorithm starts by selecting the method with the highest accuracy based on a cross validation on the training data. Then the best method in terms of classification accuracy with a significant different outcome to the previously picked method (at a significance level of 5%) is selected. This process is repeated until no more methods are found. The optimization of the k-value as well as the reliability measure used by the Ensemble classifier was entirely based on the training set of images (denoted as Ensemble^1 and Ensemble^3). We additionally combined a set of single classifiers by using knowledge of how well these methods generalize based on the performed experiments on the evaluation set. These results however have to be considered with care as the manual combination of methods prevents a fair comparison to the other methods and are only used to assess how much

room for improvement exists for the Ensemble of classifiers. We denote these Ensembles as Ensemble^2 and Ensemble^4 in the corresponding tables.

5. Results

In this section we present the results of the conducted experiments. We present two result tables for each classification task (i.e. two-class and four-class). One result table displays the classification results estimated by a leave-one-out cross validation performed on the training set. The second table presents the results of classifying the evaluation set based on the previously optimized parameters and trained classifiers. Authors in a related field might not be in the position to use distinct datasets to evaluate presented methods due to a limited amount of available data. We therefore study both evaluation methods to be able to give a comprehensive view of how well certain methods generalize on an independent dataset and of how significantly methods tend to (over)-fit the extracted features and parameters towards the data.

Within the result tables we use the abbreviations "Spec." for specificity (the percentage of correctly classified images actually showing a normal mucosal state) and "Sens." to indicate the methods sensitivity (the percentage of correctly classified images showing villous atrophy). To improve the readability the results are rounded to one decimal position in the discussion. In case of the four-class scheme we use the abbreviations 0, 3A, 3B or 3C to indicate the specific Marsh class.

We display the best overall classification results among all LBP-based methods as well as the other methods (except for the Ensemble classifiers) in bold face. In case of one or more methods with the same classification accuracy we display the method with the highest sensitivity in bold face. The "k"-column indicates the number of neighbors that was used for the nearest neighbor classification. The column labeled as "Int." indicates the number of intervals used for the contrast values in case of the two dimensional LBP/C based histograms. We present the two ensembles of single methods with a corresponding superscript to unambiguously identify the specific ensembles.

In addition to the result tables we also show the results of the statistical tests for significance we performed. The check sign indicates that a statistical significant difference between two results according to McNemar's test (McNemar, 1947) was found. The value of α corresponds to the significance level of the specific test. McNemar's test considers the classification agreement

	Classification Rates				
	Spec.	**Sens.**	**Overall**	**k**	**Int.**
LBP	94.19	93.63	93.91	3	-
LTP	94.19	94.90	94.55	5	-
LBP/C	97.42	92.99	95.19	3	6
ELBP	93.55	94.27	93.91	15	-
ELTP	93.55	94.27	93.91	10	-
WT-LBP	98.06	93.63	**95.83**	5	20
DT-CWT-Corr.	90.97	92.40	**91.67**	3	-
DT-CWT-Weibull	92.26	88.54	90.38	4	-
FFT-Evolved	95.48	87.90	91.67	2	-
Gabor-Classic	89.03	91.08	90.06	5	-
WT-BBC	90.97	89.81	90.38	5	-
WT-GMRF	87.74	89.81	88.78	5	-
WT-LDB	89.68	89.81	89.74	7	-
Ensemble[1]	98.70	92.99	95.83	-	-
Ensemble[2]	98.07	94.90	96.47	-	-

Table 3: Classification Result of a Leave-One-Out Cross Validation on the Training Set (Two-Class Case).

between two classification results. The null hypothesis of marginal homogeneity states that the marginal outcomes of two considered experiments are the same. This means, considering two experiments, that the probabilities of experiment one being correct for an image while experiment two being incorrect and vice versa (experiment two being correct while experiment one being incorrect for that same image) are equal. If McNemar's test statistic is significant (the significance level used in McNemar's test is used to evaluate whether the test statistic is likely in terms of a chi-squared distribution) there is evidence to reject the null hypothesis. This implies that the difference between two classification results are considered to be statistically significant. At a significance level of 2.5 percent ($\alpha = 0.025$) there is a confidence level of 97.5 percent that the differences between two classification results were not caused by random variation.

5.1. Results of the Two-Class Scheme for Classification

Tables 3 and 4 present the results of the experiments based on the two-class scheme for classification. Comparing the results using a leave-one-out cross validation and the classification of the evaluation set, we see that the classification accuracy drops by an average of 8.6 percentage points in case of the LBP-based methods as well as the non-LBP-based methods. This

	Classification Rates				
	Spec.	**Sens.**	**Overall**	**k**	**Int.**
LBP	79.47	87.25	83.33	3	-
LTP	75.50	93.96	84.66	5	-
LBP/C	82.12	92.62	87.33	3	6
ELBP	80.13	92.62	86.33	15	-
ELTP	79.47	92.62	86.00	10	-
WT-LBP	85.43	90.60	**88.00**	5	20
DT-CWT-Corr.	83.44	81.21	82.33	3	-
DT-CWT-Weibull	87.42	76.51	82.00	4	-
FFT-Evolved	83.44	81.21	82.33	2	-
Gabor-Classic	80.13	80.54	80.33	5	-
WT-BBC	80.13	85.23	82.67	5	-
WT-GMRF	75.50	84.56	80.00	5	-
WT-LDB	78.81	86.58	**82.67**	7	-
Ensemble[1]	85.43	91.95	88.67	-	-
Ensemble[2]	85.43	90.60	88.00	-	-

Table 4: Classification Results of the Trained Methods on the Evaluation Set (Two-Class Case).

is interesting as the LBP-based methods all use feature subset selection as compared to the other methods were only FFT-Evolved applies an additional process of feature optimization. This indicates that the selected feature subsets generalize well on an independent dataset. The decrease in classification rate is assumed to be caused by a bias within the training data caused by possibly multiple texture patches of a single patient in combination with the leave-one-out cross validation. In general the LBP-based methods performed better on the evaluation set (85.9%) as compared to the non-LBP-based methods (81.8%). The better overall accuracy of the LBP-based methods is explained by a higher average sensitivity of approximately 9.3 percentage points. The best result of a single method based on the evaluation set is achieved by the WT-LBP operator with 88.0 percentage points overall accuracy. Compared to the classification accuracy of the LOOCV this method's accuracy drops by 7.8 percentage points which is below the average decrease. By using the Ensemble classifier the result could be slightly improved to 88.7 percentage points. Interestingly the manually combined ensemble could not further improve the classification rates.

Table 5 displays the outcomes of the conducted statistical significance tests based on the classification results of the evaluation set. We see that there is no statistically significant difference between the WT-LBP operator

and the other LBP-based operators at a significance level of 0.025. Considering a significance level of 0.05 there is a significant difference between the results of the WT-LBP and the basic LBP operator.

	$\alpha = 0.025$			$\alpha = 0.05$		
	WT-LBP	**Ens.**[1]	**Ens.**[2]	**WT-LBP**	**Ens.**[1]	**Ens.**[2]
LBP	-	✓	-	✓	✓	✓
LTP	-	-	-	-	-	-
LBP/C	-	-	-	-	-	-
ELBP	-	-	-	-	-	-
ELTP	-	-	-	-	-	-
WT-LBP	-	-	-	-	-	-
DT-CWT-Corr.	✓	✓	✓	✓	✓	✓
DT-CWT-Weibull	✓	✓	✓	✓	✓	✓
FFT-Evolved	✓	✓	✓	✓	✓	✓
Gabor-Classic	✓	✓	✓	✓	✓	✓
WT-BBC	✓	✓	✓	✓	✓	✓
WT-GMRF	✓	✓	✓	✓	✓	✓
WT-LDB	✓	✓	✓	✓	✓	✓
Ensemble[1]	-	-	-	-	-	-
Ensemble[2]	-	-	-	-	-	-

Table 5: Results of McNemar's Test for Significance among the Results of the Trained Methods on the Evaluation Set for the Two-Class Case.

Compared to the non-LBP-based methods the differences are all statistically significant. As a consequence of the single methods selected, there are no statistically significant differences between the Ensemble classifiers[1] [2] and the LBP-based methods except for the basic LBP-operator. Statistically significant differences to the non-LBP-based methods can be seen at both significance levels. The standard deviation of the LBP/C method among all evaluated interval numbers (2 to 22) during the training of was 1.2 percentage points with a mean classification accuracy of 86.6 percent. The mean accuracy and standard deviation of the WT-LBP method was 87.3 percentage points and 1.6 percentage points respectively.

5.2. Results based on the Four-Class Scheme for Classification

Tables 6 and 7 present the results of the classification based on the four-class scheme for classification. By analogy to the two-class scheme for classi-

[1]Ensemble[1] combines DT-CWT-Weibull, FFT, LBP/C, LTP, WT-BBC and WT-LBP
[2]Ensemble[2] combines DT-CWT-Weibull, LTP, WT-LBP and WT-LDB

fication we can see a decrease of classification accuracy when using a distinct set for evaluation. However, in the four-class case the decrease is significantly higher with an average of 19.4 percentage points in case of methods based on LBP and 16.1 percentage points for the non-LBP-based methods.

	Classification Rates						
	0	**3A**	**3B**	**3C**	**Overall**	**k**	**Int.**
LBP	96.77	68.00	62.50	50.98	78.53	8	-
LTP	95.48	72.00	60.71	60.78	79.81	1	-
LBP/C	96.13	74.00	83.93	45.10	82.05	4	15
ELBP	94.84	56.00	71.43	54.90	77.88	7	-
ELTP	96.77	62.00	71.43	50.98	79.16	11	-
WT-LBP	97.42	76.00	78.57	50.98	**83.01**	4	12
DT-CWT-Corr.	95.48	64.00	67.86	56.86	**79.17**	4	-
DT-CWT-Weibull	92.26	60.00	66.07	41.00	74.04	7	-
FFT-Evolved	83.23	72.00	73.21	56.86	75.32	1	-
Gabor-Classic	92.26	74.00	67.86	41.18	76.60	8	-
WT-BBC	92.26	48.00	67.86	43.14	72.76	5	-
WT-GMRF	93.55	58.00	66.07	35.29	73.40	5	-
WT-LDB	88.39	64.00	67.86	43.14	73.40	4	-
Ensemble[3]	98.70	78.00	89.29	25.49	81.77	-	-
Ensemble[4]	98.71	76.00	78.57	78.57	83.01	-	-

Table 6: Classification Result of a Leave-One-Out Cross Validation on the Training Set (Four-Class Case).

This indicates that the features selected by the histogram subset selection algorithm slightly over-fits the model towards the data. On average, the classification rates of the LBP-based methods are 60.6 percent compared to 58.9 percent achieved by the non-LBP-based methods. The low classification accuracy is explained by the classification rates of the Marsh type 3 subclasses. Marsh-3C has the lowest average classification accuracy with a mean below 30 percentage points for all methods. The best result was achieved by the basic LBP operator with 66.3 percent (a drop in overall accuracy of only 12.2 percentage points). The WT-LBP operator achieves a result of 63.7 percent. The best non-LBP based method is DT-CWT-Weibull also with 63.7 percent. It is interesting that the automatically selected Ensemble[3] of classifiers could not improve the classification accuracy and reaches only 62.3 percent. This result can be explained by the single methods used for the Ensemble:

[3]Ensemble[3] combines Gabor-Classic, LBP/C and WT-LBP

WT-LBP, LBP/C and Gabor-Classic. The algorithm selected these methods because they performed well on the training set using LOOCV and had statistically significant different results. However in case of the classification using the evaluation set, LBP/C dropped by 24.4 percentage points. Also the best performing method (LBP) was not selected because the performance in the leave-one-out cross validation of the training set was below average. In contrast to this, the manually selected Ensemble[4] improved the classification accuracy to an overall of 68.0 percent. Although the manual selection is unfair to some degree by using prior information of how well certain methods generalize, we see that there is still room for improvement. The standard deviation of the LBP/C method among all evaluated interval numbers (2 to 22) during the training was 1.0 percentage points with a mean classification accuracy of 80.9 percent. The mean accuracy and standard deviation of the WT-LBP method was 81.9 percent and 0.8 percentage points respectively. Considering table 8 we see that only few statistical significantly different results were produced by the WT-LBP and the Ensemble classifiers. It is interesting that the WT-LBP was statistical significantly different to two of the other Wavelet-based methods as well as LTP and LBP/C. This is interesting as these two methods (LTP and LBP/C) are incorporated in the WT-LBP method.

5.3. Result Discussion and Interpretation

A general remark is that with respect to the absolute values of classification accuracy it should be noted that the results shown are obtained with a k-nn classifier. Previous experiments with the employed feature extraction techniques have shown that these results can be further improved by employing SVM or Bayes classifiers (Hegenbart et al., 2009; Vécsei et al., 2009).

By using a distinct set of texture patches for evaluation of trained methods we avoid the problem of over-fitting the parameters towards the given data. We saw that in the four-class case some amount of over-fitting happened when using leave-one-out cross validation in combination with parameters and feature optimization. We also saw that care has to be taken when interpreting results of a cross validation as the constructed image data might be biased because of multiple texture patches extracted for a single patient. We

[4]Ensemble[4] combines DT-CWT-Weibull, ELTP, LBP, WT-LBP and WT-BBC

	Classification Rates						
	0	**3A**	**3B**	**3C**	**Overall**	**k**	**Int.**
LBP	90.07	48.88	46.55	30.43	**66.33**	8	-
LTP	78.15	22.22	22.41	32.61	52.00	1	-
LBP/C	86.09	20.00	31.03	34.78	57.66	4	15
ELBP	85.43	35.55	44.83	17.39	59.66	7	-
ELTP	88.74	46.66	48.28	21.74	64.33	11	-
WT-LBP	87.41	24.44	51.72	39.13	63.66	4	12
DT-CWT-Corr.	86.09	46.67	27.59	17.39	58.33	4	-
DT-CWT-Weibull	88.08	35.56	48.28	30.43	**63.66**	7	-
FFT-Evolved	70.20	33.33	46.55	30.43	54.00	1	-
Gabor-Classic	87.42	31.11	53.45	26.09	63.00	4	-
WT-BBC	84.77	60.00	32.76	19.57	61.00	8	-
WT-GMRF	82.78	46.67	29.31	17.39	57.00	5	-
WT-LDB	80.79	46.67	17.24	26.09	55.00	4	-
Ensemble[3]	96.02	20.00	53.45	4.35	62.33	-	-
Ensemble[4]	94.04	51.11	53.45	53.45	68.00	-	-

Table 7: Classification Results of the Trained Methods on the Evaluation Set (Four-Class Case).

	$\alpha = 0.025$			$\alpha = 0.05$		
	WT-LBP	**Ens.**[3]	**Ens.**[4]	**WT-LBP**	**Ens.**[3]	**Ens.**[4]
LBP	-	-	-	-	-	-
LTP	✓	✓	✓	✓	✓	✓
LBP/C	-	-	✓	✓	-	✓
ELBP	-	-	✓	-	-	✓
ELTP	-	-	-	-	-	-
WT-LBP	-	-	-	-	-	-
DT-CWT-Corr.	-	-	✓	-	-	✓
DT-CWT-Weibull	-	-	-	-	-	-
FFT-Evolved	✓	✓	✓	✓	✓	✓
Gabor-Classic	-	-	-	-	-	-
WT-BBC	-	-	✓	-	-	✓
WT-GMRF	-	-	✓	✓	-	✓
WT-LDB	✓	✓	✓	✓	✓	✓
Ensemble[3]	-	-	✓	-	-	✓
Ensemble[4]	-	✓	-	-	✓	-

Table 8: Results of McNemar's Test for Significance among the Results of the Trained Methods on the Evaluation Set for the Four-Class Case.

suggest, if possible, to use a modification to the leave-one-out cross validation known as leave-one-patient out (LOPO) cross validation. It is possible

that the selected feature subsets and optimized parameters are suboptimal for classification, caused by a biased texture patch set due the leave-one-out cross validation. We expect that by using leave-one-patient out cross validation for feature optimization more stable features for classification could be found.

We can make the following observations. The proposed ELTP operator does improve the results of the LTP operator and the ELBP operator in case of the evaluation set in the four-class scheme. The results of those two methods are comparable in the two-class scheme. The proposed Wavelet-based WT-LBP operator delivered the best overall classification results in the two-class case and was among the best methods in the four-class case. Obviously, the combination of first derivate- and second derivative based information in this operator is a successful strategy. We also observe, that LBP-based operators outperform non-LBP-based feature extraction techniques in terms of obtained top and average results. This indicates that indeed LBP-based schemes are very well suited for the classification of our datasets.

Considering the results of McNemar's test we saw that the agreement among the LBP-based methods was not significantly different in a statistical meaning. However, as this test only considers the homogeneity of marginal frequencies of two classification results, a negative test result does not necessarily mean, that a method reaching a higher accuracy is not superior to a method with a lower accuracy.

6. Conclusion

We have found statistically significant differences in classification accuracy among different settings, especially between the two and four-classes case. The performance of the used methods builds a solid basis for future work in case of the two-class scheme for classification. In case of the four-classes case however we saw that the used features fail to discriminate between the Marsh-3 subtypes. Overall classification rates in the range of 60 to 65 percent requires more effort to justify a clinical deployment. We saw that using information about how well certain methods generalize an improved ensemble yielding robust features that improved the classification rates in the four-class case could be found. Although comparing this result to the result of the other methods lacks fairness to some degree, it indicates that there is room for further improvement. Ensari (2010) states that the Marsh classification, as modified by Oberhuber et al. (1999), might lead to

increased intraobserver and interobserver variations. Ensari suggests to use a new classification scheme based on Corazza and Villanacci (2005) using only 3 classes by combining Marsh type 3A and 3B. By using this scheme, automated classification might be improved. Also more advanced techniques using feature subset construction such as suggested by Šajn and Kukar (2010) in combination with a more realistic leave-one-patient-out cross validation to increase feature reliability should be considered towards the improvement of classification accuracy. Considering the discriminative power visible features among the Marsh type-3 subclasses, advanced techniques used in endoscopy such as narrow band imaging (NBI, Gross and Wallace (2006)) could possibly be beneficial to automated classification accuracy. For the two-class problem (distinguishing areas affected by celiac disease and unaffected areas) the obtained classification accuracy builds a solid basis for future work towards employment in a clinical study.

The results show that the LBP-operator family exhibited better result accuracy compared to a wide range of other feature extraction techniques. The proposed Wavelet-based operator (WT-LBP), combining the LTP operator using an adaptive threshold and the LBP/C operator using an empirical distribution function for quantization of the contrast values, was among the best operators in all experiments. We saw that combining the first derivative- and second derivative information based operators using the Wavelet transform is beneficial to the feature discrimination and is able to improve the classification results.

Acknowledgements

This work has been supported by the Austrian National Bank "Jubiläumsfonds", project no. 12991. We acknowledge the help of Georg Wimmer in computing some feature vectors and significance data.

7. Bibliography

Alexandre, L., Nobre, N., Casteleiro, J., May 2008. Color and position versus texture features for endoscopic polyp detection. In: Proceedings of the International Conference on BioMedical Engineering and Informatics, 2008 (BMEI'08). Vol. 2. Sanya, Hainan, China, pp. 38–42.

Ameling, S., Wirth, S., Paulus, D., Lacey, G., Vilarino, F., June 2009. Texture-based polyp detection in colonoscopy. In: Bildverarbeitung für die Medizin 2009. No. 15 in Informatik aktuell. Springer Berlin, pp. 346–350.

Cammarota, G., Cesaro, P., Martino, A., et al., January 2006. High accuracy and cost-effectiveness of a biopsy-avoiding endoscopic approach in diagnosing coeliac disease. Alimentary Pharmacology and Therapeutics 23 (1), 61–69.

Cammarota, G., Cuoco, L., Cesaro, P., et al., January 2007. A highly accurate method for monitoring histological recovery in patients with celiac disease on a gluten-free diet using an endoscopic approach that avoids the need for biopsy: a double-center study. Endoscopy 2007 39 (1), 46–51.

Cammarota, G., Martino, A., Pirozzi, G., 2004. Direct visualization of intestinal villi by high-resolution magnifying upper endoscopy: a validation study. Gastrointestinal Endoscopy 60 (5), 732–738.

Chand, N., Mihas, A. A., January 2006. Celiac disease: Current concepts in diagnosis and treatment. Journal of Clinical Gastroenterology 40 (1), 3–14.

Ciaccio, E. J., Tennyson, C. A., Lewis, S. K., Krishnareddy, S., Bhagat, G., Green, P. H., 2010. Distinguishing patients with celiac disease by quantitative analysis of videocapsule endoscopy images. Computer Methods and Programs in Biomedicine 100, 39–48.

Corazza, G. R., Villanacci, V., 2005. Coeliac disease. Journal of Clinical Pathology 58 (6), 573–574.

de Wouwer, G. V., Livens, S., Scheunders, P., Dyck, D. V., 1997. Color Texture Classification by Wavelet Energy Correlation Signatures. In: Proceedings of the 9th International Conference on Image Analysis and Processing (ICIAP'97). Springer, Florence, Italy, pp. 327–334.

Ensari, A., 2010. Gluten-sensitive enteropathy (celiac disease): Controversies in diagnosis and classification. Archives of Pathology and Laboratory Medicine 134 (6), 826–836.

Fasano, A., Berti, I., Gerarduzzi, T., Not, T., Colletti, R. B., Drago, S., Elitsur, Y., Green, P. H. R., Guandalini, S., Hill, I. D., Pietzak, M., Ventura,

A., Thorpe, M., Kryszak, D., Fornaroli, F., Wasserman, S. S., Murray, J. A., Horvath, K., February 2003. Prevalence of celiac disease in at-risk and not-at-risk groups in the united states: a large multicenter study. Archives of internal medicine 163, 286–92.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition, 2nd Edition. Morgan Kaufmann.

Gasbarrini, A., Ojetti, V., Cuoco, L., Cammarota, G., Migneco, A., Armuzzi, A., Pola, P., Gasbarrini, G., mar 2003. Lack of endoscopic visualization of intestinal villi with the immersion technique in overt atrophic celiac disease. Gastrointestinal endoscopy 57, 348–351.

Geman, S., Geman, D., 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721–741.

Gonzalez, R., Woods, R., 2002. Digital Image Processing – Second Edition. Prentice-Hall.

Gross, S. A., Wallace, M. B., December 2006. Hold on Picasso, narrow band imaging is here. American Journal of Gastroenterology 101 (12), 2717–2718.

Häfner, M., Gangl, A., Liedlgruber, M., Uhl, A., Vécsei, A., Wrba, F., 2009a. Combining Gaussian Markov random fields with the discrete wavelet transform for endoscopic image classification. In: Proceedings of the 17th International Conference on Digital Signal Processing (DSP'09). Santorini, Greece, pp. 177–182.

Häfner, M., Gangl, A., Liedlgruber, M., Uhl, A., Vécsei, A., Wrba, F., 2009b. Pit pattern classification using multichannel features and multiclassification. In: T.P. Exarchos, A. Papadopoulos, D. F. (Ed.), Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications. IGI Global, Hershey, PA, USA, pp. 335–350.

Häfner, M., Kwitt, R., Uhl, A., Gangl, A., Wrba, F., Vécsei, A., Sep. 2008. Computer-assisted pit-pattern classification in different wavelet domains for supporting dignity assessment of colonic polyps. Pattern Recognition 42 (6), 1180–1191.

Häfner, M., Kwitt, R., Uhl, A., Gangl, A., Wrba, F., Vécsei, A., Dec. 2009c. Feature-extraction from multi-directional multi-resolution image transformations for the classification of zoom-endoscopy images. Pattern Analysis and Applications 12 (4), 407–413.

Hegenbart, S., Kwitt, R., Liedlgruber, M., Uhl, A., Vécsei, A., Sep. 2009. Impact of duodenal image capturing techniques and duodenal regions on the performance of automated diagnosis of celiac disease. In: Proceedings of the 6th International Symposium on Image and Signal Processing and Analysis (ISPA '09). Salzburg, Austria, pp. 718–723.

Huang, X., Li, S., Wang, Y., 2004. Shape localization based on statistical method using extended local binary pattern. In: Proceedings of the 3rd International Conference on Image and Graphics (ICIG'04). Hong Kong, China, pp. 1–4.

Iakovidis, D. K., Maroulis, D. E., Karkanis, S. A., October 2006. An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy. Computers in Biology and Medicine 36 (10), 1084–1103.

Jain, A., Zongker, D., 1997. Feature selection: Evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 153–158.

Karkanis, S., Sep. 2003. Computer-aided tumor detection in endoscopic video using color wavelet features. IEEE Transactions on Information Technology in Biomedicine 7 (3), 141–152.

Kwitt, R., Uhl, A., 2007. 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 '07). Rio de Janeiro, Brasil, pp. 1–8.

Liedlgruber, M., Uhl, A., Oct. 2007. Statistical and structural wavelet packet features for pit pattern classification in zoom-endoscopic colon images. In: Dondon, P., Mladenov, V., Impedovo, S., Cepisca, S. (Eds.), Proceedings of the 7th WSEAS International Conference on Wavelet Analysis & Multirate Systems (WAMUS'07). Arcachon, France, pp. 147–152.

Liedlgruber, M., Uhl, A., Sep. 2009. Endoscopic image processing - an overview. In: Proceedings of the 6th International Symposium on Image and Signal Processing and Analysis, ISPA '09. Salzburg, Austria, pp. 707–712.

Liu, P., Ding, Z., May 2009. A blind image watermarking scheme based on wavelet tree quantization. In: Proceedings of the 2009 Second International Symposium on Electronic Commerce and Security, ISECS '09. Nanchang, China, pp. 218–222.

Mäenpää, T., 2003. The local binary pattern approach to texture analysis - extensions and applications. Ph.D. thesis, University of Oulu.

Mäenpää, T., Ojala, T., Pietikäinen, M., Soriano, M., 2000. Robust texture classification by subsets of local binary patterns. Pattern Recognition, International Conference on 3, 3947.

Malik, J., Belongie, S., Shi, J., Leung, T., 1999. Textons, contours and regions: Cue integration in image segmentation. In: ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2. IEEE Computer Society, Washington, DC, USA, p. 918.

Mallat, S., Jul. 1989. A theory for multiresolution signal decomposition: The wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (7), 674–693.

Manjunath, B. S., Ma, W. Y., Aug. 1996. Texture features for browsing and retrieval of image data. IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (8), 837–842.

Marsh, M., 1992. Gluten, major histocompatibility complex, and the small intestine. a molecular and immunobiologic approach to the spectrum of gluten sensitivity ('celiac sprue'). Gastroenterology 102 (1), 330–354.

McNemar, Q., June 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12 (2), 153–157.

Niveloni, S., Florini, A., Dezi, R., et al., March 1998. Usefulness of videoduodenoscopy and vital dye staining as indicators of mucosal atrophy of

celiac disease: assessment of interobserver agreement. Gastrointestinal Endoscopy 47 (3), 223–229.

Oberhuber, G., Granditsch, G., Vogelsang, H., November 1999. The histopathology of coeliac disease: time for a standardized report scheme for pathologists. European Journal of Gastroenterology and Hepatology 11, 11851194.

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

Ojala, T., Pietikäinen, M., Mäenpää, T., July 2002. Multiresolution Gray-Scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (7), 971–987.

Petroniene, R., Dubcenco, E., Baker, J., March 2005. Given capsule endoscopy in celiac disease: evaluation of diagnostic accuracy and interobserver agreement. The American Journal of Gastroenterology 100 (3), 685–694.

Saito, N., Coifman, R., Jul. 1994. Local discriminant bases. In: Laine, A., Unser, M. (Eds.), Wavelet Applications in Signal and Image Processing II. Vol. 2303 of SPIE Proceedings. San Diego, CA, pp. 2–14.

Su, Y., Tao, D., Li, X., Gao, X., 2009. Texture representation in aam using gabor wavelet and local binary patterns. In: SMC'09: Proceedings of the 2009 IEEE international conference on Systems, Man and Cybernetics. IEEE Press, Piscataway, NJ, USA, pp. 3274–3279.

Tan, X., Triggs, B., oct 2007. Enhanced local texture feature sets for face recognition under difficult lighting conditions. In: Analysis and Modelling of Faces and Gestures. Vol. 4778 of LNCS. Springer, pp. 168–182.

Vécsei, A., Fuhrmann, T., Liedlgruber, M., Brunauer, L., Payer, H., Uhl, A., 2009. Automated classification of duodenal imagery in celiac disease using evolved fourier feature vectors. Computer Methods and Programs in Biomedicine 95, 68–78.

Vécsei, A., Fuhrmann, T., Uhl, A., 2008. Towards automated diagnosis of celiac disease by computer-assisted classification of duodenal imagery. In: Proceedings of the 4th International Conference on Advances in Medical, Signal and Information Processing (MEDSIP '08). Santa Margherita Ligure, Italy, pp. 1–4, paper no P2.1-009.

Šajn, L., Kononenko, I., January 2008. Multiresolution image parametrization for improving texture classification. EURASIP J. Adv. Signal Process 2008, 137:1–137:12.

Šajn, L., Kukar, M., 2010. Image processing and machine learning for fully automated probabilistic evaluation of medical images. Computer Methods and Programs in Biomedicine In Press, Corrected Proof, –.

Wang, Y., chun Mu, Z., Zeng, H., dec. 2008. Block-based and multi-resolution methods for ear recognition using wavelet transform and uniform local binary patterns. In: Proceedings of the 19th International Conference on Pattern Recognition (ICPR'08). pp. 1–4.

Yokoi, K., 2007. Illumination-robust change detection using texture based features. In: MVA. pp. 487–491.

Zuiderveld, K., 1994. Contrast limited adaptive histogram equalization. In: Heckbert, P. S. (Ed.), Graphics Gems IV. Morgan Kaufmann, pp. 474–485.