On the Computer Assisted Diagnosis of Endoscopic Data with Indication for Celiac Disease

by
DI Sebastian Hegenbart

Cumulative dissertation submitted to the
Faculty of Natural Sciences, University of Salzburg
in partial fulfillment of the requirements
for the Doctoral Degree.

Thesis Supervisor
Univ.-Prof. Mag. rer. nat. Dr. rer. nat. Andreas Uhl

Department of Computer Sciences
University of Salzburg
Jakob Haringer Str. 2
5020 Salzburg, AUSTRIA

Salzburg, October 2014

Abstract

Celiac disease is a complex autoimmune disorder in genetically predisposed individuals of all age groups triggered by the introduction of food containing traces of gluten. Associated complications include osteoporosis, infertility and other autoimmune diseases such as type 1 diabetes, autoimmune thyroid disease and autoimmune liver disease. Once diagnosed, the only treatment is a strict life-long gluten free diet. A reliable diagnosis is therefore of high interest.

Systems for automated diagnosis are an emerging option for medical intervention and endoscopy in particular. Such a system could potentially save costs and manpower while simultaneously increasing the safety of the procedure. The research presented in this thesis was focused towards the development of methods for a computer assisted system for automated diagnosis of celiac disease.

The intrinsic properties of data recorded during flexible endoscopy suggest the interpretation of the automated diagnosis of celiac tissue as a texture classification problem. Consequently, a substantial part of our research was focused on texture classification methods. We established a solid knowledge base for subsequent work by studying the properties of various different feature representations, including scale-invariant methods. To accommodate for the nature of endoscopic data, we subsequently developed a scale- and rotation-adaptive feature representations based on Local Binary Patterns (LBP), which proved to be highly discriminative and robust in endoscopic environments.

Towards the development of a system for computer assisted diagnosis, a significant part of our research was aimed at evaluating characteristic properties of duodenal images and videos. In particular, techniques for the implicit handling of endoscopic image degradations as well as varying gastrointestinal regions and camera-scales were studied.

The nature of medical data differs from the characteristics of data used in more classical texture classification scenarios. As a result of asymmetric patient distributions, medical data sets are often subject to an intrinsic bias, violating assumptions used in cross-validation protocols. In order to establish a solid basis for the evaluation of developed methodologies, we studied the effects of different cross-validation schemes combined with feature optimization on the predictive accuracy.

Finally, we implemented crowd-sourcing into a medical image classification context, focusing on the wide spread problem of limited amounts of available data, annotated by domain experts. By the means of noisy, non-expert labeled data, a classifier was trained which showed to be competitive as compared to a system based on a limited amount of pristine labels.

Abstract (German)

Zöliakie ist eine komplexe Autoimmunerkrankung von genetisch prädisponierten Individuen, welche durch Einnahme glutenhaltiger Nahrungsmittel ausgelöst werden kann. Mit Zöliakie assoziierte Erkrankungen umfassen Osteoporose, Unfruchtbarkeit sowie andere Autoimmunerkrankungen, wie Typ 1 Diabetes, autoimmune Schilddrüsenerkrankungen sowie autoimmune Lebererkrankungen. Die einzig bekannte Behandlungsmöglichkeit nach einer Diagnose besteht aus einer lebenslangen glutenfreien Diät. Eine zuverlässige Diagnose ist daher unerlässlich.

Computergestützte Systeme zur automatisierten Diagnose sind eine aufstrebende Option für medizinische Eingriffe, wie zum Beispiel Endoskopie. Solche Systeme besitzen Potenzial um Kosten, Zeit als auch Arbeitskraft einzusparen. Gegebenenfalls könnte ein solch assistierendes System sogar die Sicherheit der medizinischen Behandlung erhöhen. Der Fokus, der in dieser Dissertation präsentierten Forschung, ist daher auf die Entwicklung von Methoden für ein computergestütztes System zur automatisierten Diagnose von Zöliakie gerichtet.

Aufgrund der intrinsischen Eigenschaften endoskopischer Daten, bieten sich Methoden aus der klassischen Texturklassifikation zur automatisierten Diagnose gastrointestinalen Gewebes an. Infolgedessen bestand ein substanzieller Teil unserer Forschung aus der Entwicklung und Validierung von Texturklassifikationsmethoden. Dies beinhaltete eine umfassende Auswertung der Eigenschaften diverser Verfahren zur Merkmalsextraktion, inklusive skalierungsinvarianter Methoden, welche als solide Basis für weitere Entwicklungen auf diesem Gebiet diente. Um den intrinsischen Eigenschaften endoskopischer Daten Rechnung zu tragen, wurden infolgedessen skalierungs- sowie rotationsinvariante Merkmalsextraktionsverfahren, basierend auf Local Binary Patterns (LBP), entwickelt, welche sich durch hohe Unterscheidbarkeit und Robustheit in endoskopischen Szenarien auszeichneten.

Ein weiterer fundamentaler Teil unserer Arbeit hinsichtlich der Entwicklung eines computergestützten Systems, bezog sich auf die Untersuchung charakteristischer Eigenschaften von Videos und Bildern aus dem Duodenum. Techniken für den impliziten Umgang mit endoskopischen Bildstörungen sowie variierender gastrointestinaler Regionen und verschiedener Kameradistanzen wurden ebenso untersucht.

Die Eigenschaften medizinischer Daten unterscheiden sich üblicherweise von denen gewöhnlicher Daten, welche in klassischeren Texturklassifikationsszenarien eingesetzt werden. Aufgrund der asymmetrischen Verteilung gesunder und erkrankter Patienten sind medizinische Daten oft einer intrinsischen statistischen Verzerrung ausgesetzt, aufgrund derer Annahmen von Kreuzvalidierungsverfahren verletzt werden. Hinsichtlich der Schaffung einer fundierten Grundlage zur Auswertung entwickelter Methoden, wurde der Einfluss verschiedener Kreuzvalidierungsverfahren in Kombination mit Merkmalsoptimierungsverfahren auf die Genauigkeit der getroffenen Vorhersagen untersucht.

Schlussendlich setzten wir Crowdsourcing in einem medizinischen Bildklassifikationskontext ein, um das allgemeine Problem der unausreichenden Menge verfügbarer, annotierter, medizinischer Daten zu lösen. Mithilfe einer, von nicht-medizinisch-geschulten Personen erstellten ungenauen Grundwahrheit, konnte ein Klassifikationssystem mit einer Leistung, vergleichbar zu einem System basierend auf einer kleineren Anzahl an Daten mit exakter Grundwahrheit, konstruiert werden.

Acknowledgments

The work towards this thesis was a challenging experience. I faced numerous situations which required a significant amount of stamina and faith in my ideas and concepts to keep going. In such times (particularly), it might have not been easy to bear with me. I therefore want to thank my partner Linda for her enduring support and my kids Fi and Tobias for providing the motivation for long working hours. Without the wholehearted support of my parents Barbara and Reinhard during my entire academic career, my work would not have been possible.

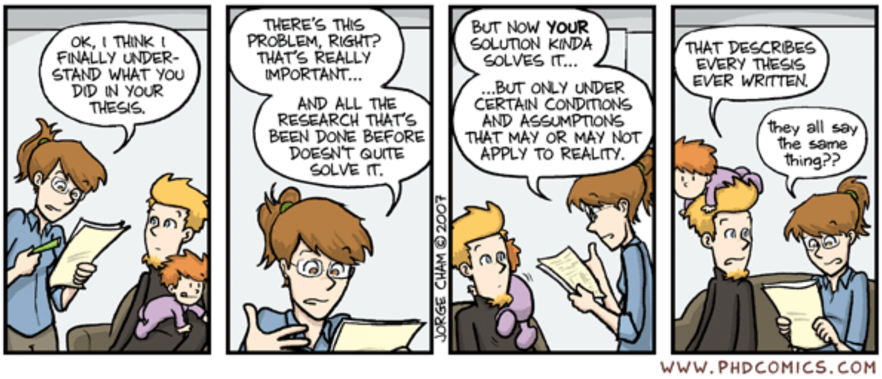
I am very grateful for the opportunity to work in this fascinating field of science, which was given to me by my advisor Andreas Uhl. The work in a medical context would not have been possible without the outstanding engagement by ”our” medical expert Andreas Vécsei, who not only provided a large amount of medical data but repeatedly visited our research lab for valuable discussion and training.

Finally, I want to give credit to all my colleagues in the *wavelab* research group for numerous exciting discussions and valuable input but also for becoming my friends.

This thesis has been funded in part by the Austrian Science Fund, project no. 24366 and the Austrian National Bank ”Jubiläumsfonds”, project no. 12991.

Salzburg, October 2014 *Sebastian Hegenbart*

Contents


OK, I THINK I FINALLY UNDERSTAND WHAT YOU DID IN YOUR THESIS.
THERE'S THIS PROBLEM, RIGHT? THAT'S REALLY IMPORTANT...
AND ALL THE RESEARCH THAT'S BEEN DONE BEFORE DOESN'T QUITE SOLVE IT.
JORGE CHAM © 2007
BUT NOW YOUR SOLUTION KINDA SOLVES IT...
...BUT ONLY UNDER CERTAIN CONDITIONS AND ASSUMPTIONS THAT MAY OR MAY NOT APPLY TO REALITY.
THAT DESCRIBES EVERY THESIS EVER WRITTEN.
they all say the same thing??
WWW.PHDCOMICS.COM

1. Introduction

This cumulative dissertation covers my research performed at the University of Salzburg as a member of the *wavelab* group lead by Andreas Uhl. My research was focused towards the development of methods for a computer assisted system for automated diagnosis of celiac disease in standard flexible endoscopy. Systems for automated diagnosis are an emerging option for endoscopic treatments (e.g. [39, 2, 1, 38]) and could potentially be used to save costs, time and manpower while simultaneously increasing the safety of the procedure. Computer assisted diagnosis systems could be used to develop less invasive approaches avoiding biopsy. Studies by Cammarota et al. [4, 5], investigating such endoscopic techniques, report reliable results.

This thesis is structured as follows. Section 1.1 will give an overview of celiac disease, its gastrointestinal manifestations and the clinical routine used in diagnosis. Section 1.2 puts the presented research into context of related work in the field of automated diagnosis of celiac disease in endoscopic data. To improve the readability, the presented contributions are arranged into three categories:

- Methods for Texture Classification (Section 2)
- Characteristic Properties of Duodenal Images and Videos (Section 3)
- Classification and Performance Prediction in Medical Data (Section 4)

Section 5 gives a summarizing conclusion of the research presented in this thesis. The actual publications can be found in Section 6.

Please note, that a substantial amount of the presented work was a joint effort with other researchers. I will therefore use the plural form when referring to our research during the text. A breakdown of the contributions of each author for each publication can be found in the appendix.

1.1. Celiac Disease and Endoscopy

Celiac disease is a complex autoimmune disorder in genetically predisposed individuals of all age groups triggered by the introduction of food containing traces of gluten. The gastrointestinal manifestations invariably comprise an inflammatory reaction within the mucosa of the

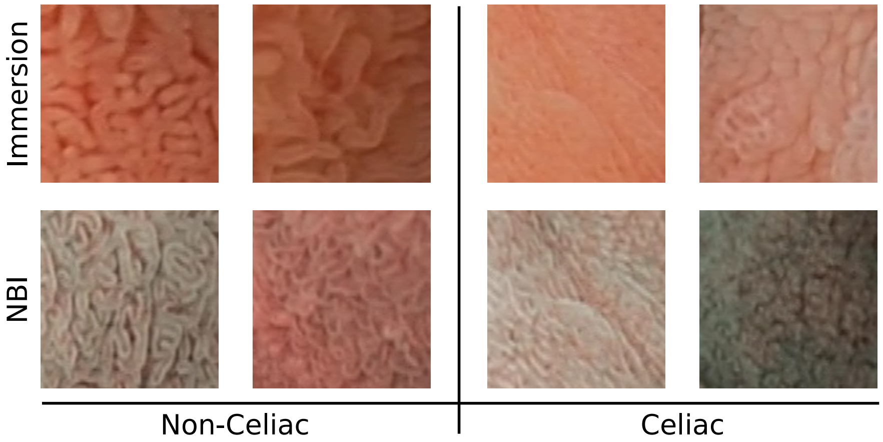


Figure 1.1.: Healthy Mucosal Tissue and Tissue affected by Villous Atrophy.

small intestine caused by a dysregulated immune response. During the course of the disease, hyperplasia of the enteric crypts occurs and the mucosa eventually loses 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 continuously been adjusted upwards. Fasano et al. [15] 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. Figure 1.1 illustrates the visualized appearance of healthy duodenal mucosa and tissue affected by villous atrophy.

The clinical routine for screening subjects with indications for celiac disease includes serological assays using endomysial antibody (EMA) and tissue transglutaminase (tTG) tests, possibly followed by upper endoscopy to perform duodenal biopsy for a histological confirmation. Guidelines recommend more then four endoscopic duodenal biopsies. Due to the patchy nature of villous atrophy [3, 36], duodenal biopsies could potentially miss abnormalities (regions affected by the disease). The correct targeting of biopsies is therefore essential for the histological confirmation of the disease. Computer assisted technology could potentially be used to guide the targeting of biopsies, consequently improving the accuracy of the diagnosis.

The most common scheme used to assess the severity of celiac disease in duodenal tissue is the modified Marsh-Oberhuber classification [42]. This histological classification scheme identifies six classes, ranging from class Marsh-0 (no visible change of villi structure) up to class Marsh-3C (absent villi). Figure 1.2 illustrates the Marsh staging schematically.

- **Marsh 0-2**: No visible changes of villi structure
- **Marsh 3A**: Mild villous atrophy
- **Marsh 3B**: Marked villous atrophy
- **Marsh 3C**: Absent villi

Besides standard upper endoscopy, several new endoscopic approaches for diagnosing celiac disease have been applied [7]. The modified immersion technique [4] is based on the instillation

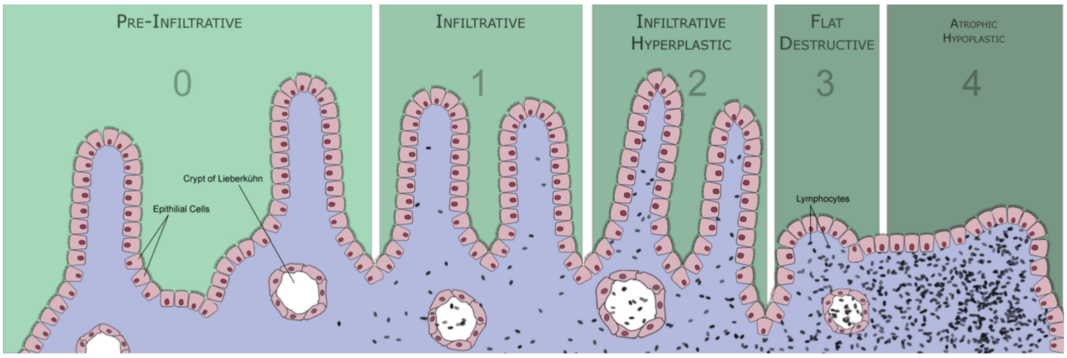


Figure 1.2.: Schematic Marsh Staging of Mucosal Tissue Affected by Celiac Disease [52].

of water into the duodenal lumen for better visualization of the villi. Furthermore, magnifying endoscopy (standard endoscopy with additional magnification) has been investigated [6]. Narrow band imaging (NBI [13]) has recently been used to enhance the contrast of vascular patterns on the mucosal surface. Valitutti et al. [48] recently proposed the use of NBI combined with the water immersion technique. Confocal endomicroscopy is a novel technology allowing real-time in vivo microscopy which has been shown to be a promising technique to diagnose celiac disease in endoscopy by Leong et al. [41].

A drawback of standard upper endoscopy using a flexible endoscope is the limited range. As a way of inspecting a much larger area of the intestine, wireless capsule endoscopy (WCE [45]) is used. In WCE, a small capsule equipped with a camera is swallowed by the patient. The capsule records images of the mucosal tissue during its passage through the intestine. In contrast to the high resolution of the data provided by fully-fledged flexible endoscopes, images captured during WCE are usually at significantly lower resolutions due to obvious restrictions on the hardware's size. WCE sequences are characterized by slow, monotonic movement. This is a result of the passive camera movement through the intestine caused by peristalsis. As a consequence of the nature of WCE, spatio-temporal features are frequently used to analyze the visualized gastrointestinal tissue.

Data captured during flexible endoscopy on the other hand, is characterized by rapid, nonmonotonic movement. The high resolution of the data combined with unpredictable movement, suggests to approach the automated diagnosis as a classical still image texture classification scenario.

Due to the missing guidance of the capsule and the lacking option of performing biopsy, WCE and flexible endoscopy are quite complementary. As a result of the different characteristics of the provided data, it is unlikely, that all methods developed for WCE can be applied in flexible endoscopy and vice versa.

1.2. Related Work

At this point in time, most of the research performed on the automated diagnosis of celiac disease in endoscopic data is done by two independent research groups. In a complementary fashion, the group around Edward J. Ciaccio has been focusing heavily on WCE imagery, employing spatio-temporal features as well as using the intestinal motility to assess the degree of villous atrophy. The group around Andreas Uhl has been working on data from flexible endoscopy, approaching the problem from a texture classification perspective.

Promising spatio-temporal features identified by Ciaccio et al. [8] include the statistics of local tissue brightness in a temporal context of multiple WCE images. In another work [12], salient information is computed based on the identification of the dominant period in a series of WCE images. Celiac tissue is identified using spatial statistics of a series of basis images.

A promising concept is based on the analysis of the motility of the duodenal wall [10]. Based on this approach areas suspect of being affected by celiac disease could be identified. The effects of villous atrophy on the visualized periodicity of peristalsis were also used to identify affected regions [9]. This was done by the means of dominant frequency analysis.

In an approach that was also evaluated in flexible endoscopy, Ciacco et al. assess the degree of villous atrophy by explicit measurement of the length of mucosal fissures [11]. This approach however requires a calibration of the surface dimension which is a challenging task in flexible endoscopy due to varying camera-scales. Finally, shape from shading was explored to analyze the luminal macro-architecture in a sequence of WCE images [11].

Contributions of the group around Uhl have shown strong empirical evidence in multiple experimental studies that feature representations used in more classical texture classification scenarios [51, 50] are feasible in the context of automated diagnosis of celiac disease. To accommodate for the characteristics of flexible endoscopy, scale- and viewpoint-invariant feature representations have been used [47] and developed in particular for the classification of celiac disease [18].

The effects of endoscopic lens distortions as well as the benefits of distortion correction have been extensively studied [25, 16, 23, 26]. The authors identified that interpolation artifacts and varying camera-scales are the main restriction on distortion correction performance [17]. As a consequence, distortion adaptive classification [20], distortion compensated features [19] as well as intrinsic distortion correction [22] have been used. Finally, the effects of endoscopic lens distortion on the diagnosis accuracy of domain experts have been studied [21].

Recently, Grisan et al. [24] were the first to use confocal endomicroscopy for the automated diagnosis of celiac disease. In their approach, statistical features and LBP were used on a pyramidal decomposition of images.

2. Contributions: Methods for Texture Classification

Due to the characteristics of flexible endoscopy (rapid motion, motion blur, non-monotonic transition of the gastrointestinal tract, high resolution of data), it is difficult to employ features in a temporal context. Consequently, we approached the automated diagnosis from a more classical texture classification perspective. As a result, a substantial amount of our research was focused on methods for texture classification.

Initial work towards the development of a computer assisted system for diagnosis of celiac disease was focused on studying the properties and applicability of various feature representations used in the field of pattern recognition.

We learned that local texture operators based on Local Binary Patterns (LBP [43]) are very promising for the classification of celiac disease. Varying camera-scales were identified to significantly affect the classification performance. Consequently scale-invariant features representations were studied.

Finally, a scale- and orientation-adaptive feature representation based on LBP was developed.

Publications (sorted chronologically)

Vécsei, A., Amann, G., Hegenbart, S., Liedlgruber, M., and Uhl, A. Automated Marsh-like Classification of Celiac Disease in Children using Optimized Local Texture Operators. *Computers in Biology and Medicine 41*, 6 (2011), 313 – 325

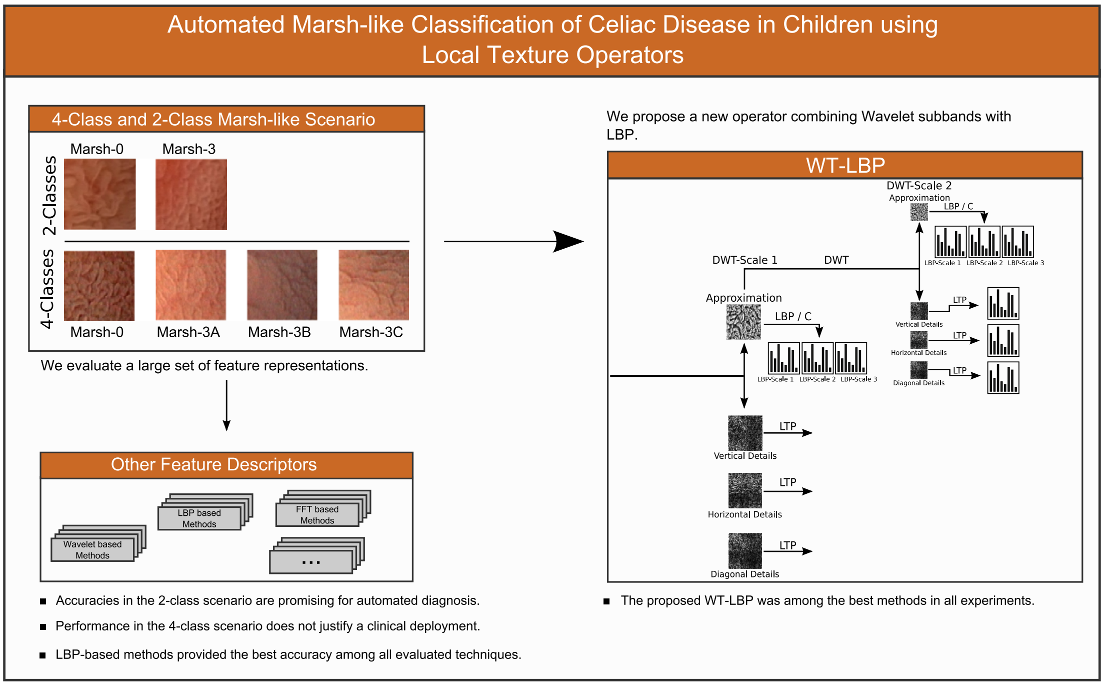
Hegenbart, S., Uhl, A., Vécsei, A., and Wimmer, G. Scale Invariant Texture Descriptors for Classifying Celiac Disease. *Medical Image Analysis 17*, 4 (2013), 458 – 474

Hegenbart, S., and Uhl, A. A Scale-Adaptive Extension to Methods based on LBP using Scale-Normalized Laplacian of Gaussian Extrema in Scale-Space. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '14)* (2014), pp. 4352 – 4356

Hegenbart, S., and Uhl, A. An Orientation-Adaptive Extension to Scale-Adaptive Local Binary Patterns. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR'14)* (2014), pp. 1120 – 1125

Hegenbart, S., Uhl, A., and Vécsei, A. A Scale- and Orientation-Adaptive Extension of Local Binary Patterns. Tech. Rep. 2014-05, Department of Computer Sciences, University of Salzburg, Austria, 2014. `http://uni-salzburg.at/index.php?id=38565`; Submitted to Elsevier Journal on Pattern Recognition (October 2014): Under Review

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



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

We evaluate a large set of feature representations.

We propose a new operator combining Wavelet subbands with LBP.

- Accuracies in the 2-class scenario are promising for automated diagnosis.
- Performance in the 4-class scenario does not justify a clinical deployment.
- LBP-based methods provided the best accuracy among all evaluated techniques.
- The proposed WT-LBP was among the best methods in all experiments.

In this initial publication, we were the first to study the automated diagnosis of celiac disease in flexible endoscopy, based on a reduced 4-class Marsh-like classification system. The main focus of our work was on the evaluation of various feature representations, in particular local texture operators based on LBP.

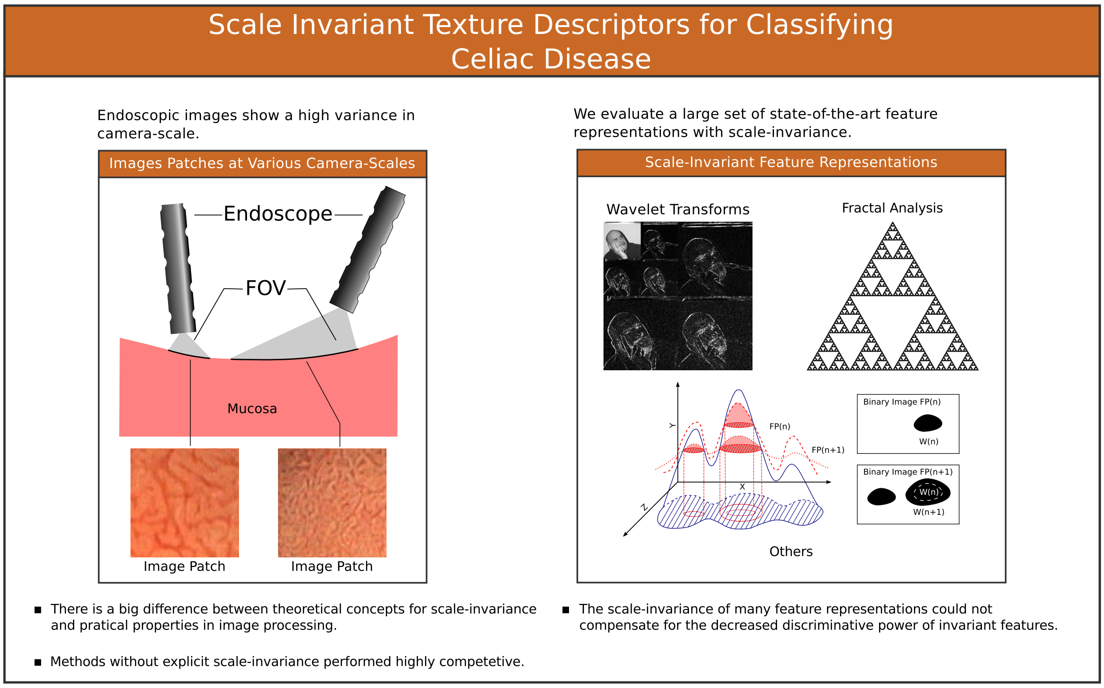
The experimental evaluation indicated, that LBP-based feature representations are highly competitive in this scenario. We consequently put more emphasize on employing LBP-based methods for automated classification in subsequent work.

We also proposed two new LBP-based methodologies, particularly optimized for endoscopic environments. The WT-LBP (Wavelet Transform - LBP) method was designed to combine different wavelet subbands with appropriate LBP-based operators, while the ELTP (Extended Local Ternary Patterns) method combines the benefits of the robust Local Ternary Patterns (LTP [46]) with the highly discriminative Extended Local Binary Patterns (ELBP [37]), using an adaptive thresholding based on local image statistics.

During experimentation, we gained strong empirical evidence, that the modified Marsh-Oberhuber scheme poses an ill-suited problem for visual texture classification. This is a consequence of the non-distinct visual appearance of tissue in classes of type Marsh-3A to Marsh-3C in endoscopic imagery. We concluded, that more simplified, visually focused systems such as proposed by Ensari [14] may be better suited for automated diagnosis.

Due to the available histological ground-truth following the Marsh-Oberhuber classification scheme, we subsequently focused on the most clinically-relevant binary categorization between tissue affected by villous atrophy and healthy mucosa.

2.2. Scale Invariant Texture Descriptors for Classifying Celiac Disease



Scale Invariant Texture Descriptors for Classifying Celiac Disease

Endoscopic images show a high variance in camera-scale.

Images Patches at Various Camera-Scales

We evaluate a large set of state-of-the-art feature representations with scale-invariance.

Scale-Invariant Feature Representations

- There is a big difference between theoretical concepts for scale-invariance and pratical properties in image processing.
- Methods without explicit scale-invariance performed highly competetive.
- The scale-invariance of many feature representations could not compensate for the decreased discriminative power of invariant features.

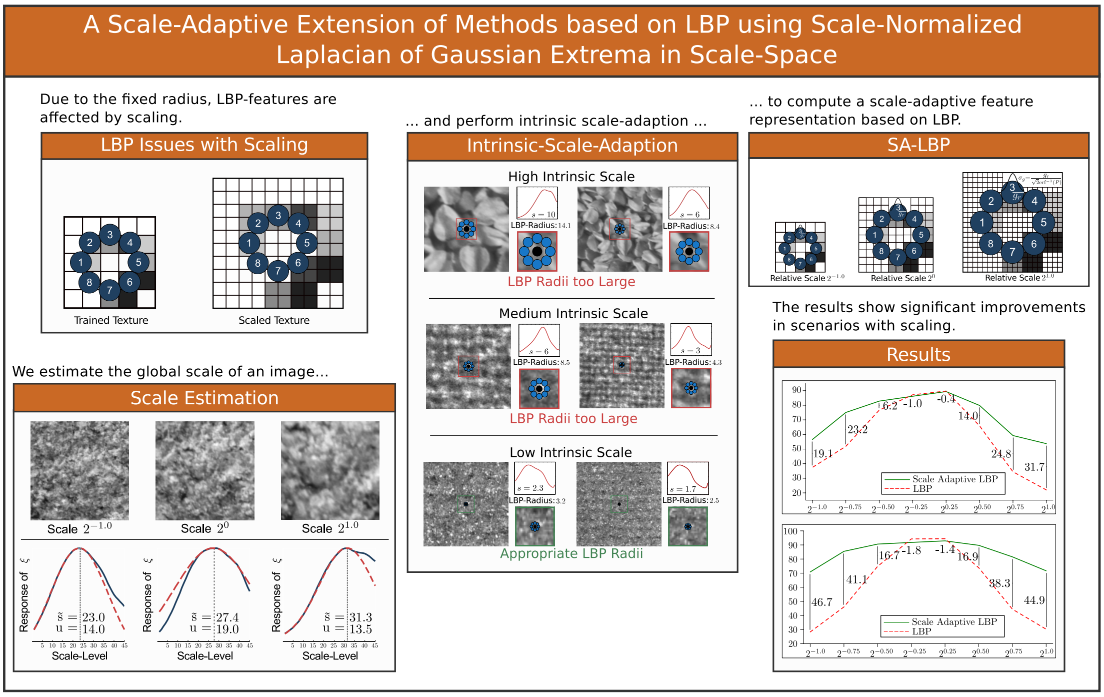
As a consequence of the manual guidance, rapid changes of scenery and a high variation of viewpoint and camera-scale is very common in endoscopic sequences. Earlier work [32] (see Section 3.2) indicated that varying camera-scales affect the performance of automated diagnosis. We therefore studied the benefits of scale-invariant feature representations in the context of celiac disease classification.

A large set of state-of-the art techniques were evaluated with emphasis on multi-scale and multi-orientation wavelet transform based methods. In the same work, we proposed the computation of an affine-invariant LTP-based feature representation using local texture scale and shape information employing Laplacian of Gaussian (LoG) maxima and multi-scale second moment matrices in a scale-space representations of endoscopic images.

We learned that the scale-invariance properties of a large number of feature representations are based on theoretical concepts and assumptions, which rarely hold in practice. As a consequence, the scale-invariance of such features is questionable. Even more, scale-invariant feature representations often exhibit a decreased discriminative power as compared to other, non-invariant feature representations. Consequently, the experimental evaluation showed, that features specifically designed for scale-invariance performed comparable to non-invariant features.

The proposed affine-invariant computation of LTP proved to be highly competitive though and motivated us to put more effort into combining highly discriminative LBP based features with a scale- and orientation-adaptive computation in subsequent work.

2.3. A Scale-Adaptive Extension to Methods based on LBP using Scale-Normalized Laplacian of Gaussian Extrema in Scale-Space



A Scale-Adaptive Extension of Methods based on LBP using Scale-Normalized Laplacian of Gaussian Extrema in Scale-Space

Due to the fixed radius, LBP-features are affected by scaling.

We estimate the global scale of an image...

... and perform intrinsic scale-adaption ...

... to compute a scale-adaptive feature representation based on LBP.

The results show significant improvements in scenarios with scaling.

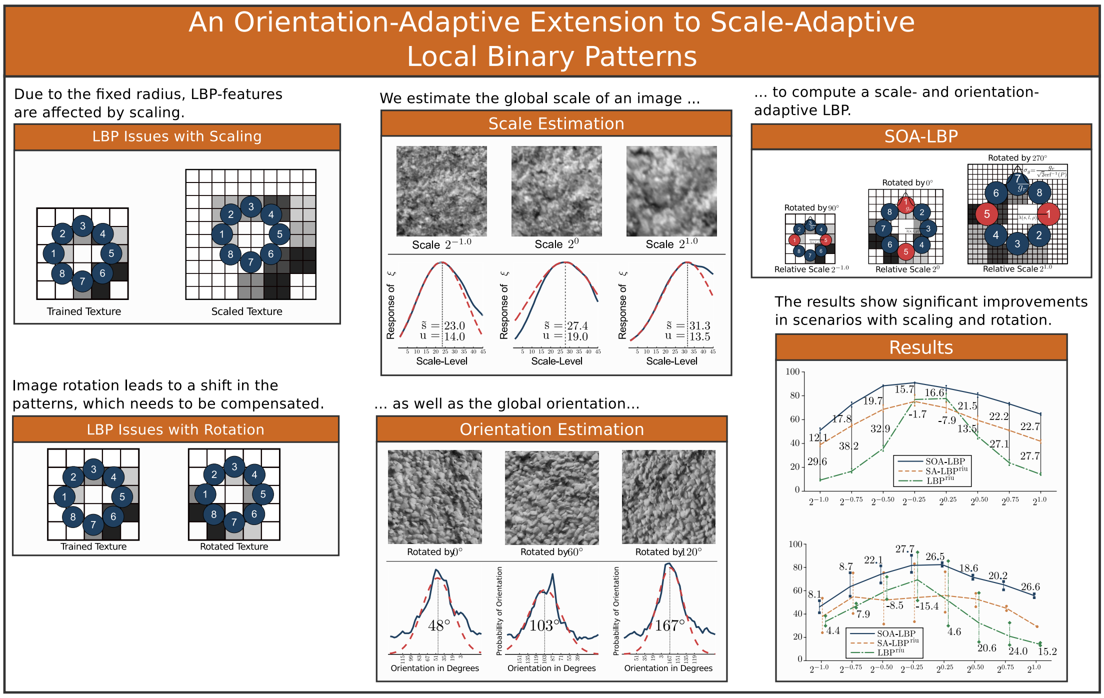
A main restriction of LBP based features is the sensitivity to affine transformations. This is a direct consequence of the fixed-scale radius and the fixed sampling area dimension of the pixel neighborhood. Locally computed patterns implicitly encode the underlying micro structures of a texture at a scale directly related to the camera-scale of an image. As a result, the LBP feature representation can not compensate for different camera-scales, a common thing in endoscopy.

Based on the prior observation, that general scale-invariant feature representations exhibit a decreased discriminative power and that theoretical concepts and assumptions in scale-invariance rarely hold in practice, we developed a general framework for computation of scale-invariant LBP, combining highly discriminative LBP based features with a reliable scale-adaptive computation.

We propose a scale-invariant LBP representation, based on the estimation of the global texture scale. The distribution of scale-normalized LoG responses in a scale-space representation is used for scale-estimation. Intrinsic-scale-adaption is performed to compute features independent of the intrinsic texture scale, leading to a significantly increased discriminative power for a large amount of texture classes. The experimental results showed significantly improved classification accuracies compared to standard LBP-based methods in scenarios with large scale differences for two publicly available texture databases (KTH-TIPS and Kylberg).

Due to the heavy focus on the methodology in this work, we used publicly available databases for the experimental evaluation to allow reproducibility in the pattern recognition community. Although no celiac data was used in this particular publication, the presented methodology was extended in subsequent work and was successfully used to improve the classification accuracy of celiac tissue in scenarios with varying scales [33] (see Section 2.5).

2.4. An Orientation-Adaptive Extension to Scale-Adaptive Local Binary Patterns



An Orientation-Adaptive Extension to Scale-Adaptive Local Binary Patterns

Due to the fixed radius, LBP-features are affected by scaling.

We estimate the global scale of an image ...

... to compute a scale- and orientation-adaptive LBP.

Image rotation leads to a shift in the patterns, which needs to be compensated.

... as well as the global orientation...

The results show significant improvements in scenarios with scaling and rotation.

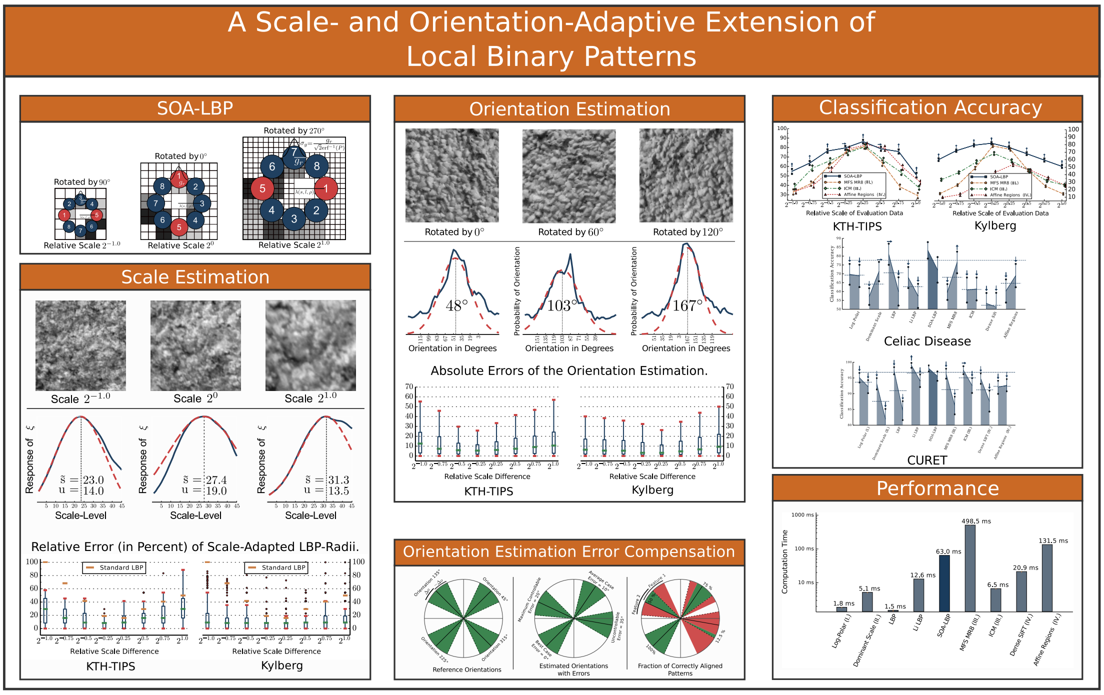
Rotation of an image is reflected as a circular shift in the individual patterns of LBP, which affects the distribution of patterns in a non-linear fashion. As a consequence, the standard LBP feature representation requires either an implicit or explicit alignment of patterns to compensate for image rotations. This is generally done at the encoding level. A major limitation of such encoding level based approaches however is the highly limited angular resolution.

Based on our prior work on the scale-adaptive computation of LBP, we solved this limitation by an explicit alignment of patterns at the extraction level, using a robust estimate of global texture orientation.

We estimate the global orientation of a texture by computing multi-scale second moment matrices at a dense grid. The orientation at a specific location is determined as the angle between the major axis of the ellipse represented by the second moment matrix and the vertical axis of the coordinate system (the axes of the image). We then estimate the global orientation of an image, based on the distribution of local orientations, computed at all coordinates of the sampled grid.

The experimental evaluation showed significantly improved classification accuracies in scenarios with scaling and rotation. Again, publicly available databases were used instead of the celiac data in favor of reproducibility.

2.5. A Scale- and Orientation-Adaptive Extension of Local Binary Patterns



A Scale- and Orientation-Adaptive Extension of Local Binary Patterns

SOA-LBP

Orientation Estimation

Classification Accuracy

Scale Estimation

Performance

Orientation Estimation Error Compensation

In our final work on rotation- and scale-invariant LBP, we proposed a multi-resolution feature representation, improving the general descriptive power by reducing the required amount of low-pass filtering for adapting the sampling area and adding the capability of describing underlying micro structures at multiple scales.

We improved the orientation-adaptive representation by applying an error compensation technique based on the accumulation of LBP distributions at multiple orientations. Exploiting the systematic error introduced by image rotation, this technique allows to compensate orientation estimation errors of up to 20 degrees.

In a detailed experimental evaluation, the effects of scaling and rotation on the proposed method are studied in reference to a representative set of scale-invariant features. Experimentation is based on four different data sets, including endoscopic data with indication for celiac disease.

The proposed method was significantly superior to all evaluated methods in case of large scale differences. The proposed multi-resolution feature representation was more than competitive in scenarios with tiny scale differences. Experimentation based on the noisy CURET data and the endoscopic data with indication for celiac disease showed, that the proposed methodology provides discriminative and reliable features in such difficult scenarios.

Experimental evidence indicates, that the proposed methodology is current state-of-the art in classifying celiac disease in scenarios with varying camera-scales.

3. Contributions: Characteristic Properties of Duodenal Images and Videos

Endoscopic sequences provide a challenging scenario for automated diagnosis. Image degradations such as blur, bubbles, specular reflections and noise are common in such data. Due to the manual guidance by a clinician, rapid, non-monotonic movement of the camera is common during flexible endoscopy. Additionally, the camera-scale and viewpoint towards the tissue shows a significant variation. As a consequence, the visualized appearance of the intestine varies in an unpredictable manner.

We approached the challenging endoscopic scenario using a one-class support vector machine (SVM) to implicitly handle the most common endoscopic image degradations. We further evaluated that approach in the context of multiple-gastrointestinal regions and camera-scales.

Finally, we studied the effects of interlaced scanning and evaluated the benefits of suited de-interlacing techniques on the performance of automated diagnosis.

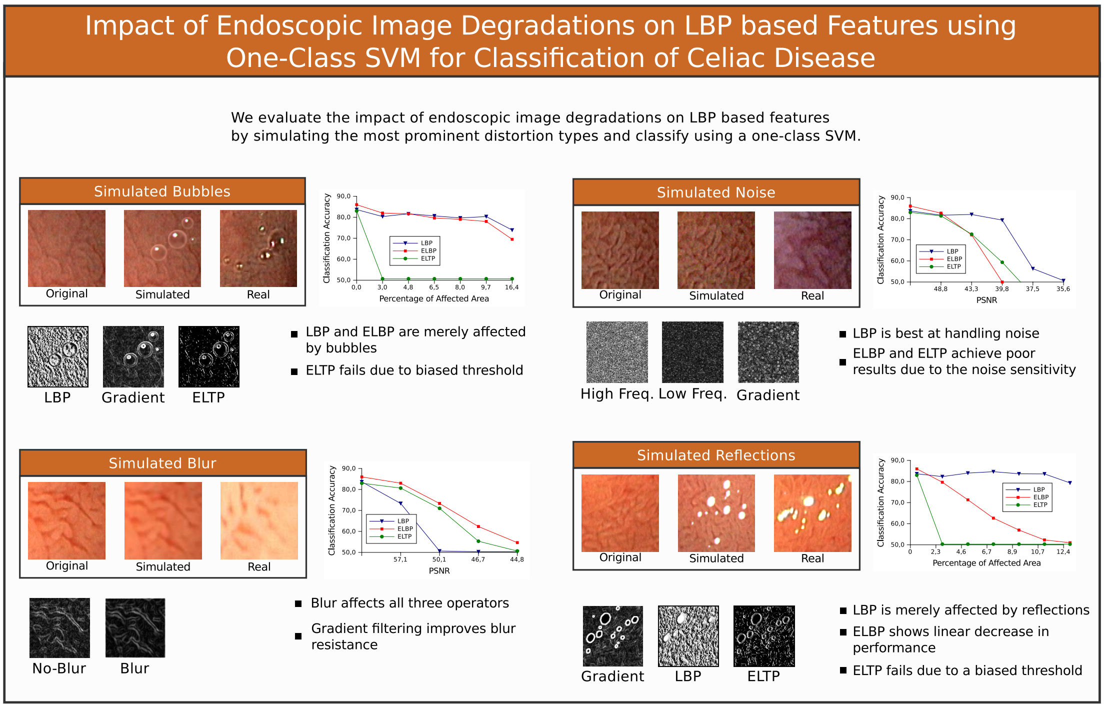
Publications (sorted chronologically)

Hegenbart, S., Uhl, A., and Vécsei, A. Impact of Endoscopic Image Degradations on LBP based Features using One-Class SVM for Classification of Celiac Disease. In *Proceedings of the 7th International Symposium on Image and Signal Processing and Analysis (ISPA'11)* (Dubrovnik, Croatia, 2011), pp. 715 – 720

Hegenbart, S., Uhl, A., and Vécsei, A. On the Implicit Handling of Varying Distances and Gastrointestinal Regions in Endoscopic Video Sequences with Indication for Celiac Disease. In *Proceedings of the IEEE International Symposium on Computer-Based Medical Systems (CBMS'12)* (2012), pp. 1 – 6

Hegenbart, S., Uhl, A., Vécsei, A., and Wimmer, G. On the Effects of De-Interlacing on the Classification Accuracy of Interlaced Endoscopic Videos with Indication for Celiac Disease. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems (CBMS'13)* (2013), pp. 137 – 142

3.1. Impact of Endoscopic Image Degradations on LBP based Features using One-Class SVM for Classification of Celiac Disease



Impact of Endoscopic Image Degradations on LBP based Features using One-Class SVM for Classification of Celiac Disease

We evaluate the impact of endoscopic image degradations on LBP based features by simulating the most prominent distortion types and classify using a one-class SVM.

Simulated Bubbles

- LBP and ELBP are merely affected by bubbles
- ELTP fails due to biased threshold

Simulated Noise

- LBP is best at handling noise
- ELBP and ELTP achieve poor results due to the noise sensitivity

Simulated Blur

- Blur affects all three operators
- Gradient filtering improves blur resistance

Simulated Reflections

- LBP is merely affected by reflections
- ELBP shows linear decrease in performance
- ELTP fails due to a biased threshold

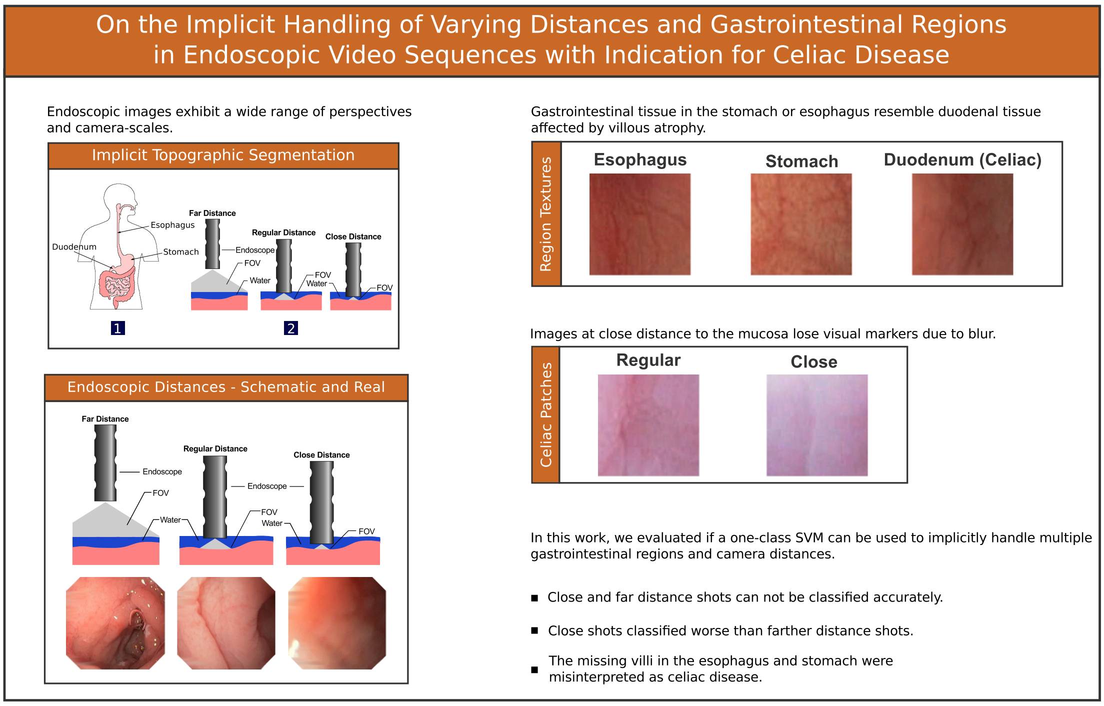
Endoscopic image degradations such as blur, bubbles, noise and specular reflections are a common nuisance in endoscopic data. Although it is safe to assume that this sort of degradations has a negative impact on automated diagnosis, it is generally unclear to what degree the classification accuracy is affected.

An explicit approach to handle such degradations is based on informative frame identification, possibly in combination with some sort of segmentation technique. Consequently, the performance of a system based on such an approach is highly affected by the reliability of this pre-processing step. As an alternative to informative frame identification and segmentation, we investigated the implicit handling of such image degradations, employing a one-class SVM classifier trained specifically on tissue affected by villous atrophy.

In an experimental evaluation using LBP-based methods for feature extraction, the most common image degradations were simulated to allow a fine grained analysis of the individual effects on the classification accuracy. The results give evidence, that certain types of image degradations, such as bubbles and reflections, only affect a subset of LBP-based methods and can actually be compensated using the proposed approach. Blur and noise had the most impact on LBP-based features.

We concluded, that unconstrained classification of celiac disease based on LBP using a one-class SVM is feasible to some degree. Extreme cases of image distortions might require an additional step of informative frame identification however. By relaxing the needs for informative frame identification to extreme cases, the general reliability of a fully automated system could possibly be increased.

3.2. On the Implicit Handling of Varying Distances and Gastrointestinal Regions in Endoscopic Video Sequences with Indication for Celiac Disease


On the Implicit Handling of Varying Distances and Gastrointestinal Regions in Endoscopic Video Sequences with Indication for Celiac Disease

Endoscopic images exhibit a wide range of perspectives and camera-scales.

Implicit Topographic Segmentation

Endoscopic Distances - Schematic and Real

Gastrointestinal tissue in the stomach or esophagus resemble duodenal tissue affected by villous atrophy.

Images at close distance to the mucosa lose visual markers due to blur.

In this work, we evaluated if a one-class SVM can be used to implicitly handle multiple gastrointestinal regions and camera distances.

- Close and far distance shots can not be classified accurately.
- Close shots classified worse than farther distance shots.
- The missing villi in the esophagus and stomach were misinterpreted as celiac disease.

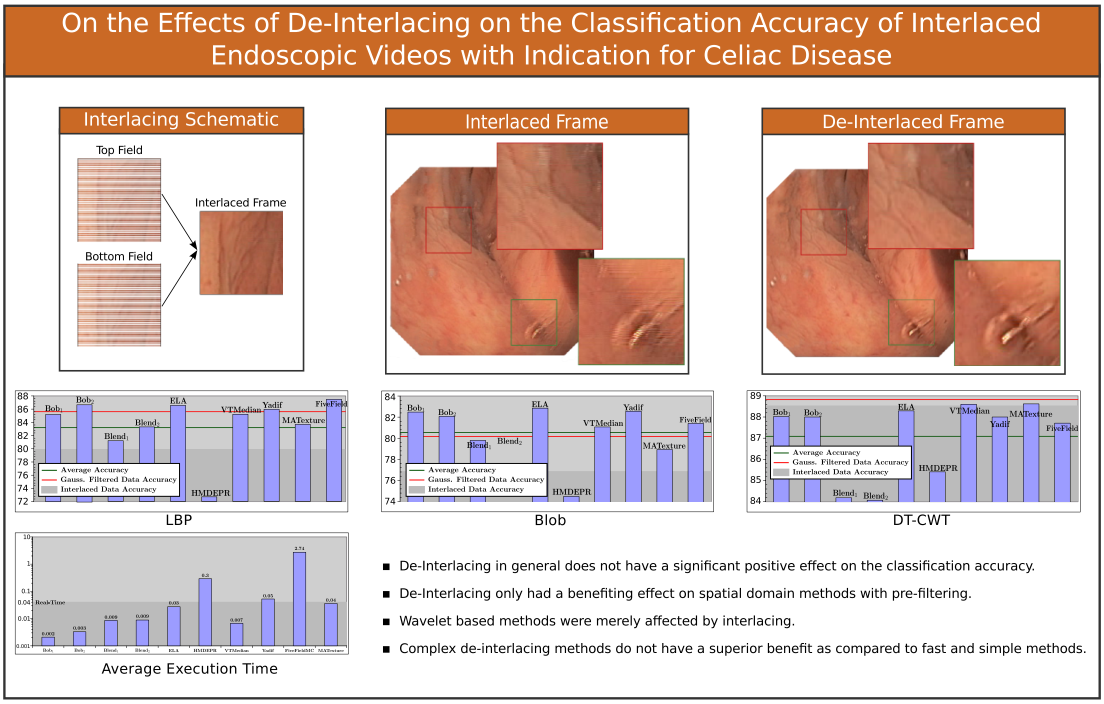
Sequences in flexible endoscopy are commonly non-monotonic in terms of the passage through gastrointestinal regions. As a consequence of the visual appearance of esophageal and gastric tissue in endoscopic data, it is possible that the missing villi structures in these regions are misinterpreted as celiac disease by an automated system.

Additionally, inappropriate camera-scales either lead to a blurred visualization of the mucosa (close distance) or to a missing visualization of small spatial structures (far distance) and could consequently be unsuited for visual classification. We therefore evaluated the impact of varying camera distances and gastrointestinal regions on the classification pipeline (employing a one-class SVM) proposed previously.

The experiments indicated, that the visualization of mucosal tissue at close and far distances is unsuited for classification with LBP-based methods. Even more, the missing villi structures in the stomach and esophagus were misinterpreted as celiac disease by our classification system. It is clear that the scaling-affected LBP-based methods are unsuited for feature extraction in these extreme scenarios. As a consequence, we focused on scale-invariant and scale-adaptive feature representations (as presented in Section 2) in subsequent work.

In order to handle the non-monotonic visualization of multiple gastrointestinal regions in a fully automated system, an additional mechanism for tracking the endoscope's location or identification of the intestinal region is required.

3.3. On the Effects of De-Interlacing on the Classification Accuracy of Interlaced Endoscopic Videos with Indication for Celiac Disease



On the Effects of De-Interlacing on the Classification Accuracy of Interlaced Endoscopic Videos with Indication for Celiac Disease

- De-Interlacing in general does not have a significant positive effect on the classification accuracy.
- De-Interlacing only had a benefiting effect on spatial domain methods with pre-filtering.
- Wavelet based methods were merely affected by interlacing.
- Complex de-interlacing methods do not have a superior benefit as compared to fast and simple methods.

Interlaced scanning is a technique that has been widely used to double the perceived frame rate without increasing the required bandwidth. This technique is still in use by endoscopic video hardware today. Various specialized de-interlacing techniques have been developed over the last decades to re-construct full frames from two interlaced half-frames.

The impact of interlaced scanning and the benefits of suited de-interlacing techniques on the classification accuracy of automated diagnosis was studied in this work.

We learned, that de-interlacing does not have a significant positive effect on the classification accuracy of endoscopic data with indication for celiac disease. The benefits of applying de-interlacing were comparable to the effects of Gaussian filtering. We were unable to identify a difference between simple and more complex de-interlacing techniques considering the classification accuracy.

As a consequence, we concluded that a system for automated diagnosis can operate on interlaced data without significant loss of accuracy.

4. Contributions: Classification and Performance Prediction in Medical Data

The nature of medical data differs from the characteristics of data used in more classical texture classification scenarios. As a consequence of the limited amount of available data and the asymmetric distributions of healthy and unhealthy patients, data sets in medical image classification often contain an unavoidable intrinsic bias.

Hence, assumptions used in cross-validation protocols, used for predicting how well developed methodologies will generalize on independent data, are violated and the predicted performance is subject to bias and error.

Typically, a substantial amount of correctly labeled medical data is required to construct systems for automated diagnosis. The limited amount of expert labeled medical data is therefore another actual problem in the field of research.

Our contributions in this section are focused on these issues of medical image classification. We studied the effects of data-bias and feature optimization on the predictive accuracy of different cross-validation techniques. This was specifically done in the context of classification of celiac disease but should apply to a variety of scenarios in medical image classification.

We finally evaluated the practical use of crowd-sourced annotations of medical data by non-experts to construct systems for automated diagnosis.

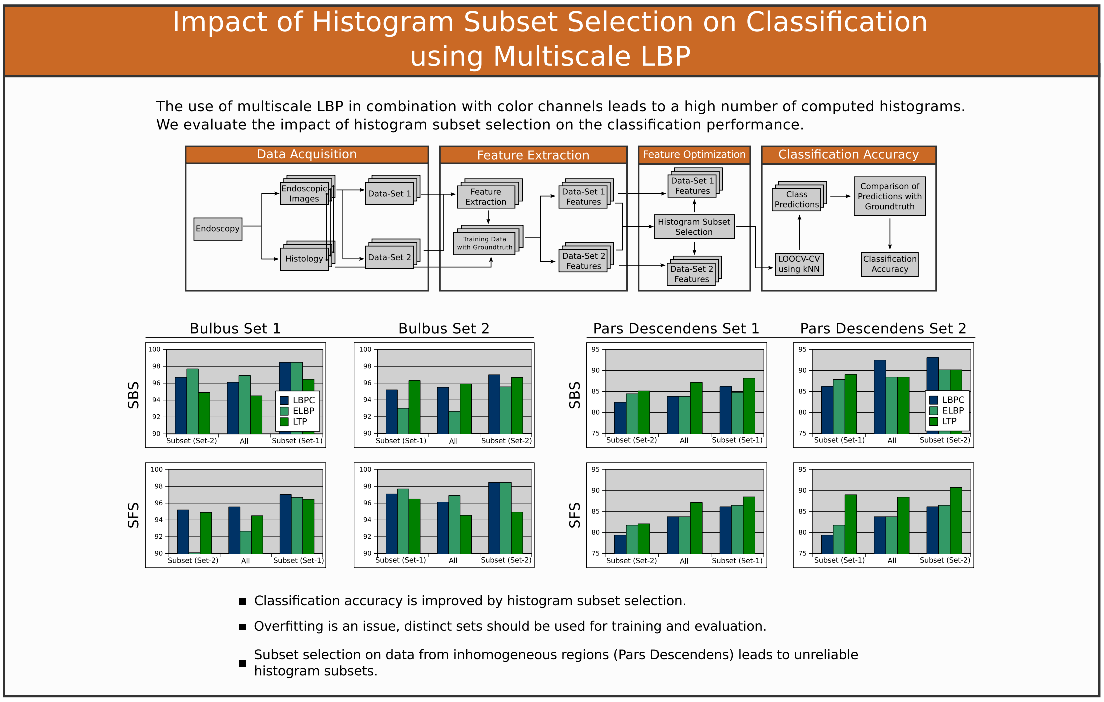
Publications (sorted chronologically)

HEGENBART, S., UHL, A., AND VÉCSEI, A. Impact of Histogram Subset Selection on Classification using Multiscale LBP. In *Proceedings of Bildverarbeitung für die Medizin 2011 (BVM'11)* (Lübeck, Germany, 2011), Informatik aktuell, pp. 359 – 363

HEGENBART, S., UHL, A., AND VÉCSEI, A. Systematic Assessment of Performance Prediction Techniques in Medical Image Classification - A Case Study on Celiac Disease. In *Proceedings of the 22nd International Conference on Information Processing in Medical Imaging (IPMI'11)* (Monastery Irsee, Germany, 2011), pp. 498 – 508

KWITT, R., HEGENBART, S., RASIWASIA, N., VÉCSEI, A., AND UHL, A. Do we Need Annotation Experts? A Case Study in Celiac Disease Classification. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'14)* (2014), pp. 454 – 461

4.1. Impact of Histogram Subset Selection on Classification using Multiscale LBP



Impact of Histogram Subset Selection on Classification using Multiscale LBP

The use of multiscale LBP in combination with color channels leads to a high number of computed histograms. We evaluate the impact of histogram subset selection on the classification performance.

- Classification accuracy is improved by histogram subset selection.
- Overfitting is an issue, distinct sets should be used for training and evaluation.
- Subset selection on data from inhomogeneous regions (Pars Descendens) leads to unreliable histogram subsets.

Multiscale-LBP [44] is used to describe underlying micro structures of an image at multiple scales, possibly improving the descriptive power of the features. The feature representation is constructed by concatenation of multiple LBP-histograms, each computed at a separate fixed LBP-radius.

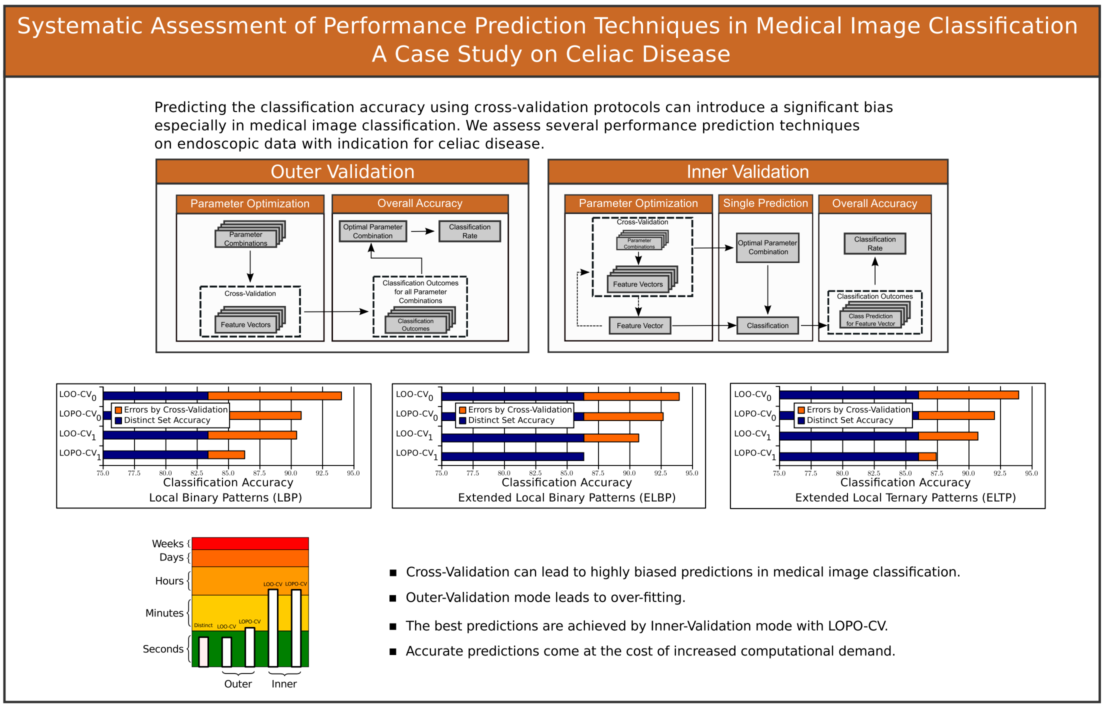
In combination with different color channels, this leads to a high number of possibly indiscriminate LBP-histograms. It is likely that not all scales and color channels are well suited for the classification of celiac disease.

Consequently, we studied the use of feature subset selection for identification of suitable and discriminative features in a multiscale-LBP feature representation

The experiments provided empirical evidence, that feature subset selection improves the classification accuracy in the context of automated diagnosis of celiac disease. Over-fitting was an issue however and cross-validation protocols should be chosen with care.

Following the experimental results, we decided to study the effects of cross-validation and feature optimization in a medical texture classification context in more detail in subsequent work (see Section 4.2).

4.2. Systematic Assessment of Performance Prediction Techniques in Medical Image Classification - A Case Study on Celiac Disease



Systematic Assessment of Performance Prediction Techniques in Medical Image Classification
A Case Study on Celiac Disease

Predicting the classification accuracy using cross-validation protocols can introduce a significant bias especially in medical image classification. We assess several performance prediction techniques on endoscopic data with indication for celiac disease.

- Cross-Validation can lead to highly biased predictions in medical image classification.
- Outer-Validation mode leads to over-fitting.
- The best predictions are achieved by Inner-Validation mode with LOPO-CV.
- Accurate predictions come at the cost of increased computational demand.

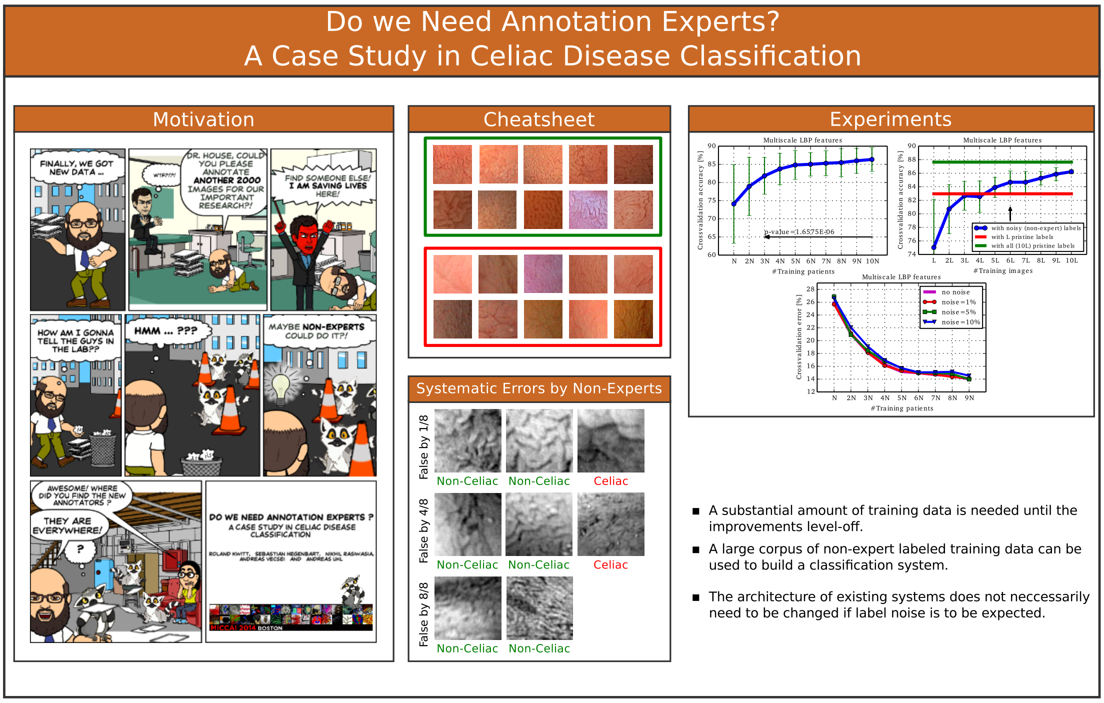
Cross-validation protocols are used in scenarios with a limited amount of available data for evaluation. As a consequence of the small number of patients, multiple samples are often used per patient to build data sets. This results in an intrinsic bias of the data set and violates certain assumptions of cross-validation protocols.

To gain more insight and to establish a solid basis for evaluation, we specifically assessed the predictive accuracy of cross-validation techniques as well as the effects of feature optimization in the classification of celiac disease.

The experiments indicated, that cross-validation can lead to highly biased predictions in medical image classification. Inappropriate validation protocols can result in a significant extent of over-fitting, providing inaccurate predictions and mislead research.

The best predictive accuracy was achieved by nested cross-validations, using a data partitioning based on a patient basis instead of an image basis. Although the computational complexity of this scheme is significantly higher as compared to running a single cross-validation, the predicted accuracies are much more realistic in this evaluation methodology.

4.3. Do we Need Annotation Experts? A Case Study in Celiac Disease Classification



Do we Need Annotation Experts?
A Case Study in Celiac Disease Classification

Motivation

Cheatsheet

Experiments

Systematic Errors by Non-Experts

- A substantial amount of training data is needed until the improvements level-off.
- A large corpus of non-expert labeled training data can be used to build a classification system.
- The architecture of existing systems does not neccessarily need to be changed if label noise is to be expected.

FINALLY, WE GOT NEW DATA ...
WTF?!?!
DR. HOUSE, COULD YOU PLEASE ANNOTATE ANOTHER 2000 IMAGES FOR OUR IMPORTANT RESEARCH?!
FIND SOMEONE ELSE! I AM SAVING LIVES HERE!
HOW AM I GONNA TELL THE GUYS IN THE LAB??
HMM ... ???
MAYBE NON-EXPERTS COULD DO IT?!
AWESOME! WHERE DID YOU FIND THE NEW ANNOTATORS ?
THEY ARE EVERYWHERE!
?
DO WE NEED ANNOTATION EXPERTS ?
A CASE STUDY IN CELIAC DISEASE CLASSIFICATION
ROLAND KWITT, SEBASTIAN HEGENBART, NIKHIL RASIWASIA, ANDREAS VECSEI AND ANDREAS UHL
MICCAI 2014 BOSTON
False by 1/8
Non-Celiac Non-Celiac Celiac
False by 4/8
Non-Celiac Non-Celiac Celiac
False by 8/8
Non-Celiac Non-Celiac
Multiscale LBP features
Crossvalidation accuracy [%]
90 85 80 75 70 65 60
N 2N 3N 4N 5N 6N 7N 8N 9N 10N
#Training patients
p-value=1.6575E-06
Multiscale LBP features
Crossvalidation accuracy [%]
90 88 86 84 82 80 78 76 74
L 2L 3L 4L 5L 6L 7L 8L 9L 10L
#Training images
with noisy (non-expert) labels
with L pristine labels
with all (10L) pristine labels
Multiscale LBP features
Crossvalidation error [%]
28 26 24 22 20 18 16 14 12
N 2N 3N 4N 5N 6N 7N 8N 9N
#Training patients
no noise
noise=1%
noise=5%
noise=10%

The general lack of medical data annotated by domain experts is a major problem in the field of research. It is unclear how methods developed and evaluated on such small datasets generalize.

Typically, a substantial amount of data is required to find well generalizing decision boundaries for classification. As a consequence, the small amount of expert labeled medical data often holds back research progress.

In this work, we proposed the use of medical data labeled by non-experts, to train classification boundaries in a celiac disease classification scenario. Following the idea of crowd-sourcing, our data was re-labeled by non-experts after a small amount of training. The noisy class labels were then used to construct a classification system.

Experimentation showed evidence, that label noise can be compensated by a sufficiently large corpus of training data, annotated by non-experts. In contrary to the explicit handling of label noise, no change in the classification architecture is required by such an approach. In scenarios where data acquisition is not the limiting factor, this could substantially broaden the use of computer aided diagnosis.

5. Conclusion

During the course of my work, we were able to build a solid foundation towards the development of a system for automated diagnosis of celiac disease. The systematic assessment of prediction errors in cross-validation protocols was essential for the subsequent development of methods in a medical context. We have shown that an intrinsic bias in medical data leads to violated assumptions in cross-validation methods. As a consequence, the predicted performance of classification systems can be subject to error. We suggest to use cross-validation partitions on a patient basis instead of an image basis to reduce this sort of bias. Feature optimization in combination with cross-validation and small data sets has to be performed with care. We propose to use a nested cross-validation scheme to avoid over-fitting effects.

We identified LBP based methods to be a very promising feature representation for automated diagnosis of celiac disease. Motivated by the characteristics of endoscopic images, we proposed two highly competitive LBP-based methods. The WT-LBP method combines multiple subbands of the wavelet transformation with appropriate LBP-operators. We also introduced a combination of LTP with ELBP using an adaptive thresholding based on local image information (ELTP).

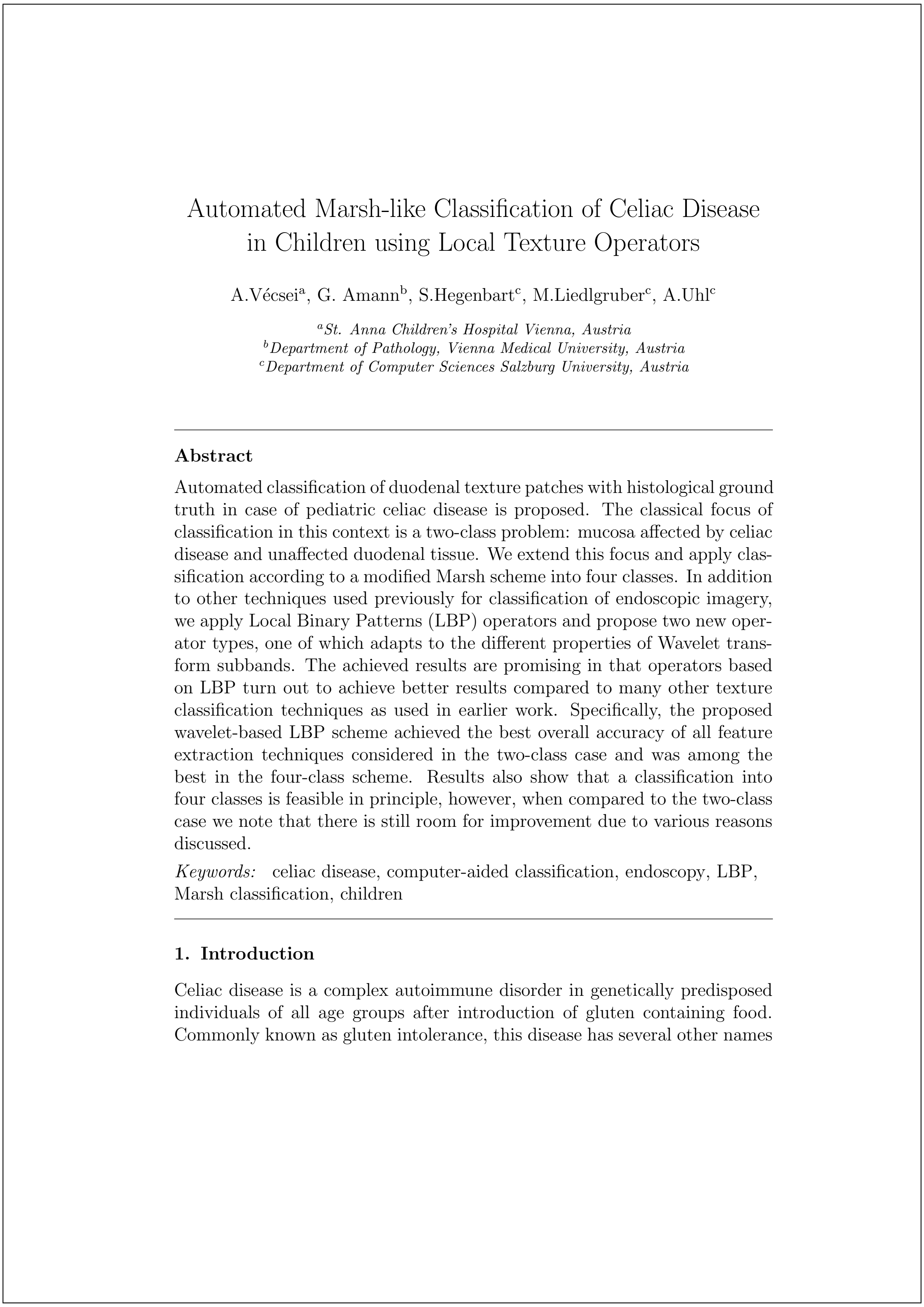
Highly varying camera-scales, as a natural characteristic of flexible endoscopy, were identified to have a significant impact on the accuracy of automated diagnosis. As a consequence, scale-invariant feature representations were extensively studied in the context of classifying celiac disease. Experimental evidence was given, that scale-invariant features do not pose a significant benefit in this context due to the generally lower discriminative power of such features. Consequently, a scale- and orientation-adaptive feature representation based on highly discriminative LBP was developed. We were able to show, that the proposed methodology is current state-of-the art in classifying celiac disease in scenarios with varying camera-scales.

We learned, that the most common image degradations in endoscopy such as bubbles, specular reflections, blur and noise can be handled implicitly, to a certain degree, using a one-class SVM with LBP-based features. The missing villous structures in esophageal and gastric mucosa however is misclassified as celiac disease. As a consequence, a methodology for identification of gastrointestinal regions will be required by a fully automated system based on this approach. We also found, that interlaced scanning used in endoscopic hardware does not significantly affect the performance of automated diagnosis and can potentially be handled without using complex de-interlacing techniques.

The limited amount of expert labeled data in medical imaging is an actual problem in the field of research. We presented empirical evidence that a large corpus of non-expert labeled training data can be used to build a classification system that performs comparable to a system trained solely on a limited number of pristine labels. Implementing crowd-sourcing in medical image classification, this could potentially broaden the use of computer aided diagnosis if data acquisition is not the limiting factor.

6. Publications

This chapter presents publications as originally published. The copyright of the original publications is held by the respective copyright holders, see the following copyright notices. In order to fit the paper dimension, reprinted publications may be scaled in size and/or cropped.



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*

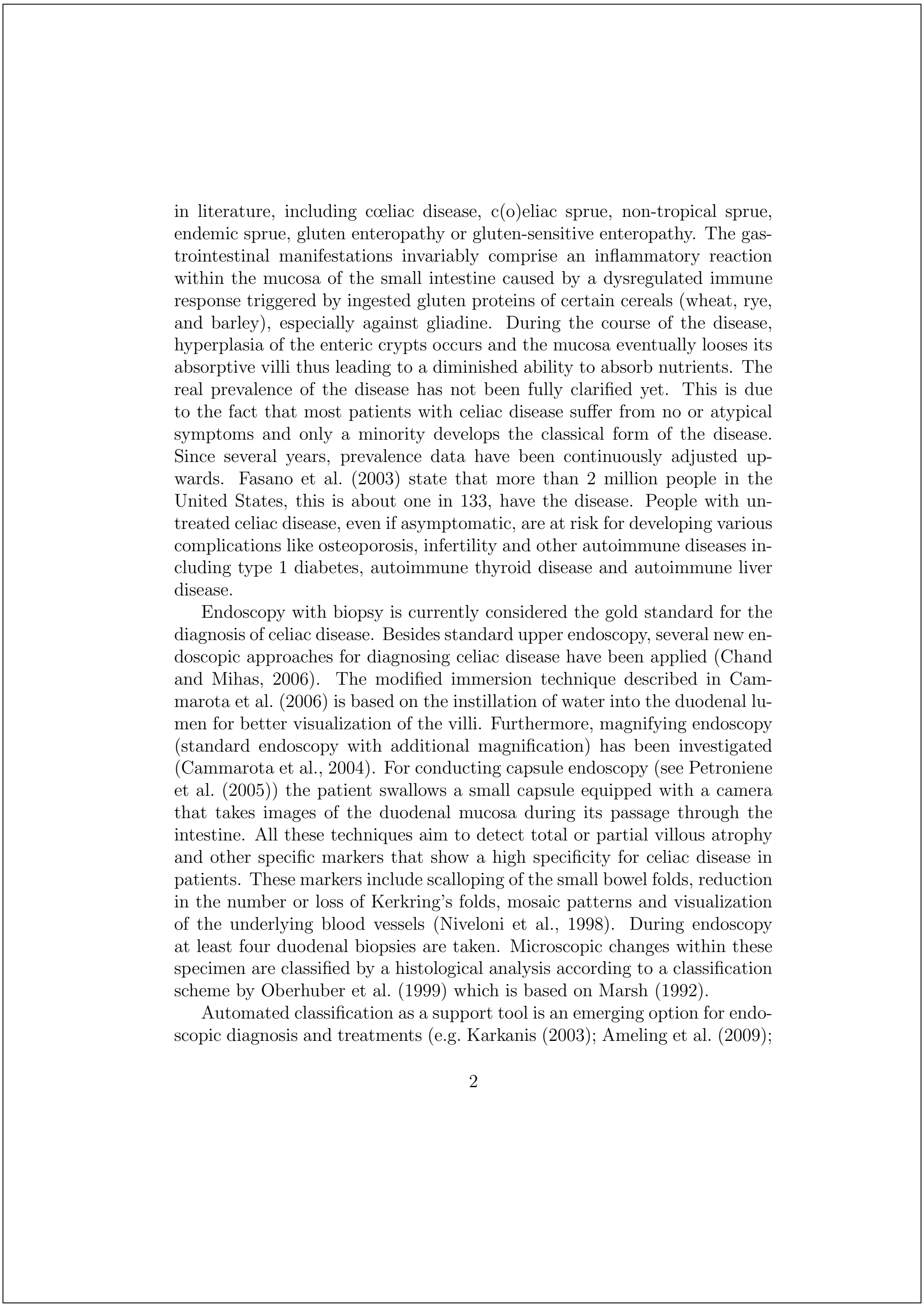
Abstract

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

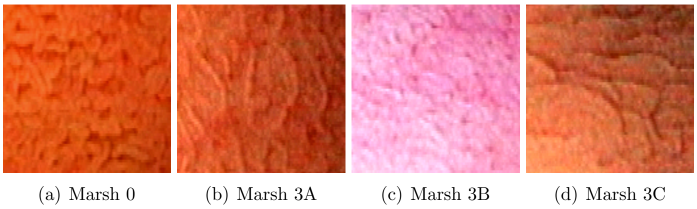
Table 1: Characteristic Changes of Mucosal Tissue caused by Celiac Disease.

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 generator 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.

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

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

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

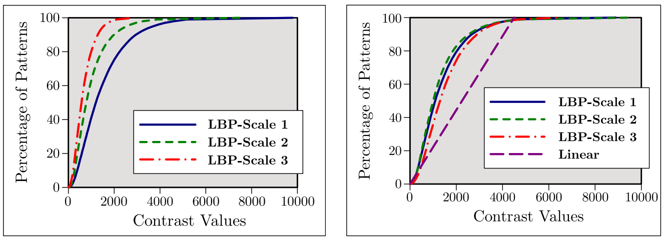
with

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

$C_{r,p}$ 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

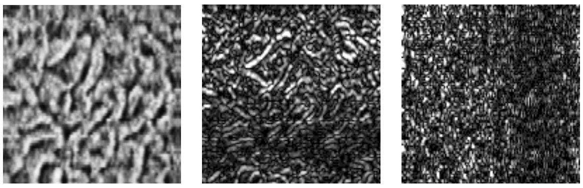
3.3. The Wavelet based LBP Operator (WT-LBP)

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

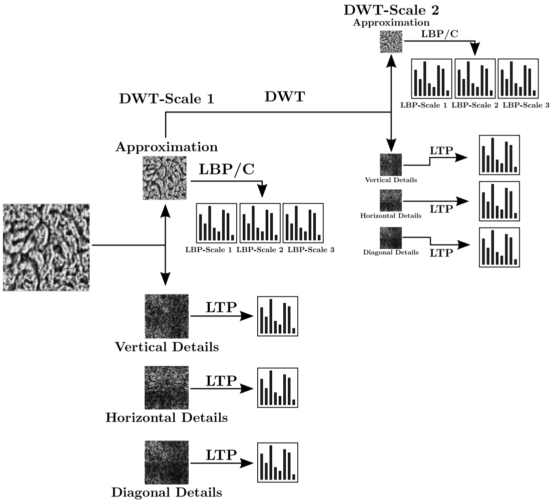


Figure 4: Two-Scale Wavelet Based LBP Operator (WT-LBP-Operator).

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 is interest-

	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).

ing 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 classification

[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

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 suggest, if possible, to use a modification to the leave-one-out cross valida-

[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).

	α = 0.025			α = 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.

tion 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.

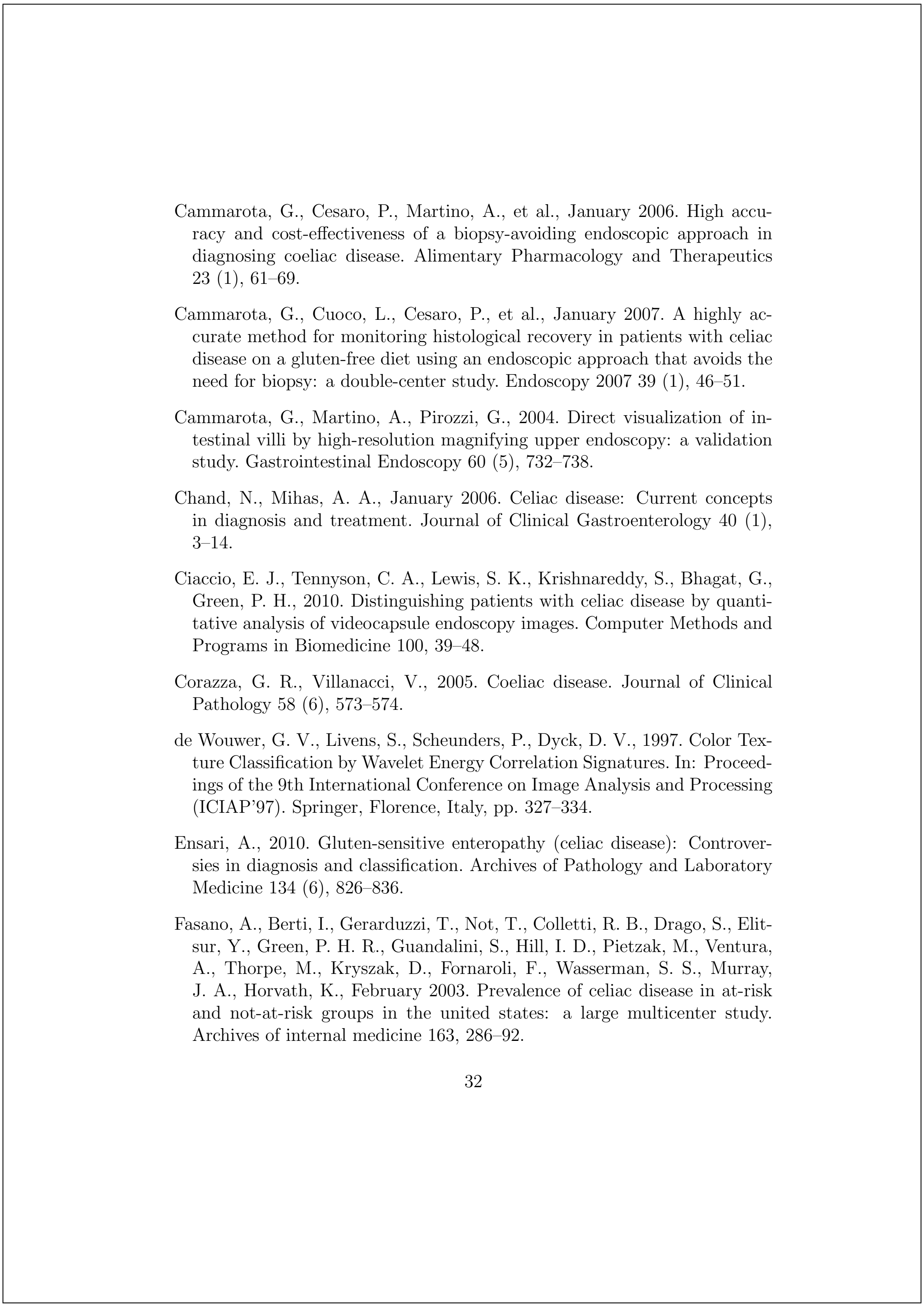
Acknowledgements

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.

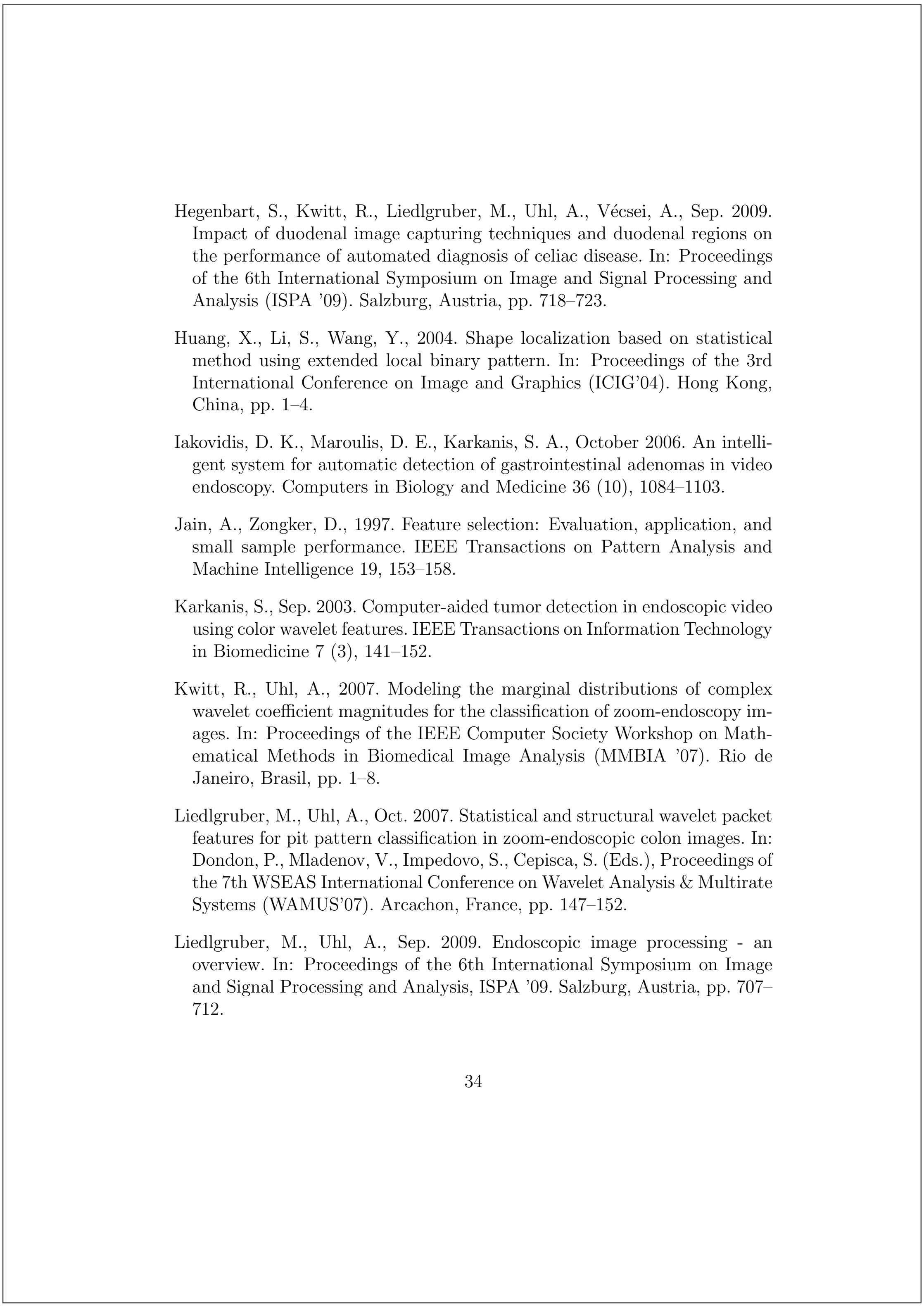
Gross, S. A., Wallace, M. B., December 2006. Hold on Picasso, narrow band imaging is here. American Journal of Gastroenterology 101 (12), 2717–2718.

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.

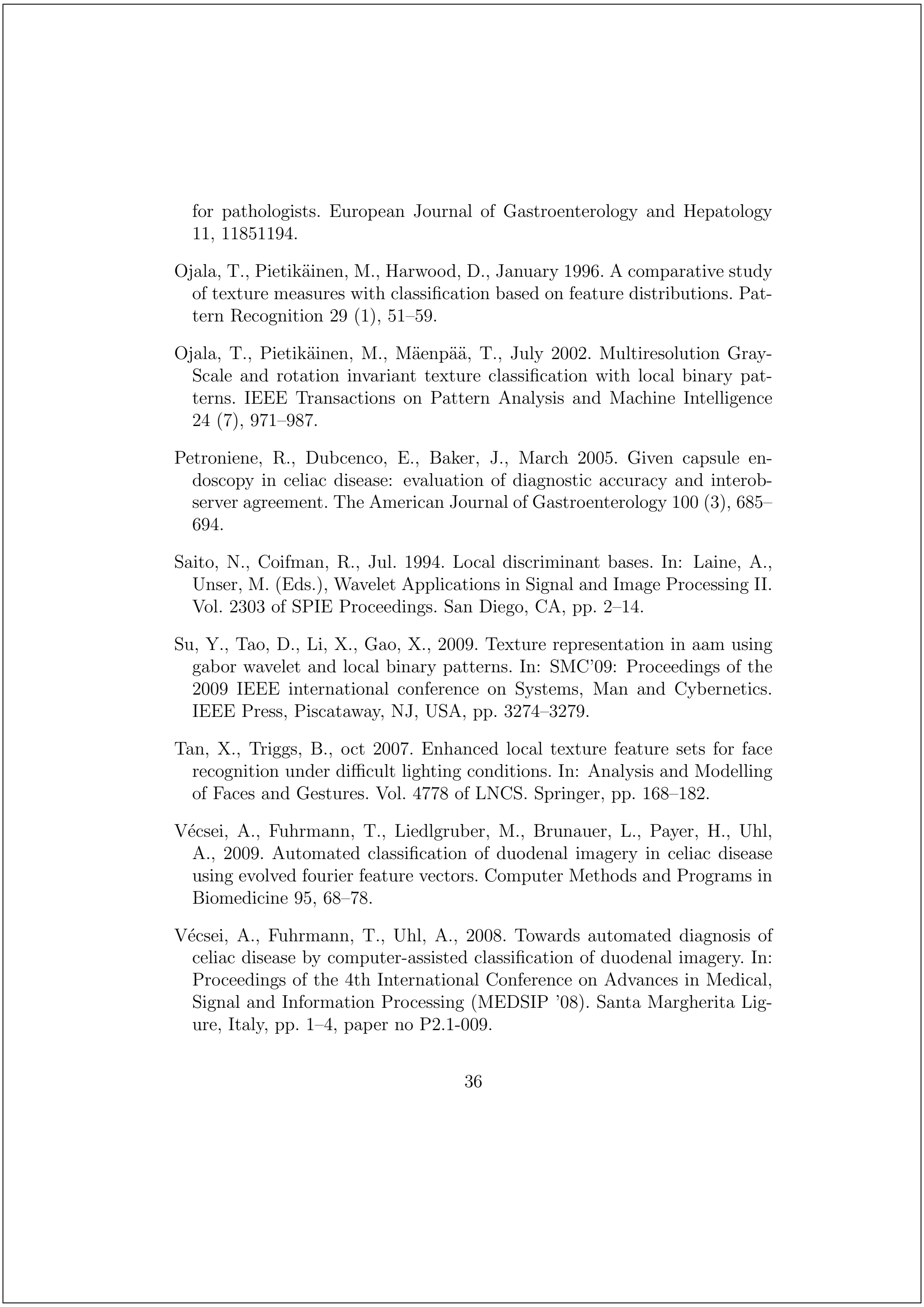
Manjunath, B. S., Ma, W. Y., Aug. 1996. Texture features for browsing and retrieval of image data. IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (8), 837–842.

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. 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.

Scale Invariant Texture Descriptors for Classifying Celiac Disease

S.Hegenbart[a], A. Uhl[a], A. Vécsei[b], G. Wimmer[a,*]

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

Abstract

Scale invariant texture recognition methods are applied for the computer assisted diagnosis of celiac disease. In particular, emphasis is given to techniques enhancing the scale invariance of multi-scale and multi-orientation wavelet transforms and methods based on fractal analysis. After fine-tuning to specific properties of our celiac disease imagery database, which consists of endoscopic images of the duodenum, some scale invariant (and often even viewpoint invariant) methods provide classification results improving the current state of the art. However, not each of the investigated scale invariant methods is applicable successfully to our dataset. Therefore, the scale invariance of the employed approaches is explicitly assessed and it is found that many of the analyzed methods are not as scale invariant as they theoretically should be. Results imply that scale invariance is not a key-feature required for successful classification of our celiac disease dataset.

Keywords: scale invariance, texture recognition, celiac disease

1. Introduction

Texture analysis is one of the fundamental issues in image processing. The majority of existing texture analysis methods work with the assumption that texture images are acquired from the same viewpoint (Zhang and Tan, 2002). This limitation makes these methods useless for applications, where textures occur with different scales, orientations, or translations. Therefore, scale and orientation invariant texture analysis approaches have been proposed (see Tan (1995) or Zhang and Tan (2002) for surveys on this topic). Invariance is important for many applications in medical image processing, since medical images are often acquired at different scales and viewpoints. This is especially true for endoscopic imagery since mucosal texture is seen from different perspectives and distances to the cavity wall depending on the relative position of the endoscopes tip and the mucosa surface. Figure 1 illustrates that, depending on the angle between endoscope and the surface (middle case) and the curvature of the surface (rightmost example), different distances between camera and surface may even occur within a single image.

In gastroscopic (and other types of endoscopic) imagery, mucosal texture is usually found with different perspective and scale (see Figure 3). That means that the mucosal texture shows different spatial scales, depending on the camera perspective and distance to the mucosal wall (see Figure 1).

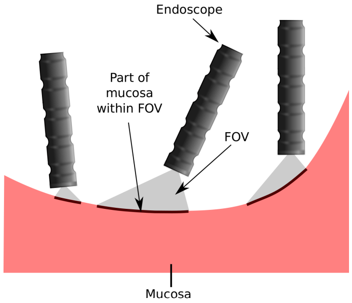


Figure 1: The field of view (FOV) depending on the endoscopic viewpoint and distance to the mucosal wall

As a consequence, endoscopic imagery typically exhibits mucosal texture with different and/or mixed spatial scales, depending on the corresponding acquisition conditions (see Figure 3 for examples from our celiac disease database).

In this work, we focus on scale invariant texture classification approaches being applied in computer-assisted diagnosis of celiac disease. While most of the used techniques in this work exhibit additional invariance to other transformations like rotation, translation, and illumination, we specifically concentrate on scale invariance for the reasons explained above. The contributions of this manuscript are as follows:

- We apply general purpose scale invariant texture descriptors for the classification of duodenal mucosa texture imagery aiming at the staging of celiac disease.

[*] Corresponding author
Email addresses: shegen@cosy.sbg.ac.at (S.Hegenbart), uhl@cosy.sbg.ac.at (A. Uhl), andreas.vecsei@stanna.at (A. Vécsei), gwimmer@cosy.sbg.ac.at (G. Wimmer)

- Several approaches have been developed to achieve scale (and often orientation) invariance for multi-scale and multi-orientation wavelet transforms. These techniques are mostly applicable to any multi-scale and multi-orientation transform. We employ the Dual-Tree Complex Wavelet Transform (DT-CWT) (Selesnick et al., 2005) instead of the originally proposed transforms and are able to show that our approach works better for the target celiac disease database than other wavelet-type transforms (like e.g. Gabor filters (Fung and Lam, 2009) or steerable pyramids (Montoya-Zegarra et al., 2007) (see Table 3). An additional benefit is the improved ability to compare the different strategies to achieve scale invariance if the underlying transform is the same in all cases.

- We propose a new affine invariant method based on Local Ternary Patterns (LTP).

- We conduct explicit experimental tests for scale invariance for all feature descriptors considered based on the Columbia-Utrecht (CUReT) (Dana et al., 1999) dataset and the Celiac Disease Scale (CDS) database (see Section 6.2.2) following ideas in Varma and Zisserman (2009), revealing that claimed scale invariance cannot be verified for many of the schemes investigated.

- Most approaches are tested for their ability of invariant texture analysis on public databases like Brodatz (Brodatz, 1966), CUReT (Dana et al., 1999), KTH-TIPS (Hayman et al., 2004), or the UIUCTex (S. Lazebnik and Ponce, 2005) database. Correspondingly, most of the considered methods have been optimized for the corresponding datasets. Hence we have adjusted some of these methods (e.g. using different parameters or replacing parts of the original algorithm) to make them applicable in a sensible manner for the classification of celiac disease (e.g., use of different measures in techniques based on fractal analysis in Section 4 or application of a different clustering strategy for the dense Scale Invariant Feature Transform (SIFT) features in Section 5).

- We show, that methods extracting highly contrast sensitive information work well for the classification of celiac disease, specifically methods based on fractal analysis.

This paper is organized as follows. In Section 2 we briefly introduce the concept of computer-assisted diagnosis of celiac disease by automated classification of duodenal mucosa texture patches and review the corresponding state-of-the-art. In Section 3, we describe strategies to achieve scale invariance for wavelet transforms including the application of the discrete cosine transform (DCT) or the discrete Fourier transform (DFT) to the feature vectors of the wavelet transforms (Häfner et al., 2010; Lo et al., 2004), re-arrangement of feature vectors (cyclic shifting, dominant scale, and slide matching) (Montoya-Zegarra et al., 2007; Lo et al., 2009; Fung and Lam, 2009), or methods that preprocess the image before the wavelet transform is being applied (Lee, 2003). Section 4 describes techniques based on fractal analysis while Section 5 covers a heterogeneous set of additional approaches to generate scale invariant texture descriptors (e.g. neural nets (Ma et al., 2010; Zhan et al., 2009), SIFT features and region detectors (Fei-Fei and Perona, 2005; Zhang et al., 2006), and multiscale blob features (Xu and Chen, 2006)) as well as a new affine invariant method we propose which is based on scale-normalized Laplacian maxima combined with Local Ternary Patterns (Hegenbart and Uhl, 2013). Experimental results with respect to classification of the celiac disease dataset and with respect to effective scale invariance (by means of the CDS database and parts of the CUReT database) are presented in Section 6. Section 7 concludes our work.

2. Computer-Assisted Diagnosis of Celiac Disease

Celiac disease is a complex autoimmune disorder in genetically predisposed individuals of all age groups after introduction of gluten containing food. 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 continuously been 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. This is why early diagnosis is of highest importance. 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 (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 for detection of total or partial villous atrophy and other specific markers that show a high specificity for celiac disease in adult patients like scalloping of the small bowel folds, reduction in the number or loss of Kerkring's folds, scalloped folds, mosaic patterns, and visualization of the underlying blood vessels (Niveloni et al., 1998).

Automated classification as a support tool is an emerging option for endoscopic treatments (e.g. (Liedlgruber and Uhl, 2011a,b)). Systems are being developed that support physicians during surgery or highlight malignant areas during an endoscopy for further inspection. Such systems could also be used for training purposes. In the context of celiac disease, an automated system identifying areas affected by celiac disease in the duodenum 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 (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.

The celiac state of the duodenum is usually determined by visual inspection during the endoscopic session followed by a biopsy of suspicious areas. During endoscopy at least four duodenal biopsies are taken. The severity of the mucosal state of the extracted tissue can be histologically staged according to a modified Marsh scheme (Oberhuber et al., 1999) which is based on Marsh (1992).According to this staging scheme, we have divided available duodenal image material into four different classes, Marsh-0 Marsh-3a, Marsh-3b and Marsh-3c (see Figure 2).

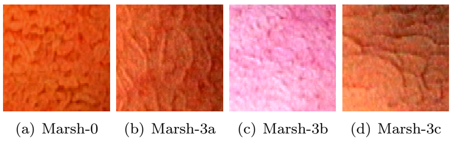

(a) Marsh-0 (b) Marsh-3a (c) Marsh-3b (d) Marsh-3c

Figure 2: Example images for the respective classes

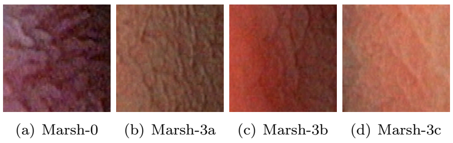

(a) Marsh-0 (b) Marsh-3a (c) Marsh-3b (d) Marsh-3c

Figure 3: Images with different perspective and scale

Marsh-0 represents a healthy duodenum with normal crypts and villi, Marsh-3a, Marsh-3b and Marsh 3c have increased crypts and mild atrophy (3a), marked atrophy (3b) or the villi are entirely absent (3c), respectively. Types Marsh-3a to Marsh-3c span the range of characteristic changes caused by celiac disease, where Marsh-3a is the mildest and Marsh-3c is the most severe form. We also consider the 2-class case, where we only differentiate between healthy (Marsh-0) and unhealthy (Marsh-3a, Marsh-3b and Marsh 3c) mucosal types, respectively.

As described in Section 1 endoscopic image material of mucosal texture is usually found at different perspective and scale (see Figure 3 for examples from our database). Therefore, the employment of scale invariant feature description techniques is a highly intuitive idea for a computer-assisted diagnosis system.

Prior approaches dealing with the computer-aided diagnosis of celiac disease using endoscopic imagery do exist but they do not focus on scale invariance. With respect to feature descriptors in previous papers, we have investigated several variants of Local Binary Pattern (LBP) based operators (Vécsei et al., 2011; Hegenbart et al., 2011), band-pass type Fourier filters (Vecsei et al., 2009), as well as histogram and wavelet-transform based features (Uhl et al., 2011a; Vécsei et al., 2008). We have also systematically compared the classification performance of two different image capturing techniques and various preprocessing schemes using a set of different feature extraction and classification methods (Hegenbart et al., 2009). Smoothness/sharpness measures have been used as features in Ciaccio et al. (2011). Techniques involving temporal information computed from videocapsule endoscopy have been described recently (Ciaccio et al., 2010b,a).

3. Scale Invariant Wavelet based Features

In this section we describe scale invariant texture descriptors, that are based on multi-scale and multi-orientation

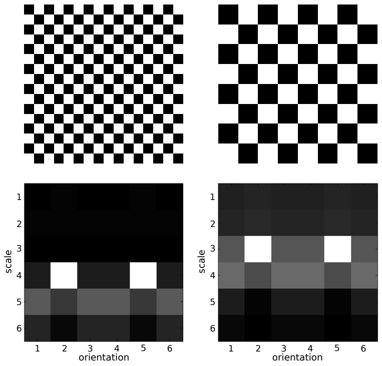


Figure 4: Cyclic shifting of the means of the subbands across the scale dimension

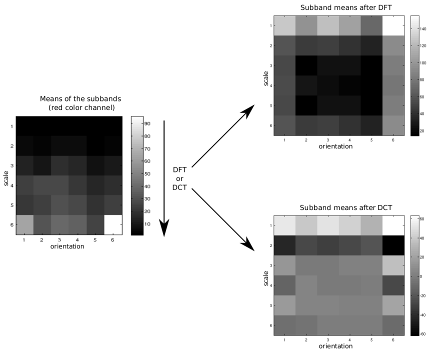


Figure 5: Computing the discrete cosine transform (DCT) or discrete Fourier transform (DFT) across the scale dimension of the subband means

transforms like the discrete wavelet transform, the Gabor wavelet transform and the dual-tree complex wavelet transform (DT-CWT). Various wavelet-based feature extraction methods have been proposed for endoscopic image analysis (since approximately 2003) (e.g. Kwitt et al. (2009); Barbosa et al. (2008, 2009); Iakovidis et al. (2004)). The subbands of these methods contain information about different scales and orientations of an image. The strategies to make these transforms invariant to scale change are to transform or reorder the corresponding transform coefficients or to find a different representation for the images before applying the respective transforms. The underlying principles how to achieve scale invariance are similar for the approaches in this section (except for the approach that re-arranges the image before the transformation). If an image is scaled, then the subbands of the scaled image are shifted across the scale dimension compared to the subbands of the unscaled image. In Figure 4 we see two checkerboard patterns in the first row, where the right pattern is a scaled version of the left one with a scale factor of two. In the second row of Figure 4, the corresponding subband means (when using DT-CWT) of the checkerboard patterns are shown. We can see that the subband means of the scaled checkerboard pattern (the right one) are shifted one scale level up compared to the subband means of the unscaled checkerboard pattern.

Most strategies used to achieve scale invariance of the methods described are applicable to any multi-scale and multi-orientation transform method. We propose to apply these different strategies to the DT-CWT as opposed to the various (wavelet) transforms published in the original papers. The advantages of this approach are as follows: First, the DT-CWT provides better results for the classification of celiac disease than any other type of multi scale transform (as we will see in Table 3). Second, the different strategies to achieve scale invariance are easier to compare if the underlying transform is the same in all cases. In fact, the results for classifying celiac disease are better for all strategies if we use the DT-CWT instead of the originally proposed (wavelet) transforms. One possible reasons for that is the shift invariance of the DT-CWT. Shift invariance is important for the classification of celiac disease, since the representation of features of an image by wavelet coefficients should not be dependent on the position of the features in the image. Another possible reason is the high redundancy of the DT-CWT (it is using two separate Discrete Wavelet Transforms (DWT) and thus has double the redundancy of the DWT), which provides extra information for the analysis.

Kingbury's DT-CWT (Selesnick et al., 2005) divides an image into six directional (15°, 45°, 75°, 105°, 135°, 165°) oriented subbands per level of decomposition. The DT-CWT analyzes an image only at dyadic scales. For some of the strategies proposed in this section, a finer scale resolution is required. The double dyadic dual-tree complex wavelet transform (D^{3}T-CWT) (Lo et al., 2009) overcomes this issue by introducing additional levels between dyadic scales. These additional levels are generated by recursively applying the DT-CWT to a downscaled version of the original image (using a factor of $2^{-0.5}$ in downscaling). For each subband we calculate the statistical features' mean (μ) and standard deviation (σ) from the absolute values of the subband coefficients. We denote a statistical feature of a subband $S_{l,d}$ with scale level $l \in \{1, \ldots, L\}$ and orientation $d \in \{1, \ldots, D\}$ as $q_{l,d}$. The feature vector of an image is composed of these statistical features collected from the different subbands.

3.1. Applying Discrete Fourier Transform and Discrete Cosine Transform to DT-CWT

Features that are approximately scale invariant can be generated by applying the discrete Fourier transform (DFT) across the scale dimension of a feature vector of the the D^{3}T-CWT (Häfner et al., 2010) (see Figure 5):

$$Q_{n,d} = \frac{1}{\sqrt{L}} \sum_{l=1}^{L} q_{l,d}\, e^{\frac{-i\, 2\pi(l-1)(n-1)}{L}},$$

with $n \in \{1,\ldots,L\}\,, d \in \{1,\ldots,D\}$.
The vector $\mathrm{fv}_{SI} = \{|Q_{1,1}|,\ldots|Q_{L,1}|, |Q_{1,2}|,\ldots|Q_{L,2}|, \ldots|Q_{L,D}|\}$ provides a texture feature that is nearly invariant to scale. The feature curve of a feature vector shifts if input texture is scaled (see Figure 4). If a feature curve $q^m_{l,d}$ is a cyclic shifted version of the old one ($q^m_{l \pmod{L+1},d} = q_{l+m \pmod{L+1},d}$, $m \in \{1,\ldots,L\}$), then applying DFT to the feature curves followed by taking the magnitude of it provides the same results for the old and new feature curve ($|Q_{n,d}| = |Q^m_{n,d}|$, where $Q^m_{n,d}$ is defined like $Q_{n,d}$, but with using $q^m_{l,d}$ instead of $q_{l,d}$). The reason for that follows from the Shift Theorem of the DFT: $Q^m_{n,d} = Q_{n,d}\, e^{\frac{2\pi i(n-1)m}{L}}$ (with $|e^{\frac{2\pi i(n-1)m}{L}}| = 1$). The problem is, that the Shift Theorem is only valid if the input signal $q_{l,d}$ is periodic, but it is questionable why these statistical features should be periodic. However if the statistical features are close to zero at both ends, the approach provides good scale invariance. In Figure 5 we can see the means of the subband coefficients from the red color channel of an image of the celiac disease database (we separately apply the DFT to the features of the three color channels of the RGB color space). We can see that the coarse end ($l = 6$) has the highest means, which are absolutely not close to zero. This of course questions input periodicity.

Another possibility is to consider only the real part of the DFT, which is a cosine transform. This leads us to the application of the Discrete Cosine Transform (DCT) across scale dimension (see Häfner et al. (2010)). The DCT is not invariant to cyclic shifts of the feature curve and so it is not theoretically clear if the DCT enhances the scale invariance of the DT-CWT in general. Even if the DCT would be invariant to cyclic shifts, this would not enhance scale invariance since the input signal $q_{l,d}$ is not periodic. Results in Häfner et al. (2010) indicate that the DCT enhances the scale invariance at least for small differences of scales (maximum scale factor ≈ 1.4), tests for bigger scale differences were not made.

A related approach (Lo et al., 2004) is to resize each D^{3}T-CWT subband to the size of the original image. In this way we get a local feature vector for each pixel consisting of the subband coefficients (absolute values) at the position of the pixel. Like in the approach before, the DFT is applied across the scale dimension of these local feature vectors (see Figure 5), but this time to each local feature vector instead of the statistical features of the subbands:

$$Q^{local}_{n,d}(x,y) = \frac{1}{\sqrt{L}} \sum_{l=1}^{L} S_{l,d}(x,y)\, e^{\frac{-i\, 2\pi(l-1)(n-1)}{L}},$$

with $n \in \{1,\ldots,L\}\,, d \in \{1,\ldots,D\}$. For each "transformed subband" $Q^{local}_{n,d}$ we compute the statistical features mean and standard deviation. Since the operations for achieving scale invariance are applied to the local subband coefficients (instead to the global statistical subband features mean and standard deviation like in the approaches before) we denote this method as "D^{3}T-CWT with DFT (local)". In extending Lo et al. (2004) we additionally use the DCT instead of the DFT and denote this method as "D^{3}T-CWT with DCT (local)". A feature vector of DT-CWT or DT-CWT with DCT has a length of 216 (6 orientations × 6 scale levels × 3 color channels × 2 statistical features per subband). In case of DT-CWT with DFT, for each direction d, the following features form complex conjugates: $Q_{2,d} = Q^*_{6,d}$ and $Q_{3,d} = Q^*_{5,d}$. That means 2 of the 6 scale levels (scale levels 5 and 6) are redundant, which reduces the length of the feature vector to 144 elements. If using D^{3}T-CWT instead of DT-CWT, the feature vector length is doubled.

3.2. Cyclic Shifting of Local Features

Instead of computing the DFT across the scale dimension of the local feature vectors of the D^{3}T-CWT, these vectors are cyclically shifted across the scale dimension (in the original approach (Lo et al., 2009), the local feature vectors are additionally shifted across the orientation dimension to achieve orientation invariance, but since we are primarily interested in scale invariance we omit that). First we square each element of the local feature vectors and apply subsequently a circular-correlation (only across the scale dimension) of the squared feature vector with a specific mask M (see Figure 6).

The result of this process is a correlation vector, which is as long as the number of scale levels is. Now the original feature vector is cyclically shifted in the scale dimension, so that the first scale level of the new local feature vector is the scale level of the original local feature vector, in which the correlation vector had its maximum (see Figure 6). Then the subbands (consisting of the corresponding feature values) are modeled by a Rayleigh distribution (Lo et al., 2009) and the parameters of this distribution are used to form the final feature vector of an image. Since we use only one statistical feature per subband, the number of features per image is half the number of features using the D^{3}T-CWT (216).

3.3. The Dominant Scale Approach

The accumulated energies of the scales $l \in \{1\ldots,L\}$ are computed across the orientations $d \in \{1,\ldots,D\}$ (Montoya-Zegarra et al., 2007):

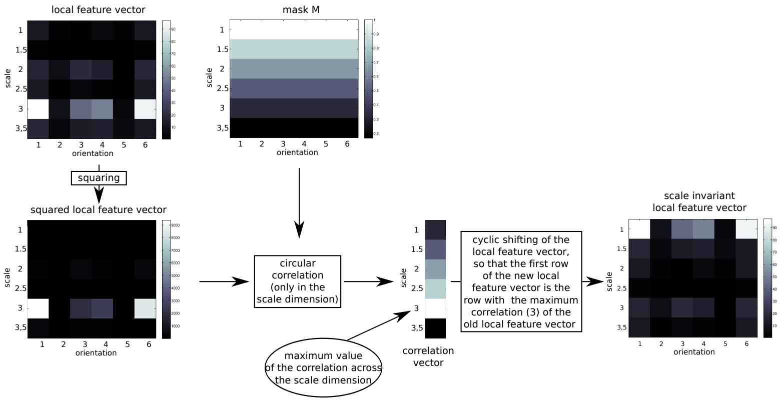

local feature vector
mask M
squaring
squared local feature vector
circular correlation (only in the scale dimension)
maximum value of the correlation across the scale dimension
correlation vector
cyclic shifting of the local feature vector, so that the first row of the new local feature vector is the row with the maximum correlation (3) of the old local feature vector
scale invariant local feature vector
scale
orientation
1
1.5
2
2.5
3
3,5

Figure 6: Cyclic shifting of local feature vectors

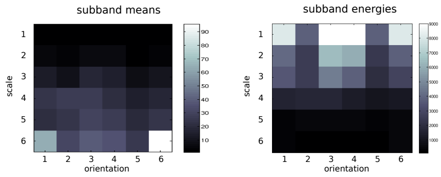


Figure 7: Comparison of the subband means and energies

$$E(l) = \sum_d \sum_x \sum_y |S_{l,d}(x, y)|.$$

A scale invariant representation is achieved by computing the dominant scale (DS) of the images followed by feature alignment. The dominant scale is defined as the scale with the highest accumulated energy $E(l)$. Now the feature vector, consisting of mean and standard deviation of the subbands, is circularly shifted, such that the features of the dominant scale are the first ones in the feature vector.

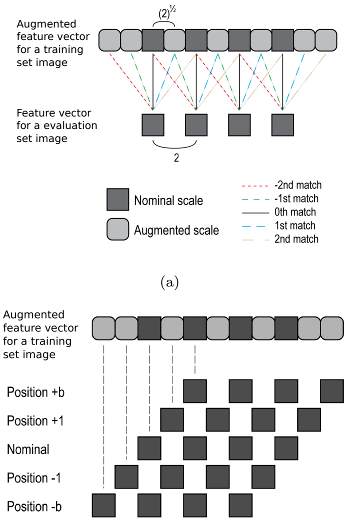
We face a problem when applying this method to our database (which is identical for the original approach (Montoya-Zegarra et al., 2007) using steerable pyramid decomposition and for our version using the DT-CWT). Due to subsampling, subbands at increasing scale have a lower number of coefficients. On the other hand, the absolute values of coefficients in higher scales are distinctively higher than those in lower levels (see the subband means in Figure 7). Nevertheless, the dominant scale will almost ever occur at scale level $l = 1$ due to the high number of coefficients.

In fact, when using the DT-CWT on our data set, the DS is always at scale level $l = 1$, and for the steerable pyramid decomposition the DS is not at scale level $l = 1$ for 17 images only (out of 612 images). That is why we adapted the dominant scale approach by using subband means instead of the subband energies. Using this approach, for 38 images the DS is not at scale level $l = 6$ which improves the situation only slightly. Feature vector values are almost always monotonically increasing with the scale level (subband means) or monotonically decreasing with the scale level (energy of the subbands) and therefore shifting the features across the scale dimension according to the DS does not make a big difference. The original approach of Montoya-Zegarra et al. (2007) also determines the dominant orientation, but as we are more interested in scale invariance we omit this process (results have been deteriorated when using this approach). The length of the feature vector is equal to that of the DT-CWT (216).

3.4. The Slide Matching Approach

Slide matching (Fung and Lam, 2009) was originally proposed for the Gabor transform but is used with the D^{3}T-CWT in our context. The original approach is first made orientation invariant by summing up the means and standard deviations of the subbands' coefficients with same scale level. In adapting the proposed approach to our scenario (Fung and Lam, 2009), we compute the scale levels $1, 1.5, 2, \ldots, 6$ of the training set and the scale levels $2, 3, 4, 5$ of the evaluation set. The distance between an image of the training set and an image of the evaluation set is the distance that is minimized by sliding the feature vectors along the scale dimension against each other (see Figure 8).

Since we are primarily interested in scale invariance we also use a modified version of slide matching without summing up the subbands of the same scale level. Consequently, we have for each scale level 12 (6 orientations, 2 parameters per subband) instead of 2 features for the slide matching process. Therefore in case of the original version the length of the feature vector is equal to that of


(2)^½
Augmented feature vector for a training set image
Feature vector for a evaluation set image
2
Nominal scale
Augmented scale
-2nd match
-1st match
0th match
1st match
2nd match
Augmented feature vector for a training set image
Position +b
Position +1
Nominal
Position -1
Position -b

(a)

(b)

Figure 8: (a) Different scale factors are used for the training set images and for the evaluation set images. Each node denotes two elements, a sum of means and a sum of standard deviations. (b) The sliding of evaluation set image feature vector along augmented training set image vector

the DT-CWT (216) and in case of the modified version, the length of the feature vector is only a sixth of that of the DT-CWT (36).

3.5. The Log-Polar Approach

The log-polar transformation maps points from the Cartesian plane (x, y) to points in the log-polar plane (ξ, η) (see Figure 9). In this coordinate system, scaling and rotation is converted to translations.

Scale invariance (and orientation invariance) can be achieved by analyzing the transformed image with a shift invariant transform like the adaptive row shift invariant wavelet packet transform (as originally proposed in Lee (2003)) or the DT-CWT used in this work. Unlike in Lee (2003), we compute subband means and standard deviations as features (instead of energies) and do not use the best basis algorithm. Obviously, the length of the feature vector is equal to that of the DT-CWT (216).

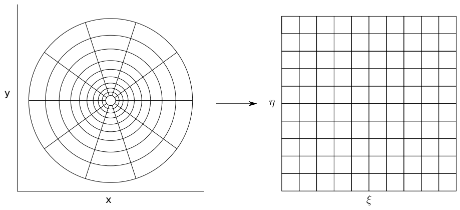
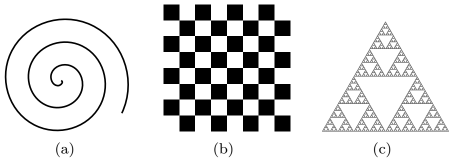


Figure 9: The log-polar transformation



(a)

(b)

(c)

Figure 10: Fractal dimension D in 2D space. (a) Smooth spiral curve with $D = 1$, (b) the checkerboard with $D = 2$ and (c) the Sierpinski-Triangle with $D \approx 1.6$

4. Scale Invariant Methods based on Fractal Analysis

For a point set E defined on $\mathbb{R}^2$, the fractal dimension of E is defined as

$$dim(E) = \lim_{\delta \to 0} \frac{\log N(\delta, E)}{-\log \delta},$$

where $N(\delta, E)$ is the smallest number of sets with diameter less than δ that cover E. The set is made up of closed disks of radius δ or squares of side length δ. In Figure 10 we present some examples for the fractal dimensions of different objects.

Intuitively, the fractal dimension is a statistical quantity that gives a global description of how complex, how irregular or how rough a geometric object is. However, the fractal dimension alone, as defined before, does not provide a rich description. It is just a single value.

The local fractal dimension or also called the local density function, used in two of the three methods presented in this section, provides a more powerful and adaptive description. Let μ be a finite Borel regular measure on $\mathbb{R}^2$. For $x \in \mathbb{R}^2$, denote $B(x, r)$ as the closed disk with center x and radius $r > 0$. $\mu(B(x, r))$ is considered as an exponential function of r, i.e. $\mu(B(x, r)) = cr^{D(x)}$, where $D(x)$ is the density function and c is some constant. The local density function (or also called local fractal dimension) of x is defined as

$$D(x) = \lim_{r \to 0} \frac{\log \mu(B(x, r))}{\log r}.$$

The density function measures the “non-uniformness” of the intensity distribution in the region neighboring the considered point.

The local density D is invariant under the bi-Lipschitz map, which includes view-point changes and non-rigid deformations of texture surface as well as local affine illumination changes. A bi-Lipschitz function f must be invertible and satisfy the constraint $c_1||x-y|| \leq ||f(x)-f(y)|| \leq$

$c_2||x-y||$ for $c_2 \geq c_1 > 0$. Consequently, local fractal dimensional based approaches are especially interesting for developing scale-invariant feature descriptors, and so also for the classification of celiac disease.

By choosing different measures μ, the local density function can be adapted to different image processing tasks. As we will see, measures based on derivative information work best for our dataset, since their contrast sensitiveness is a good feature to differentiate between images with or without celiac disease. The reason for that is that images of patients with celiac disease have less or entirely no villi and therefore a lower amount of contrast compared to images of patients without celiac disease.

4.1. The Multi-Fractal Spectrum

First, the local fractal dimension is computed for each pixel of an image (Xu et al., 2009b). Let E_α be the set of all image points x with local density in the interval α:

$$E_\alpha = \{x \in \mathbb{R}^2 : \, D(x) \in \alpha\}.$$

Usually this set is irregular and has a fractal dimension $f(\alpha) = dim(E_\alpha)$.

We denote the convolution $*$ between an image $I = I(x,y)$ and a Gaussian kernel $G_\sigma = G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{x^2+y^2}{2\sigma^2}}$ (i.e. Gaussian blur) as follows:

$$I(x,y,\sigma) = I(\sigma) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(x+u, y+v) G(u,v,\sigma) dudv$$

and $I_x(\sigma)$ is the first derivative of $I(\sigma)$ in the direction of x.

In the original approach (Xu et al., 2009b) three different types of measures $\mu(B(x,r))$ are defined for the computation of the local density:

$$\begin{aligned} \mu(B(x,r)) = & \quad \textstyle\int_{B(x,r)} I(\sigma) \, dx \\ \mu(B(x,r)) = & \quad \textstyle\int_{B(x,r)} \sum_{k=1}^{4} (f_k * (I(\sigma)^2)^{\frac{1}{2}} \, dx \\ \mu(B(x,r)) = & \quad \textstyle\int_{B(x,r)} |(I_{xx}(\sigma) + I_{yy}(\sigma))| \, dx, \end{aligned} \tag{1}$$

where $\{f_k, \; k = 1,2,3,4\}$ are four directional operators (derivatives) along the vertical, horizontal, diagonal, and anti-diagonal directions. The feature vector of an image I consists of the concatenation of the fractal dimensions $f(\alpha_i)$ for the three different measures $\mu(B(x,r))$.

In case of our dataset, it turned out that the Laplacian measure (equation (1)) is the only one of the three measures which leads to sensible results. The reason for that is very probably the comparatively highest contrast sensitiveness of the Laplacian measure. So contrasting to the original proposal (Xu et al., 2009b), the employed feature vector only has entries of the third type. We use 14 non-overlapping intervals α and so the length of the feature vector per image is 14.

4.2. Fractal Analysis using Filter Banks

Instead of partitioning the local densities of images in sets $E(\alpha)$'s and computing their fractal dimensions, we first convolve the images with the MR8 filter bank (Varma and Zissermann, 2005; Geusebroek et al., 2003), a rotationally invariant, nonlinear filterbank with 38 filters but only 8 filter responses, and compute local fractal dimension afterwards (Varma and Garg, 2007; Uhl et al., 2011b). Filters can smooth over image noise and lead to more robust features. However, they also have the drawback of lowering the level of bi-Lipschitz invariance.

Let us introduce the measures

$$\mu(B(x,r)_i) = \quad \int \textstyle\int_{B(x,r)} |f(i)| \; dx \tag{2}$$

$$\mu(B(x,r)_i) = \quad \int \textstyle\int_{B(x,r)} |(S_x + S_y) * (G_\sigma * f(i))| \, dx, \tag{3}$$

where $f(i)$ is the i-th MR8 filter response image with $1 \leq i \leq 8$. $S_x = [-1,0,1;-2,0,2;-1,0,-1]/4$ and $S_y = -S(x)^T$ are Sobel filters. The first measure (Equation 2) is the measure originally proposed in Varma and Garg (2007), while the second measure (Equation 3) is proposed in Uhl et al. (2011b), where the original fractal method of Varma et al. (Varma and Garg, 2007) has been optimized for the celiac disease database. We follow the optimized approach using the second measure and computing the local density for each of the 8 filter responses (in Varma and Garg (2007), only 5 of the 8 filter responses are used, in our experiments the results are better using all 8 responses). For each pixel of an image, we result in an 8-dimensional local density vector. For each class of the training set we aggregate the local density vectors of the images of this class and learn cluster centers (called textons) by k-means clustering.

The next step is to learn models for each image of the training and evaluation sets. Given an image, its corresponding model is generated by first convolving it with the filter bank, computing the local density of each filter response and then labeling each local density vector with the texton that lies closest to it. Distances between two frequency histograms (models) are measured using the χ^2 statistic. The length of the feature vector per image is the number of classes of the according image database multiplied by 10 (10 clusters per class).

4.3. Fractal Dimensions for Orientation Histograms

Similar to SIFT features (see next section), this method (Xu et al., 2009a) is based on computing local orientation histograms. First the gradient magnitude and the orientation of a given pixels neighborhood are computed. The orientation histogram from the neighborhood of the given pixel is formed by discretization of orientations by weighing the gradient magnitude (see Figure 11). The histogram is then assigned to one of 29 orientation histogram templates, which are constructed based on the spatial structure of the orientation histogram (the number of significant image gradient orientations and their relative positions).

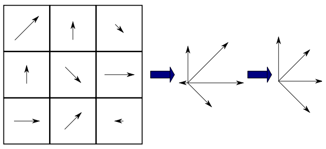


Figure 11: The process of construction and discretization of the orientation histogram when using a neighborhood of size 3 × 3

We now have for each pixel (for a given neighborhood size) a value between 1 and 29, depending on the template it is assigned to. By setting a pixel to one if it is assigned to template i ($i \in \{1, \ldots, 29\}$) and to zero otherwise, 29 binary images are generated, from which we compute the fractal dimensions (by means of the box-counting method [1]). This process is applied for eight different neighborhood sizes (scale levels). In order to get better robustness to scale changes, finally a wavelet transform (a redundant tight wavelet frame system) is applied across the scale dimension (the different neighborhood sizes) of the fractal dimensions. The final feature vector of an image consists of the detail and approximation coefficients of the wavelet transform. This feature vector can be viewed as the information about the changes with respect to scale (the neighborhood sizes represent the scale levels). According to (Xu et al., 2009a), this enhances the scale invariance, since the scale changes are often consistent across multiple scales for natural textures. The length of the feature vector per image is 1160 (29 orientation histogram templates × 8 neighborhood sizes × 5 (2 different high-pass filters each with 2 decomposition levels and a low-pass filter using only the second decomposition level).

5. Further Approaches

In this section we will present approaches that are neither based on wavelet transforms nor on fractal analysis. The first two approaches are based on the widely used SIFT features (Lowe, 1999) and affine invariant region detectors (Zhang et al., 2006), two approaches work with neural networks (Ma et al., 2010; Zhan et al., 2009) and one approach analyzes characteristics of connected regions (blobs) (Xu and Chen, 2006). Finally we cover the affine invariant Local Ternary Patterns which are based on the analysis of multi-scale second moment matrices in a Laplacian scale space (Hegenbart and Uhl, 2013).

5.1. SIFT Features and Region Detectors

The Scale Invariant Feature Transform (SIFT) (Lowe, 1999) is probably the most popular feature used in computer vision (Vedaldi and Fulkerson, 2008). SIFT detects salient image regions (keypoints) and extracts discriminative yet compact descriptors of their appearance. SIFT keypoints are invariant to viewpoint changes like translation, rotation, and rescaling of an image.

First an image is convolved with Gaussian filters at different scales σ. By means of detecting the maxima/minima of the Difference of Gaussians (DoG), local scale space extrema are found. The DoG is given by

$$DoG(x, y, \sigma) = I(k\sigma) - I(\sigma) \ .$$

Local scale space extrema which have low contrast or are poorly localized are eliminated, the rest are used as keypoints. Using the Gaussian filtered image (with keypoint scale σ), gradient magnitudes and orientations are computed from the neighboring region of the keypoint to form an orientation histogram. Now the gradient information is rotated according to the dominant orientation of the orientation histogram and weighted by a Gaussian function. A local descriptor uses 16 histograms, aligned in a 4 × 4 grid, each with 8 orientation bins. This results in a feature vector containing 128 (16*8) elements for each keypoint of an image.

The original approach Lowe (1999) is suited for object recognition, in this work however we are interested in texture classifications, SIFT keypoints do not make sense in our context. We apply two different ways to deal with that problem:

1. We use dense SIFT features (Fei-Fei and Perona, 2005), that means that we compute SIFT descriptors for each pixel of an image.
2. We use a region detector that is suited for texture images and then apply the SIFT descriptor to the detected regions (Zhang et al., 2006).

A region detector suited for texture recognition is the Harris detector (S. Lazebnik and Ponce, 2005; Mikolajczyk and Cordelia, 2004). The Harris detector is based on the second moment matrix M_I. This matrix must be adapted to scale changes to make it independent of the image resolution:

$$M_I(\sigma, \gamma) = G_\sigma * \begin{pmatrix} I_x^2(\gamma) & I_x(\gamma)I_y(\gamma) \\ I_x(\gamma)I_y(\gamma) & I_y^2(\gamma) \end{pmatrix},$$

The Harris corner measure μ is defined as follows:

$$\mu(\sigma, \gamma) = \det(M_I(\sigma, \gamma)) - \alpha \ \text{trace}^2(M_I(\sigma, \gamma)).$$

Note that μ simultaneously lives in two scale spaces (caused by the Gaussian kernel G) with parameters γ and σ. The inner scale γ, which is less critical than the outer scale σ, is set to a constant value. Local maxima of this measure determine the location of Harris interest points, and then the Laplacian scale selection procedure is applied at these locations to find their characteristic (outer) scale σ. The Laplacian scale selection finds the characteristic scale at a given point (x, y) by maximizing the Laplacian-of-Gaussian:

$$L(x, y; \sigma) = \sigma^2 |I_{xx}(\sigma) + I_{yy}(\sigma)|. \qquad (4)$$

[1] rsbweb.nih.gov/ij/plugins/fraclac/FLHelp/BoxCounting.htm

The elliptic region around a found location is described by its principal axes corresponding to the eigenvectors of M_I and axis length depending on the eigenvalues. For affine invariance, a region is normalized by mapping it onto a unit circle and using a rotational invariant descriptor, the SIFT descriptor.

So for both ways, using dense SIFT features or using the Harris detector (combined with the affine invariant mapping and the SIFT descriptor), we get features from the SIFT descriptor as output. For both approaches we follow the strategy applied in Section 4.2, as opposed to the classical dense SIFT approach (Fei-Fei and Perona, 2005). For each class of the training set we aggregate the SIFT descriptors of the images of this class and learn cluster centers (textons) by k-means clustering. Given an image, its corresponding model is generated by labeling its SIFT descriptors with the texton that lies closest to it. Distances between two frequency histograms (models) are measured using the χ^2 statistic. For both approaches, the length of the feature vector per image is the number of classes of the according image database multiplied by 10 (10 clusters per class).

It should be noted that instead of using the Harris detector it would be possible to use other region detectors (e.g. Laplacian (Zhang et al., 2006) and Hessian region detectors (Mikolajczyk and Schmid, 2002)) and descriptors (e.g. SPIN and RIFT features (Zhang et al., 2006)), the principle of the approach however remains the same. Following the terminology in the original papers, we denote the approach using the dense SIFT features as “Dense SIFT Features” and the approach using the Harris detector as “Local Affine Regions”.

5.2. Pulse-Coupled Neural Networks based Methods

Pulse-coupled neural networks (PCNN’s) (Ranganath et al., 1995) are neural models proposed by modeling a cat’s visual cortex. PCNN is a neural network algorithm that produces a series of binary pulse images when stimulated with an image. The intersecting cortical model (ICM) (Ma et al., 2010) and the spiking cortical model (SCM) (Zhan et al., 2009) are two methods derived from the PCNN, which are faster and provide better or similar results as compared to the PCNN (Ma et al., 2010; Zhan et al., 2009).

The ICM model consists of two coupled oscillators, a small number of connections and a non-linear function. F is the state oscillator and Θ the threshold oscillator. Together they constitute the neurons pulse sequence Y. The mathematical model of ICM is described as follows:

$$\begin{aligned} F_{ij}(n) = &\quad fF_{ij}(n-1) + I(i,j) + \\ &\quad \sum_{kl} M_{ijkl} Y_{kl}(n-1) \\ \Theta_{ij}(n) = &\quad g\Theta_{ij}(n-1) + hY_{ij}(n-1) \\ Y_{ij}(n) = &\quad \begin{cases} 1 \text{ for } F_{ij}(n) > \Theta_{ij}(n) \\ 0 \text{ otherwise.} \end{cases} \end{aligned}$$

where f, g and h are scalars, $M = [0.5, 1, 0.5; 1, 0, 1; 0.5, 1, 0.5]$ is the connection function through which the neurons communicate, I is the input image and $n \in \{1, \ldots, N\}$. The pair (i, j) stands for the position of the neuron in the map and (k, l) is that of its neighboring neurons. The outputs of ICM are N binary images, which represent features like texture, edges, and segments.

The mathematical model of the SCM is described as follows:

$$\begin{aligned} F_{ij}(n) = &\quad fF_{ij}(n-1) + I(i,j) + \\ &\quad I(i,j) \sum_{kl} M_{ijkl} Y_{kl}(n-1) \\ \Theta_{ij}(n) = &\quad g\Theta_{ij}(n-1) + hY_{ij}(n-1) \\ Y_{ij}(n) = &\quad \begin{cases} 1 \text{ for } F_{ij}(n) > \Theta_{ij}(n) \\ 0 \text{ otherwise.} \end{cases} \end{aligned}$$

As for ICM, the outputs of SCM are N binary images. The final feature vectors of the SCM and ICM, respectively, consist of the entropies of the $N = 37$ binary output images.

The authors (Ma et al., 2010; Zhan et al., 2009) state that their approaches (ICM and SCM) are scale invariant (and rotation and translation invariant), however their manuscripts miss a valid justification for this statement. They reference a further publication (Johnson, 1994) in which scale invariance is explained. The problem is that there a special kind of PCNN is considered and that scale invariance is only shown for objects on a uniform background, not for textures.

5.3. Multiscale Blob Features

In order to derive multiscale blob features (Xu and Chen, 2006), we apply a series of flexible threshold planes to a textured image and then use the topological and geometrical attributes of the generated blobs in the obtained binary images to describe image texture. The flexible threshold planes FP are determined by Gaussian blurring:

$$FP(x, y; \sigma, b) = b + I(x, y, \sigma) \ .$$

where σ^2 is the variance to control the spread of the window and b is the bias. By applying the flexible threshold planes to the grayscale image I, we obtain binary images

$$g_b(x, y; \sigma) = \begin{cases} 1 & \text{if } I(x, y) > FP(x, y; \sigma, b) \\ 0 & \text{otherwise} \end{cases}$$

In each binary image, all 1-valued pixels and 0-valued pixels are grouped into two sets of connected regions called blobs (see Figures 12,13).

The original approach uses two features to describe an image, the number of blobs and the shapes of the blobs. The shape feature indicates how compact the blobs of an

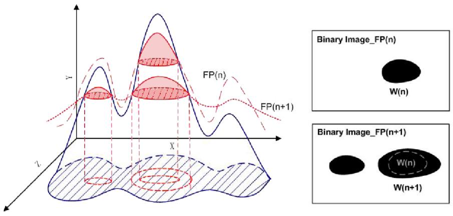
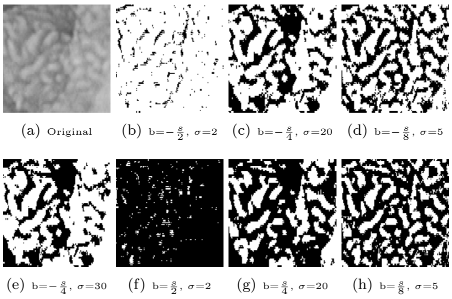


Figure 12: Process of extracting binary blobs



(a) Original (b) b$=-\frac{s}{2}$, σ=2 (c) b$=-\frac{s}{4}$, σ=20 (d) b$=-\frac{s}{8}$, σ=5

(e) b$=-\frac{s}{4}$, σ=30 (f) b$=\frac{s}{2}$, σ=2 (g) b$=\frac{s}{4}$, σ=20 (h) b$=\frac{s}{8}$, σ=5

Figure 13: Binary blob images with different sigmas and biases denoted by b, where s denotes the standard deviation of the original image. 0-valued pixels are displayed as black pixels and 1-valued pixels are displayed as white ones

image are (the compactness of a blob is here defined as the maximum of the distances from the pixels of a blob to the centroid of the blob, divided by the square root of the number of pixels of the blob). The multiscale blob features are invariant to rotation and gray-level scaling (the bias b of $FP(x, y; \sigma, b)$ is depending on the standard deviation of the input image). The shape features are invariant to spatial scaling within a small range (the compactness is similar for spatial scaling within a small range), but the number of blobs change to some extent. As opposed to the original approach (Xu and Chen, 2006), we separately classify the images for the two blob features, since the shape feature is scale invariant and the number of blobs is not. The length of the final feature vector is 480 (30 values for σ × 8 values for b × 2 (black and white regions of a binary image)).

5.4. *Affine Invariant Local Ternary Patterns*

The Local Binary Pattern approach as introduced by Ojala et al. (1996) as well as the Local Ternary Patterns method proposed by Tan and Triggs (2007) are not affine invariant. An extension to the method suggested by Mäenpää (2003) uses multiple Gaussian filters with varying sizes to improve the support area of the operator. This extension adds multi resolution to the operator but misses a scale selection mechanism. We propose an affine invariant method based on Local Ternary Patterns that employs scale-normalized derivatives of local scale space maxima for scale selection. We compute the multi-scale second moment matrices at given scales to analyze textures according to their affine shape along an ellipse. The method shares the idea of using the scale space framework with methods such as the SIFT feature detector and other region detectors as discussed in Section 5.1. The idea of combining scale space maxima with Local Binary Patterns has also been explored by Li et al. (2012).

Instead of using the DoG (difference of Gaussian) approximation to the Laplacian of Gaussians as used by SIFT we construct the scale space by computing the scale-normalized Laplacian (see Equation 4) of each image I at each location $x \in \mathbb{N}^2$ at different scales with $\sigma = \frac{1.5}{\sqrt{2}}^k, k \in \{1, \ldots, 20\}$ denoted as $(\overline{\triangle} I(x; \sigma))$. The initial scale is chosen such that it corresponds to the standard radius of LBP (1.5).

Due to the fact that not all locations in an image attain a local maximum and a maximum between scales, we compute a scale mask to improve the reliability of the scale estimation. This is especially useful when textures are not strictly periodic and attain multiple scales as is the case in celiac disease. The computation involves the detection of local maxima at each scale. We exploit the fact that pixels in close spatial proximity to a maximum are at the same or a relatively close scale to the detection scale of the corresponding maximum.

We compute the multi-scale second moment matrices at each location x of an image I which is attaining a local scale space maximum. We use the detection scale of the maximum as the local scale t, the integration scale $s = \sqrt{2}t$ is depending on the detection scale.

$$\mu(x; t, s) = \int_{\xi \in \mathbb{R}^2} (\nabla I)(x - \xi; t)(\nabla I)^T(x - \xi; t)\, g(\xi; s)\, d\xi.$$

With $(\nabla I)(x; t)$ denoting the gradient of the scale space orientation at scale t and g denoting a Gaussian function. The second moment matrix summarizes the gradient distribution of the area around a pixel location at a given scale. The eigenvalues of the matrix characterize the length of the axes of an ellipse (up to some constant multiplier) while the eigenvectors describe the orientation of the axes. Due to the fact that the orientation of the ellipse described by the second moment matrix is normal to the detected blob we compute the inverse of the second moment matrix. The inverse results in a rotation by ninety degrees without modifying the ratios of the axis lengths.

The absolute sizes of the axes given by the second moment matrices are unknown, we therefore normalize the ellipses such that the circumference is equal to the circumference of the detected maxima treated as circle (the radius at scale σ is $\sqrt{2}\sigma$). To do so, we apply Ramanujan's formula for approximating the circumference of the ellipse (with axes a and b) and solve the quadratic equation for a constant scaling factor c

$$\sqrt{2}\sigma 2\pi = \pi\left(3(ac+bc) - \sqrt{(3ac+bc)(ac+3bc)}\right).$$

The axes of the ellipse are then re-scaled by the appropriate solution of c.

After the computation of the support area of a local maximum (the ellipse given by the second moment matrix at the position and scale of the maximum), all locations within this area are assigned to the scale and response of that maximum (we call this a corresponding maximum for a location). In our methodology it is possible that a single location contains multiple possible scales and responses, this is due to the fact that the support areas of multiple maxima might overlap.

We observed that the reliability of a location attaining the same scale as a corresponding maximum decreases with spatial distance. To compensate for this, we compute the reliability of a scale at a distance d from a corresponding maximum using a Gaussian probability density function choosing σ_p such that the reliability of a given scale at a distance of the length of the semi minor axis b of the corresponding maximum's normalized support area is 0.5. We additionally use the responses of all corresponding maxima for a pixel location to ensure that maxima with lower responses have lower reliabilities.

$$r(x;l) = \frac{\overline{\triangle} I(x;l)}{\max\limits_{t} \overline{\triangle} I(x;t)} e^{\frac{-d^2}{2\sigma_p^2}} \quad \text{and} \quad \sigma_p = \left(\frac{-(\frac{b}{2})^2}{2\,\log(0.5)}\right)^{\frac{1}{2}}$$

The reliability measure is finally used to assign a weight controlling the contribution of each computed pattern to the histogram . Once the scales of each location have been determined we apply an adaptive Gaussian filter to the data, prior to computing the LTP code at a location. We select the width of the Gaussian filter such that the area covered by the operator in relation to the local scale is the same across all scales. This gives invariance to uniform scaling. The width of the Gaussian filter (f_w) is selected in a way that 90 percent of the area of the Gaussian function are in the area of the computed filter.

$$f_w = \frac{\sqrt{2}\sigma 2\pi}{n} \quad \text{and} \quad \sigma_{\text{Gauss}} = \frac{f_w\,0.5}{\sqrt{2}\,\text{erf}^{-1}(0.9)}$$

For n being the number of considered LTP neighbors, σ being the scale at the location.

To compute a pattern at a location we estimate the second moment matrices using the detection scales of all corresponding maxima at that location. We distribute the sample points of the operator such that they lie along the normalized ellipses described by the second moment matrices. By using this approach non uniform scaling of the data can be compensated, because this type of transformation would change the shape of the ellipses accordingly. We then distribute sample points so that the distance in terms of arc length between adjacent points is equal, giving n-equidistant points along the ellipse. To speed up the computation we define four support points on the ellipse which lie on the ends of the major and minor axes respectively. The definition of support points limits the method to distribute a number of $4N+4$ equidistant points along the ellipse but reduces the computation to N points. We use the fact that all ellipses can be described as a scaled and rotated version of a canonical ellipse. To distribute the points on a canonical ellipse in parametric form, the positions of N points in the first quadrant are computed and symmetries are exploited to gain the other $3N$ points. To find the offset on the x-axis of the n-th point (Δx_n) from the center of the ellipse the equation

$$\frac{n}{N+1}\int_0^a \sqrt{1+\left(\frac{dy}{dx}\right)^2}\,dx \;=\; \int_0^{\Delta x_n} \sqrt{1+\left(\frac{dy}{dx}\right)^2}\,dx$$

is solved for Δx, where a is the length of the horizontal semi-axis, N is the number of points to distribute per quadrant and the second additive term is the derivative of the canonical implicit equation of an ellipse.

The definition of support points also provides the possibility of defining a fixed starting point for the computation of the patterns. Due to the ambiguous orientation of an ellipse we define two starting points (computing two patterns per position) to compensate. These are by definition the points on the intersection of the major axis with the ellipse.In case of ellipses that are close to a circle this definition becomes unreliable, we therefore treat second moment matrices with a ratio of eigenvalues $\frac{\lambda_{\min}}{\lambda_{\max}} >= 0.95$ as a circle treating the vertical axis as the major axis. By defining a starting point we are able to compensate rotations, this is due to the fact that this kind of affine transformation is reflected by the orientation of the computed ellipses. Please see Hegenbart and Uhl (2013) for a more thorough explanation of the method. The feature vector of an image consists of a single histogram with 59 bins.

6. Experimental Results

We use the software provided by the Robotic Research Group [2] for region detection (Harris detector) and description (SIFT) in Section 5.1, the VLFEAT implementation (Vedaldi and Fulkerson, 2008) for the dense SIFT features in Section 5.1 and the implementation of Geusebroek et al. (2003) for the MR8 filter in Section 4.2. For the remaining algorithms custom implementations from earlier work (Kwitt and Uhl, 2007; Uhl et al., 2011b) (DT-CWT, Fractal Analysis using Filter Banks) or specifically developed for this work have been used (all using Matlab except for the affine invariant LTP method which we developed using Java).

[2] http://www.robots.ox.ac.uk/ vgg/research/affine

The original manuscripts employ a wide variety of different classifiers. Since higher developed classifiers (e.g. the SVM classifier) will mostly produce better results than more simple classifiers (e.g. the k-NN classifier) and since the focus lies on scale invariant feature extraction strategies and not on classification methods, all methods are classified using the k-NN classifier. The advantage of that approach is the better comparability of the results with respect to feature expressiveness.

Classification accuracy is computed using an evaluation set and a training set. An image from the evaluation set is classified into the class, where most of the k nearest neighbors from the training set belong to. The k for the k-NN classifier, used to classify the evaluation set, is optimized on the training set (the k with the highest overall classification rate (OCR) using leave–one–out cross–validation (LOOCV) on the training set).

The algorithms using k-means clustering provide different results each run. For these methods we provide average results from 10 runs per method.

6.1. Celiac Disease

We use a database of duodenal endoscopic images employed in earlier work (Hegenbart et al., 2011) to enable easier comparison. Table 1 lists the number of image samples and patients per class. To avoid overfitting and to test the methods in in a practice-related context, the images of a patient are either all in the evaluation set or all in the training set. In this way it is impossible that the nearest neighbors of an image and the image itself come from the same patient. This is important to avoid any bias in the result, the setup of this data set resembles in a way how LOPO (Leave–one–patient–out) and LOOCV (Leave–one–out cross–validation) work.

The original endoscopic images (which are of size 620 × 530 or 520 × 510 depending on the used endoscope) often exhibit only small areas that permit a distinction between healthy mucosa and mucosa affected by celiac disease. This is due to the facts, that endoscopic images in general show a high amount of distortions such as bubbles, specular reflections and occlusions due to the geometric properties of the duodenum. Additionally, the distribution of villous atrophy caused by the disease could be restricted to certain areas within the visible area (this is known as patchy distribution of celiac disease). Therefore we extract non overlapping patches of size 128 × 128 (under supervision of a physician). Our celiac disease database consists of these patches.

We observed that the overall classification rate (OCR) varies significantly depending on the chosen number of nearest neighbors of the k-NN classifier. Therefore, we use a second measure for the 2-class case to evaluate the methods, the area under the ROC (receiver operating characteristic) curve (AUC) (Bradley, 1997). We generate the ROC curve by considering the class membership of the 20 nearest neighbors for each image of the evaluation set, where the area under the ROC curve is calculated by trapezoidal integration (Bradley, 1997). The AUC uses the information how many of the 20 nearest neighbors of each evaluation set image are positive (celiac disease) or negative (healthy), whereas the OCR only uses the information if more or less of the k nearest neighbors are positive than negative.

Data set	Training set				
Marsh type	0	3a	3b	3c	Total
Number of images	155	50	56	51	312
Number of patients	66	6	7	8	87
Data set	Evaluation set				
Marsh type	0	3a	3b	3c	Total
Number of images	151	45	58	46	300
Number of patients	65	5	6	8	84

Table 1: Number of image samples per Marsh type (ground truth based on histology)

The results in Table 2 are sorted according to the OCR results of the 2-class case. In the last four rows of the table we display methods not designed to be scale invariant, DT-CWT, D^{3}T-CWT and two earlier results using the same database (Hegenbart et al., 2011). One method is the original LBP approach, the other one, denoted "WT–LBP", is the best performing approach for this dataset so far. (Except for the approach 'Fractal Analysis using Filter Banks" , which is proposed in Uhl et al. (2011b)) WT–LBP is a combination of Local Binary Patterns and the discrete wavelet transform (for details see Hegenbart et al. (2011)).

The two fractal methods using the local density function perform best for our celiac disease database, especially "Fractal Analysis using Filter Banks" works very good. Also the second fractal method ("Multi-Fractal Spectrum") performs reasonably well. However, the AUC of the latter fractal method is not high compared to other methods. This is because this method has the highest OCRs when we consider many nearest neighbors ($30 \leq k \leq 70$ in the kNN classifier), while the AUC only uses information of the 20 nearest neighbors (all other methods have their highest OCRs for ks between one and thirty). The third method using fractal features, "Fractal Dimensions for Orientation Histograms", also provides useful results. Overall, the considered fractal methods are quite well suited for classifying celiac disease.

When we consider the results of different strategies for achieving scale invariance using the DT-CWT or the D^{3}T-CWT, we see that DCT computed across the scale dimension of the statistical subband features (DT-CWT and D^{3}T-CWT with DCT) can clearly enhance the results compared to the the DT-CWT or the D^{3}T-CWT without any further feature manipulation. All other modifications of the DT-CWT or D^{3}T-CWT decrease the results. The results of the methods, where operations for achieving scale invariance are applied to the local subband

Method	2-class case		4-class case
	OCR	AUC	OCR
Fractal Analysis using Filter Banks	91.7	95.0	65.8
Multi-Fractal Spectrum	89.0	90.5	62.0
D^3T-CWT with DCT	88.3	92.9	63.0
Multiscale Blob Features (number)	86.3	89.9	57.7
DT-CWT with DCT	86.0	92.7	63.0
Affine Invariant LTP	85.6	92.4	61.3
Fractal Dim. for Orientation Histograms	84.0	90.6	62.7
Dense SIFT Features	83.6	87.5	62.0
D^3T-CWT with DCT (local)	82.3	89.3	60.0
Cyclic shifting of Local Features	81.0	88.4	61.7
Log-Polar Approach	80.0	86.9	57.0
Dominant Scale Approach	78.3	87.5	56.7
D^3T-CWT with DFT (local)	78.3	86.4	55.7
Slide matching (original)	76.3	81.7	57.3
Slide matching (modified)	74.7	85.5	62.3
Local Affine Regions	70.9	88.1	56.3
Multiscale Blob Features (shape)	70.8	76.2	54.3
ICM	67.7	71.7	52.3
D^3T-CWT with DFT	66.0	70.2	50.0
SCM	64.0	66.4	51.3
DT-CWT	84.7	90.2	60.3
D^3T-CWT	82.3	90.1	58.0
WT-LBP	88.0	–	63.7
LBP	84.0	–	61.4

Table 2: Results of the different methods in OCR (%) and AUC. In the 4–class case we only present the OCR

coefficients of the D^3T-CWT ("D^3T-CWT with DCT (local)", "D^3T-CWT with DFT (local)", and "Cyclic Shifting of Local Features"), are in the middle of the results range and give pretty similar OCR. The results of the methods, where operations for achieving scale invariance are applied to the global statistical subband features (e.g. mean and standard deviation) of the D^3T-CWT, differ a lot. Some are better than their local counterparts ("DT-CWT and D^3T-CWT with DCT"), some are worse ("Slide Matching" and "D^3T-CWT with DFT"), while the others ("Log-Polar" and "Dominant Scale Approach") give comparable OCR.

"Multiscale Blob Features" using the scale dependent number of blobs as feature works well whereas using the scale invariant shape of the blobs as feature did not provide useful rates for classifying celiac disease. This result, together with the well performing DT-CWT and D^3T-CWT techniques without any technique for further scale invariance being applied, questions the importance of scale invariance in general for our dataset. "Dense SIFT Features", which do not use any keypoint selection, provides a clearly better result compared to "Local Affine Regions" using a keypoint selection strategy specifically tuned for textured data. "Affine Invariant LTP" shares the same idea of key point detection with "SIFT" and "Local Affine Regions" , the computed scales mask however increases the reliability in case of non periodic textures. Overall it provides a better performance as compared to these two methods.

The results of the neural nets approaches ("ICM" and "SCM") and the two slide matching variants are not competitive at all.

If we compare results of the 4-class case to the 2-class case, then we see that the differences among the results are smaller in case of the 4-class case. The ranking among the different approaches is similar to the 2-class case. Overall, the OCRs in the 4-class case are not suited for any application scenario and do not improve over earlier work.

Having applied DT-CWT and D^3T-CWT instead of the originally proposed transforms for several techniques, we shed light on the reason for this decision. In Table 3 we show classification results of the wavelet based methods, which originally did not use CWTs. We compare the OCRs (2–class case) of the methods using the originally proposed wavelet transforms with the results using DT-CWT variants instead of the original transforms.

As we can see in Table 3, using CWTs works distinctly better for classifying celiac disease as compared to the originally proposed transforms.

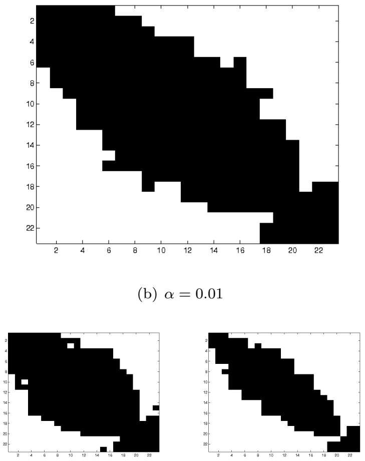
Finally, we want to assess statistical significance of our results. The aim is to analyze if the images from the celiac disease database are classified differently by the various methods considered or if all techniques fail for the same set of images. We use the McNemar test (McNemar, 1947), to test if two methods are significantly different for a given

Methods	original	CWT
Log-Polar Approach	58.0	80.0
Dominant Scale Approach	74.3	78.3
Slide matching Approach	74.0	76.3

Table 3: Results of the wavelet based methods in % (2–class case). The column "original" shows the results using the originally proposed wavelet transforms, the column "CWT" shows the results using CWTs instead of the original transforms

1.) Fractal A. u. Filter Banks
2.) Multi Fractal Spectrum
3.) D3t-CWT with DCT (global)
4.) M. Blob Feat. (number)
5.) DT-CWT with DCT (global)
6.) Affine Invariant LTP
7.) DT-CWT (global)
8.) Fractal Dim. f. O. H.
9.) Dense Sift Features
10.)D3T-CWT (global)
11.)D3T-CWT with DCT (local)
12.)Cyclic shifting of Local F.
13.)Log-Polar Approach
14.)Dominant Scale Approach
15.)D3T-CWT with DFT (local)
16.)Slide matching (original)
17.)Slide matching (modified)
18.)Local Affine Regions
19.)M. Blob Feat. (shape)
20.)ICM
21.)D3T-CWT with DFT (global)
22.)SCM

(a) Methods



(b) $\alpha = 0.01$

(c) $\alpha = 0.001$

(d) $\alpha = 0.05$

Figure 14: Results of the McNemar test for the 2–class case. A white square in the i'th row and j'th column or in the j'th row and i'th column of a plot means that the i'th and the j'th method are significantly different with significance level α. If the square is black then there is no significant difference between the methods. The methods are sorted beginning with the best ones (OCR 2-class case) like in Table 2

level of significance (α) by building test statistics from incorrectly classified images. Tests were carried out for the 2-class case with three different levels of significance ($\alpha = 0.05$, $\alpha = 0.01$ and $\alpha = 0.001$). Results are displayed in Figure 14 (the methods are sorted according to the OCR results of the 2-class case), where we can observe, that methods with similar OCRs are never found to be significantly different. Figure 14 shows that only methods with clearly different OCR results are rated as significantly different. That indicates that for methods with similar OCR results, almost the same images are classified wrong, independent of the extracted features.

6.2. Testing Scale Invariance Explicitly

We have employed a set of methods, explicitly introduced to provide scale invariance, motivated by the observation that our celiac disease database contains features at various scales. We want to investigate if these methods are really as scale invariant as they theoretically should be. Further, we want to assess if the techniques' scale invariance really enhances the results for detecting celiac disease, or if the obtained results depend primarily on the general feature extraction ability, independent of scale invariance properties.

The training and the evaluation sets of the celiac disease database both contain images with features at different scales (as well as various orientations, brightnesses and viewpoints). Since each class in the training or evaluation sets has at least 45 images, for almost every image in the evaluation set there might exist images of the same class in the training set with rather similar scales. That means, that a technique does not necessarily have to be scale invariant to work well on our dataset. Therefore, for assessing the scale invariance of a method it is not adequate to test if the method works well for a database containing images with various scales. We need to use two databases, where one database contains differently scaled images as compared to the other.

A further problem of testing scale invariance of the employed approaches with the celiac disease database is that we do not have the information which actual spatial scale an image belongs to (i.e. the distance and perspective of the camera to the mucosal wall) therefore it is difficult to separate the database into two disjoint sets depending on the scale of the images. Another possibility to get two data sets with different scales would be to synthetically scale the database, but this changes the characteristics of the images too much (e.g. interpolation effects, eventual contrast changes, etc. ...).

We solved this problem by extracting patches from frames of endoscopy videos instead of extracting them from endoscopic images (like done for the celiac disease database). Since it is possible to choose any (suitable) frame of a video from an endoscopy session, it is easier to find patches with a specific distance to the mucosal wall as compared to choosing the patches of some images taken during the endoscopic session. Additionally it is easier to estimate distances in a video than estimating them by means of single images.

As a second texture database to verify the scale invariance of the employed methods, we use parts of the CUReT database (Dana et al., 1999). The advantage of this database compared to our celiac disease scale database is that we have the exact information to which scale an images actual belongs to and that images of one texture class are gathered under exactly the same scale conditions.

That is why we decided to test scale invariance by using two different databases, the celiac disease scale database and parts of the CUReT database.

6.2.1. CUReT Database

The cropped version of the CUReT database [3] contains 92 images per texture with different viewing and illumination conditions. There are four texture classes from the CUReT database (material numbers 2, 11, 12, and 14), for which additional scaled data is available (as material numbers 29, 30, 31, 32). The scale difference between these two sets is approximately 1.7. These materials are shown in Figure 15. The material classes are evenly divided into one part for the evaluation set and one part for the training set, where the images of 46 viewpoint and illumination conditions are used for the training set and the images of the remaining 46 viewpoint and illumination conditions are used for the evaluation set.

For explicitly testing scale invariance, two experiments are performed following ideas in Varma and Zisserman (2009). In the first experiment (E1), the training set consists of original textures (4 × 46 images of material numbers 2,11,12, and 14, each with 46 different viewpoint and illumination conditions), while the evaluation set consists of original textures and scaled versions of the original textures (8 × 46 images, images of material numbers 2,11,12,14, 29, 30, 31 and 32 with the remaining 46 viewpoint and illumination conditions). For this experiment, scale invariance is obviously crucial since half of the evaluation set consists of data scaled differently than the data in the training set. In the second experiment (E2), the evaluation set is like in the first experiment, but this time the training set consists of original textures as well as scaled versions of the original textures. The lower the difference between the classification results of the first and the second experiment, the higher the scale invariance of a method is.

The classification results are shown in Table 5.

6.2.2. Celiac Disease Scale Database

The celiac disease scale (CDS) database consists of patches extracted from endoscopy videos. To determine the scale invariance of the employed approaches, we divided the patches into the two categories “Regular“ and “Far“, depending of the distance to the mucosa wall. Images of the category “Regular“ have optimal distances to differentiate between ”healthy” and ”affected” tissue. Because of the larger distances, the differentiation between the two classes is harder for images of the category “Far”. The assessment of distance was performed manually based on the visibility of features (there is no ground truth about the actual distance of the endoscope to the mucosal wall).

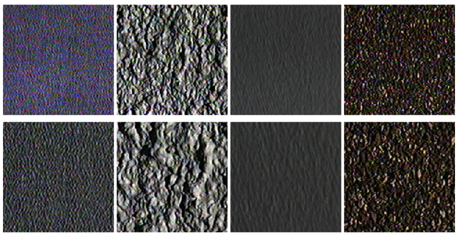


Figure 15: The top row shows one image each from material numbers 2, 11, 12, and 14 from the CUReT database, while the bottom row shows the textures with higher zoom factor (as material numbers 29, 30, 31, and 32)

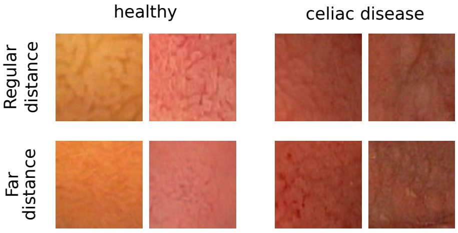


Figure 16: Example images of the CDS database gathered from regular and further distances

We only used images of sequences showing the same mucosal area at a regular distance as well as at a further distance (like done in Hegenbart et al. (2012)). That means for each extracted image of regular distance, we extracted exactly one image with further distance (and vice versa). Similar to the CUReT database, the CDS database consists of a training set (named training set ”Regular-Far”), consisting of images with far and regular distances, a second training set (training set “Regular“) consisting of images with only regular distances (the images of training set *Regular-Far* with regular distance), and an evaluation set consisting of images with regular and far distances (evaluation set “Regular-Far“) (see Figure 16).

In parallel to the celiac disease database, the images of the training sets are gathered from different patients as of these contained in the evaluation set. Table 4 lists the number of image samples and patients per class.

For explicitly testing scale invariance we perform two experiments. In the first experiment, we use training set *Regular-Far* and evaluation set *Regular-Far*. In the second experiment, we use training set *Regular* and evaluation set *Regular-Far*. Similar to the CUReT database, scale invariance is only needed for the second experiment and not for the first, since only in the second experiment the training and evaluation set are gathered under different

[3] www.robots.ox.ac.uk/vgg/research/texclass/data/curetcol.zip

Data set	Training set *Regular-Far*		
Class	healthy	celiac disease	Total
Number of images	40	40	80
Number of patients	20	12	32
Data set	Training set *Regular*		
Class	healthy	celiac disease	Total
Number of images	20	20	40
Number of patients	20	12	32
Data set	Evaluation set *Regular-Far*		
Class	healthy	celiac disease	Total
Number of images	38	38	76
Number of patients	19	10	29

Table 4: Number of image samples and patients of the CDS database

scale conditions. The Classification results are shown in Table 5. The lower the difference between the classification results of the first (E3) and the second experiment (E4), the higher is the scale invariance of a method.

6.2.3. Results of Testing the Scale Invariance

The presented results in Table 5 are the mean values of the results using a k-NN classifier with $k = 1 - 20$. In that way we balance the problem of varying results depending on the number of nearest neighbors of the k-NN classifier. The lower the difference between the classification results of experiment 1 (E1) and experiment 2 (E2) respectively experiment 3 (E3) and experiment 4 (E4), the higher is the scale invariance of a method.

The methods showing the highest degree of scale invariance for a database are marked with a “+“, the ones showing the least degree of scale invariance are marked with a “−“, and the methods showing average scale invariance are marked with a ”◦”. The ones that are hard to interpret, because they even do not work without scale changes (E1 or E3), are marked with a ”?“.

Results shown in Table 5 are quite unexpected, especially the ones of the CUReT database.

The absolute OCR results of the first (E1 respectively E3) and second experiments 2 (E2 respectively E4) are not relevant for us, but the differences between them, indicating the extent of scale invariance, are very interesting.

In case of the CDS database, some results are hard to interpret (the two slide matching approaches, D^{3}T-CWT with DFT and SCM), because even the results without scale changes (E3) are pretty near to the results of randomly classifying images (50%).

In case of the CUReT database, the three methods using fractal analysis are rated as not scale invariant (except of ”Fractal Analysis using Filter Banks“, which is rated as average). In case of the CDS database all three methods are rated as scale invariant. So the ratings with respect to scale invariance of the CUReT database are contrary to those of the CDS database.

When we consider the methods based on DT-CWT or D^{3}T-CWT, we also see a clear difference between the two databases. In case of the CUReT database the original (not scale invariant) approaches (DT-CWT and D^{3}T-CWT) are more scale invariant than their variations (except of the Dominant Scale Approach). In case of the CDS database, the methods applying a transformation to the local wavelet features or shifting them across the scale dimension (D^{3}T-CWT with DCT (local), D^{3}T-CWT with DFT (local) and Cyclic Shifting of Local Features) provide more scale invariance than the original approaches. The methods which apply the transformation to global wavelet features or shift them across the scale dimension are either average scale invariant (DT-CWT with DCT and D^{3}T-CWT with DCT), not scale invariant (Dominant Scale Approach) or they are hard to interpret, because already the results without scale change (E3) are rather low (the two slide matching approaches and D^{3}T-CWT with DFT). The Log-Polar-Approach turned out to be scale invariant for both databases.

The methods ICM and SCM are rated as average scale invariant or unratable for both databases.

The Multiscale Blob Features are rated as quite scale invariant (the shapes of the blobs) or as average scale invariant (the number of blobs) in case of the CUReT database (corresponding to to the theoretical considerations). But in case of the CDS database they are both rated as not scale invariant (especially when using the shape of the blobs).

The Dense SIFT Features are for both databases more scale invariant as compared to the Local Affine Regions. In case of the CUReT database both methods are rated as average scale invariant, in case of the CDS database the Dense SIFT are rated as quite scale invariant and the Local Affine Regions are rated as not scale invariant.

The Affine Invariant LTP is rated as quite scale invariant for both databases.

Overall, the results of the two databases with respect to the scale invariance are quite different. Of course the two databases for testing the scale invariance are very different.

The intra-class variability of the CUReT database is significantly smaller than those of the CDS database. The visual distinction between images with or without celiac disease is quite different, since there are many images that don't look like typical representatives of their class or they even look like belonging to the other class. In case of the CUReT database the visual distinction between the classes is quite easy. Another difference between the two databases is, that the images of the CUReT database are much more homogeneous than those of the CDS database (an image of the CUReT database looks similar at different positions of the image, that is usually not the case for images of the CDS database). One additional problem is, that one of the most important feature to

Method	CUReT				CDS			
	E1	E2	Diff	SI	E3	E4	Diff	SI
Fractal Analysis using Filter Banks	91.1	85.8	5.8	○	71.8	70.5	1.8	+
Multi-Fractal Spectrum	91.4	77.0	15.8	−	69.1	71.1	-2.8	+
D^3T-CWT with DCT	98.3	88.7	9.8	○	76.1	73.8	3.0	○
Multiscale Blob Features (number)	97.2	89.6	7.8	○	65.9	59.9	9.1	−
DT-CWT with DCT	98.4	87.1	11.5	○	75.3	70.9	5.8	○
Affine Invariant LTP	99.0	95.7	3.3	+	74.4	75.8	-1.9	+
Fractal Dim. for Orientation Histograms	86.9	74.1	14.7	−	71.7	70.6	1.5	+
Dense SIFT Features	71.5	67.4	5.7	○	68.1	66.7	2.1	+
D^3T-CWT with DCT (local)	97.7	92.9	4.9	○	72.2	71.1	1.5	+
Cyclic Shifting of Local Features	98.8	95.1	3.7	+	73.3	72.8	0.7	+
Log-Polar Approach	90.7	88.3	2.6	+	72.6	73.4	-1.1	+
Dominant Scale Approach	92.7	93.9	-1.3	+	72.4	64.8	11.7	−
D^3T-CWT with DFT (local)	96.3	89.7	6.9	○	72.9	73.0	-0.1	+
Slide Matching (original)	93.5	81.4	12.9	○	58.8	54.3	7.6	?
Slide Matching (modified)	97.6	75.3	22.8	−	63.2	60.5	4.3	?
Local Affine Regions	96.1	89.8	6.6	○	67.6	59.0	12.7	−
Multiscale Blob Features (shape)	96.9	93.9	3.1	+	72.0	62.0	16.1	−
ICM	90.2	81.4	9.8	○	64.8	59.8	7.7	○
D^3T-CWT with DFT	95.9	89.2	7.0	○	63.7	69.1	-8.5	?
SCM	97.9	92.6	5.4	○	60.3	56.5	6.3	?
DT-CWT	99.2	97.0	2.2	+	72.4	68.4	5.5	○
D^3T-CWT	99.1	96.8	2.3	+	73.0	69.7	4.5	○

Table 5: OCR results for the CDS and CUReT database.The columns “E1“ and “E3“ show the results of the experiments using same scale levels in training and evaluation set, and the columns “E2“ and “E4“ show the results of the experiments using different scale levels in training and evaluation set. The columns ”Diff” show the relative differences between the results of E1 and E2 respectively E3 and E4. The column “SI“ rates the scale invariance of the methods as high (+), low (−), average (○) or unratable (?)

differentiate between the two classes of the CDS database, the villi, are often less visible at bigger distances of the endoscope to the mucosal wall. This complicates the differentiation between images showing healthy mucosa from further distances to those showing celiac disease affected mucosa from closer distances Hegenbart et al. (2012).

Because of these big differences between the two databases, there are features proving to be more scale invariant for one database than for the other. The scale invariance of the extracted features of a method vary with the application of the method.

7. Conclusion

It seems that especially contrast sensitive methods work very well for the celiac disease database, specifically the fractal methods. The “Multi-Fractal Spectrum” is originally using a combination of three different measures ($\mu(B(x,r))$), but we only use the Laplacian measure, which is the most contrast sensitive of the three. The second fractal method, “Fractal Analysis using Filter Banks” behaves similarly. Other contrast sensitive methods are the third fractal method “Fractal Dimensions for Orientation Histograms” and the method “Multiscale Blob Features (number)”, both methods performed well for our celiac disease database. The affine invariant LTP method performed comparably to the best methods.

When we consider the methods using the DT-CWTs, we see that many of the techniques designed to be scale invariant perform worse than the original CWTs without any specific tuning, except for the methods applying DCT across global subband descriptors, for which it is not even theoretically clear why they should enhance scale invariance of the DT-CWT.

Our results indicate that scale invariance is not important for the classification of celiac disease, at least when considering our dataset to be representative. There is no positive correlation between the performance of the methods (in terms of OCR) and their (determined) scale invariance.

There is a big difference between theoretical concepts for scale invariance and practical scale invariance actually achieved in experiments. It also turned out that the practical scale invariance of a method is not fixed, it depends on the application the method is used for. The determined scale invariance of the methods using the CUReT database (application texture recognition) is quite different to the determined scale invariance using the CDS database (application endoscopic image classification).

In case of endoscopic image classification, it turned out that the methods which have not been designed to be scale invariant are nearly as scale invariant than those explicitly designed to be scale invariant. The affine invariant LTP method exhibited the highest degree of scale invariance.

The behavior of methods is interesting in the case of texture recognition. Our results indicate that scale variant methods turn out to be more effective than their scale invariant counterparts. This is quite surprising, since these methods were especially designed to be scale invariant for texture recognition tasks. From this point of view we have to state that techniques claimed to be scale invariant should be actually tested for this property in properly designed experiments. It contradicts good scientific practice to state properties which do not hold in actual applications.

8. Acknowledgments

This work is partially supported by the Austrian Science Fund, TRP Project 206, and the Austrian National Bank Jubiläumsfonds Project 12991.

References

Barbosa, D.J.C., Ramos, J., , Correia, J.H., Lima, C.S., 2009. Automatic detection of small bowel tumors in capsule endoscopy based on color curvelet covariance statistical texture descriptors, in: Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009 (EMBC'09), Minneapolis, Minnesota, USA. pp. 6683–6686.

Barbosa, D.J.C., Ramos, J., Lima, C.S., 2008. Detection of small bowel tumors in capsule endoscopy frames using texture analysis based on the discrete wavelet transform, in: Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008 (EMBS'08), Vancouver, British Columbia, Canada. pp. 3012–3015.

Bradley, A.P., 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition 30, 1145–1159.

Brodatz, P., 1966. Textures: A Photographic Album for Artists and Designers. Dover Publications, New York.

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

Cammarota, G., Cuoco, L., Cesaro, P., et al., 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, 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, 732–738.

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

Ciaccio, E.J., Bhagat, G., Tennyson, C.A., Lewis, S.K., Hernandez, L., Green, P.H., 2011. Quantitative assessment of endoscopic images for degree of villous atrophy in celiac disease. Digestive Disease and Science 56, 805–811.

Ciaccio, E.J., Tennyson, C.A., Lewis, S.K., Bhagat, G., Green, P.H., 2010a. Classification of videocapsule endoscopy image patterns: comparative analysis between patients with celiac disease and normal individuals. BioMedical Engineering Online 9.

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

Dana, K., Van-Ginneken, B., Nayar, S., Koenderink, J., 1999. Reflectance and Texture of Real World Surfaces. ACM Transactions on Graphics (TOG) 18, 1–34.

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., 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.

Fei-Fei, L., Perona, P., 2005. A bayesian hierarchical model for learning natural scene categories, in: Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society, pp. 524 – 531.

Fung, K.K., Lam, K.M., 2009. Rotation- and scale-invariant texture classification using slide matching of the gabor feature, in: Proceedings of Intelligent Signal Processing and Communication Systems, pp. 521–524.

Geusebroek, J.M., Smeulders, A.W.M., van de Weijer, J., 2003. Fast anisotropic gauss filtering. IEEE Transactions on Image Processing 12, 938–943.

Häfner, A., Uhl, A., Vécsei, A., Wimmer, G., Wrba, F., 2010. Complex wavelet transform variants and scale invariance in magnification-endoscopy image classification, in: Proceedings of the 10th International Conference on Information Technology and Applications in Biomedicine (ITAB'10), Corfu, Greece.

Hayman, E., Caputo, B., Fritz, M., Eklundh, J.O., 2004. On the significance of real-world conditions for material classification, in: Proceedings of the European Conference on Computer Vision, pp. 253–266.

Hegenbart, S., Kwitt, R., Liedlgruber, M., Uhl, A., Vecsei, A., 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.

Hegenbart, S., Uhl, A., 2013. An Affine Invariant Texture Descriptor based on Local Ternary Patterns. Technical Report 2013-01. Department of Computer Sciences, University of Salzburg, Austria. http://www.cosy.sbg.ac.at/research/tr.html.

Hegenbart, S., Uhl, A., Vécsei, A., 2011. Systematic assessment of performance prediction techniques in medical image classification - a case study on celiac disease, in: Proceedings of the 22nd International Conference on Information Processing in Medical Imaging (IPMI'11), Monastery Irsee, Germany. pp. 498–508.

Hegenbart, S., Uhl, A., Vécsei, A., 2012. On the implicit handling of varying distances and gastrointestinal regions in endoscopic video sequences with indication for celiac disease, in: Proceedings of the IEEE International Symposium on Computer-Based Medical Systems (CBMS'12).

Iakovidis, D., Maroulis, D., Karkanis, S., Papageorgas, P., Tzivras, M., 2004. Texture multichannel measurements for cancer precursors identification using support vector machines. Measurement 36, 297–313.

Johnson, J., 1994. Pulse-coupled neural nets: translation, rotation, scale, distortion, and intensity signal invariance for images. Applied Optics 33, 6239–6253.

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.

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

Lee, C.M.P.M.C., 2003. Log-polar wavelet energy signatures for rotation and scale invariant texture classification 25, 590–603.

Li, Z., Liu, G., Yang, Y., You, J., 2012. Scale- and rotation-invariant local binary pattern using scale-adaptive texton and subuniform-based circular shift. IEEE Transactions on Image Processing 21, 2130 –2140.

Liedlgruber, M., Uhl, A., 2011a. Computer-aided decision support systems for endoscopy in the gastrointestinal tract: A review. IEEE Reviews in Biomedical Engineering .

Liedlgruber, M., Uhl, A., 2011b. Predicting pathology in medical decision support systems in endoscopy of the gastrointestinal tract, in: Jao, C. (Ed.), Efficient Decision Support Systems – Practice and Challenges in Biomedical Related Domain. InTech, Rijeka, Croatia, pp. 195–214.

Lo, E.H.S., Pickering, M.R., Frater, M.R., Arnold, J.F., 2004. Scale and rotation invariant texture features from the dual-tree complex wavelet transform, in: Proceedings of the International Conference on Image Processing, ICIP '04, IEEE, Singapore. pp. 227–230.

Lo, E.H.S., Pickering, M.R., Frater, M.R., Arnold, J.F., 2009. Query by example using invariant features from the double dyadic dual-tree complex wavelet transform, in: CIVR '09: Proceeding of the ACM International Conference on Image and Video Retrieval, ACM, Santorini, Fira, Greece. pp. 1–8.

Lowe, D.G., 1999. Object recognition from local scale-invariant features, in: Proceedings of the Seventh IEEE International Conference on Computer on Computer Vision, IEEE. pp. 1150 – 1157.

Ma, Y., Liu, L., Zhan, K., Y.Wu, 2010. Pulse coupled neural networks and one-class support vector machines for geometry invariant texture retrieval. Image and Vision Computing 28, 1524–1529.

Mäenpää, T., 2003. The Local Binary Pattern Approach to Texture Analysis - Extensions and Applications. Ph.D. thesis. University of Oulu.

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, 330–354.

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

Mikolajczyk, K., Cordelia, S., 2004. Scale & affine invariant interest point detectors. International Journal of Computer Vision 60, 63–86.

Mikolajczyk, K., Schmid, C., 2002. An affine invariant interest point detector, in: Proceedings of the European Conference on Computer Vision, Springer Verlag. pp. 128–142.

Montoya-Zegarra, J.A., Leite, N.J., Torres, R., 2007. Rotation-invariant and scale-invariant steerable pyramid decomposition for texture image retrieval, in: Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing, pp. 121–128.

Niveloni, S., Florini, A., Dezi, R., et al., 1998. Usefulness of videoduodenoscopy and vital dye staining as indicators of mucosal atrophy of celiac disease: assessment of interobserver agreement. Gastrointestinal Endoscopy 47, 223–229.

Oberhuber, G., Granditsch, G., Vogelsang, H., 1999. The histopathology of coeliac disease: time for a standardized report scheme for pathologists. European Journal of Gastroenterology and Hepatology 11, 1185–1194.

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

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

Ranganath, H., Kuntimad, G., Johnson, J., 1995. Pulse coupled neural networks for image processing, in: Proceedings of the IEEE Southeastcon '95, 'Visualize the Future', pp. 37 –43.

S. Lazebnik, C.S., Ponce, J., 2005. A sparse texture representation using local affine region. Transactions on Pattern Analysis and Machine Intelligence 27, 1265–1278.

Selesnick, I., Baraniuk, R., Kingsbury, N., 2005. The dual-tree complex wavelet transform. Signal Processing Magazine, IEEE 22, 123–151.

Tan, T.N., 1995. Geometric transform invariant texture analysis, in: Proceedings of SPIE 2488, pp. 475 – 485.

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

Uhl, A., Vécsei, A., Wimmer, G., 2011a. Complex wavelet transform variants in a scale invariant classification of celiac disease, in: Proceedings of the 5th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2011), Las Palmas de Gran Canaria, Spain. pp. 742–749.

Uhl, A., Vécsei, A., Wimmer, G., 2011b. Fractal analysis for the viewpoint invariant classification of celiac disease, in: Proceedings of the 7th International Symposium on Image and Signal Processing (ISPA 2011), Dubrovnik, Croatia. pp. 727 –732.

Varma, M., Garg, R., 2007. Locally invariant fractal features for statistical texture classification, in: Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil.

Varma, M., Zisserman, A., 2009. A statistical approach to material classification using image patch exemplars. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31, 2032–2047.

Varma, M., Zissermann, A., 2005. A statistical approach to texture classification from single images. International Journal of Computer Vision (IJCV) 62, 61–81.

Vécsei, A., Amann, G., Hegenbart, S., Liedlgruber, M., Uhl, A., 2011. Automated marsh-like classification of celiac disease in children using an optimized local texture operator. Computers in Biology and Medicine 41, 313–325.

Vecsei, 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, S68 – S78.

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.

Vedaldi, A., Fulkerson, B., 2008. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/.

Xu, Q., Chen, Y.Q., 2006. Multiscale blob features for gray scale, rotation and spatial scale invariant texture classification, in: Proceedings of 18th International Conference on Pattern Recognition (ICPR), pp. 29–32.

Xu, Y., Huang, S.B., H. Ji, C.F., 2009a. Combining powerful local and global statistics for texture description, in: Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE, pp. 573–580.

Xu, Y., Ji, H., Fermüller, C., 2009b. Viewpoint invariant texture description using fractal analysis. International Journal of Computer Vision 83, 85–100.

Zhan, K., Zhang, H., Ma, Y., 2009. New spiking cortical model for invariant texture retrieval and image processing 20, 1980–1986.

Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C., 2006. Local features and kernels for classification of texture and object categories: A comprehensive study, in: Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on, p. 13.

Zhang, J., Tan, T., 2002. Brief review of invariant texture analysis methods. Pattern Recognition 35, 735–747.

A SCALE-ADAPTIVE EXTENSION TO METHODS BASED ON LBP USING SCALE-NORMALIZED LAPLACIAN OF GAUSSIAN EXTREMA IN SCALE-SPACE

Sebastian Hegenbart and Andreas Uhl

University of Salzburg
Department of Computer Sciences
Salzburg, Austria

ABSTRACT

Local Binary Patterns and its derivatives have been widely used in the field of texture recognition over the last decade. A restriction of methods based on LBP is the variance in terms of signal scaling. This is mainly caused by the fixed LBP radius and the fixed support area of sampling points. In this work we present a general framework to enhance the scale-invariance of all LBP flavored methods, which can be applied to existing methods with minimal effort. Based on scale-normalized Laplacian of Gaussian extrema in scale-space, the global scale of a texture in question is estimated, combined with a confidence measure, to compute scale adapted patterns. By using the notion of intrinsic scales, textures are analyzed at appropriate LBP scales. A comprehensive experimental study shows that the scale-invariance of three different LBP based methods (LBP, LTP, Fuzzy LBP) is highly improved by the proposed extension.

Index Terms— scale, adaptive, LBP, scale-space, estimation

1. INTRODUCTION

In certain scenarios, medical imaging for example, textures are captured at various perspectives and distances [1]. These variations caused by camera motion lead to a visualization of textures under different scales. Methods that are invariant in terms of signal scale can therefore improve the accuracy of an automated classification in such a setting.

Since the introduction of the Local Binary Patterns (LBP) method [2], a variety of LBP based flavors have been developed and applied in various specialized texture recognition scenarios. All LBP based methods share the limitation of being highly affected by scaling of a signal however. Ojala and Mäenpää introduced multi-resolution Local Binary Patterns [3], using a set of different radii with appropriate sampling areas. While this approach improves the discriminative power of the method, it does not employ a scale selection mechanism and hence does not improve invariance in terms of signal scaling.

The idea of combining scale-space extrema with LBP to improve scale-invariance has also been explored by Li et al. [4]. Their approach utilizes the scale of a scale-space maxima at a pixel position as the scale of the Local Binary Pattern descriptor, using a fixed number of neighbors (8) with a fixed sized neighbor sampling area.

Our experimentation has shown that scale selection, based on a single pixel location, is very prone to error, especially for non-regular textures such as shown in Figure 1. We therefore compute a global scale estimation combined with a confidence measure for the estimation to compute scale adapted patterns along a fixed grid (all pixel positions) in an image. Experimental data also shows that a direct mapping from a scale in scale-space to LBP scale is far from optimal. The fact that the estimated scale at a pixel level highly correlates to the intrinsic scale of a texture, leads to rather large LBP scales if using a direct mapping. This reduces the discriminative power of the LBP patterns due to the decreased correlation between sampling points and reference point and leads to sparse sampling if fixed sampling area dimensions are used, reducing the discriminative power even further. We solve this problem by introducing the notion of intrinsic scale, computing a mapping from estimated scale in scale-space to a much more suitable LBP scale. We adjust the sampling area dimensions according to the adapted LBP scale to improve the scale-invariance even further. In this work a general framework to enhance the scale-invariance of all LBP flavored methods, which can be applied to existing methods with minimal effort, is presented.

2. A SCALE-ADAPTIVE EXTENSION TO LBP BASED METHODS

The scale-space theory was first extensively explored in the field of signal processing by Lindeberg [5, 6]. It presents a framework to analyze signals at different scales. Let $f : \mathbb{R}^2 \mapsto \mathbb{R}$ represent a continuous signal, then the linear scale-space representation $L : \mathbb{R}^2 \times \mathbb{R}_+ \mapsto \mathbb{R}$ is defined by

$$L(\cdot;\sigma) = g(\cdot;\sigma) * f, \qquad (1)$$

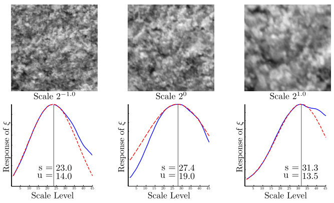


Fig. 1: Scale Estimation of a non-Regular Texture (stone2)

with initial condition $L(\cdot;0) = f$. Where $\sigma \in \mathbb{R}_+$ is the scale parameter, g is a Gaussian function and "$*$" denotes convolution. The scale-space family L is the solution to the diffusion equation (heat equation):

$$\partial_\sigma L = \sigma \left(\frac{\partial^2 L}{\partial x^2} + \frac{\partial^2 L}{\partial y^2} \right) = \sigma \triangle L. \qquad (2)$$

We construct the scale-space and compute the scale-normalized Laplacians ($\sigma^2 \, |\triangle L(\cdot;\sigma)|$, denoted as $\overline{\triangle} I(\cdot;\sigma)$) of each image I at each location $x \in \mathbb{N}^2$ at different scales with $\sigma = c\sqrt{2}^k, k \in \{-4, -3.75, \ldots, 7.75, 8\}$ and $c = 2.1214$. Note that the parameter c acts as a scaling factor of the scale-space and was initially chosen such that the center scale equals a 3 pixel radius. We however found during experimentation that the intrinsic scale of natural textures tends to be large. We therefore expanded the scale-space to cover larger scales as well.

Methods based on scale selection employing the scale-space abstraction identify image locations which are simultaneously a local extremum with respect to both the spatial coordinates and the scale-space parameter. Hegenbart et al. [1] use a local scale reliability mask to improve the reliability of the scale estimation based on such extrema. Experimentation showed however that the utilization of such locations to compute LBP based feature vectors in general leads to an insufficient number of computed patterns and a reduced discriminative power of the feature. The scale selection based on a single pixel location, such as performed by Li et al. [4], however is prone to error, as not each pixel is at a representative scale, especially for non-regular textures as shown in Figure 1. As a consequence we compute a global scale estimate combined with an uncertainty for the estimation to compute LBP patterns along a fixed grid (all pixel positions) in an image, adapted to the global scale of the texture. Let δ denote the Kronecker delta, the scale estimation response function ξ is then

$$\xi(t) := \sum_{x,y} \delta(\arg\max_{\sigma}(\overline{\triangle} L\,(x,y;\sigma)), t)\overline{\triangle} L(x,y;t). \qquad (3)$$

The global scale is identified by searching for the first local maximum of ξ which is then used as seed point for a least-squares Gaussian fit. By using the first local maximum we are capable of consistently estimating the scale of textures exhibiting more than a single dominant scale. The quality of the estimation is improved by considering only data points within a certain offset from the seed point. In our implementation an offset of ± 5 scales is used to fit the Gaussian function. Finally the average value (s) of the fitted Gaussian function is interpreted as the estimated scale where the standard deviation is used as uncertainty of the estimation (u). Due to the fact that the accuracy of the scale estimation is not perfect, we extract weighted LBP patterns at multiple scales to improve robustness, taking the uncertainty of the estimation into account. The weighted patterns contribute to the LBP histogram, based on the response of the unnormalized Gaussian function at the specific scale level. In our implementation only scale levels with a response ≥ 0.9 were used. Figure 1 illustrates the fitted Gaussian function (dashed line) to the scale estimation response function (solid line) of three textures at different scales.

2.1. Intrinsic Texture Scale

The response of the scale-normalized Laplacian of Gaussians (LoG) attains a maximum if the zeros are aligned with a circular shaped structure. Hence scales estimated, based on the LoG, correlate strongly with the scale of the dominant circular shaped structures of a texture. As a consequence, the estimated scale is highly related to an essential property of each texture, the *intrinsic scale* of a texture.

A texture exhibiting pebbles for example and a texture exhibiting sand, captured at the same distance, might have equal scales in terms of camera-scale, but different scales in terms of the scale-space, a consequence of different intrinsic scales. In contrast, sand and pebbles captured at a different camera-scales, corresponding to the difference of the textures' intrinsic scales, are equal in scale in terms of the scale-space. Scales estimated in the scale-space domain are therefore always a combination of the intrinsic texture scale and the camera-scale.

Utilizing LBP based methods, textures are described by the means of the joint distribution of underlying micro structures. The discriminative power is not directly related to the scale of the dominant structures of an image. This statement is based on the observation that LBP based methods are successfully used in classification scenarios with multiple textures and multiple different intrinsic scales, using a set of fixed sized LBP radii. Hence, a direct mapping between estimated scale in scale-space of a texture to LBP scale, introduces several problems as discussed in Section 1, without improving the descriptive capabilities of the method in general.

The identification of an intrinsic scale of a general texture is a non-trivial problem. A requirement on an intrinsic scale estimation method would be scale-invariance, a property that the LoG response in scale-space does not provide.

Based on the property that the intrinsic scale of a texture is scale-invariant, the intrinsic factors cancel each other out for two estimated scales in scale-space of the same texture. We hence estimate the scale in scale-space per texture class in the training data, denoted as *trained base scale*, as the median of all estimated scales from all images within a class.

This approach requires that all samples of a specific texture class are at a single or relatively close camera-scale in the training data. This requirement could be loosened by identifying the *trained base scale* per class and camera-scale however. A benefit of estimating the *trained base scale* per class is, that additional information such as a shape model per texture class could be computed and used for improving the feature extraction further. This is part of our current work.

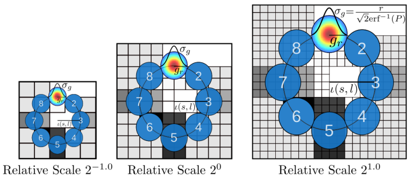
By using the *trained base scale* of a texture class, we can define a mapping from the scale-space domain to the LBP scale domain (the LBP radius). For a texture with estimated scale in scale-space s the adapted LBP radius is then computed in reference to the *trained base scale* ($\bar{s}_l$) of texture class l as

$$\iota(s,l) = b\frac{s}{\bar{s}_l}, \tag{4}$$

with b denoting the defined LBP radius at the *trained base scale* $\bar{s}_l$. Please note that the linearity of this mapping is a requirement for scale-invariance. The value of b defines the LBP scale the training textures are analyzed at. In order to be able to adapt to down-scaled textures, the value requires to be larger than the minimal LBP radius. We use $b = 3$ as default.

Considering the extraction of feature vectors for evaluation, we are not capable of identifying the corresponding *trained base scale* for such input textures, due to the inability of estimating the intrinsic scale of the texture. Hence, for each input texture a set of feature vectors is computed, one feature vector in relation to the specific *trained base scale* of each class in the training set. Feature vectors computed in relation to the same texture class will be based on a matching *trained base scales* (the input sample and the texture class are at the same intrinsic scale), canceling out the intrinsic scale factors. Feature vectors computed in relation to other texture classes (and other *trained base scales*) are computed at the wrong relative LBP scale.

The pairwise comparison of feature vectors computed at different *trained base scales* can lead to very high LBP scales. As a consequence such feature vectors exhibit a higher general similarity among all textures (due to the high amount of low-pass filtering), leading to a decreased discriminative power of the system. As a consequence, only features computed at the same *trained base scale* are compared during classification. Please note, that this poses no unfair bias or advantage to the classifier as each of the feature vectors of an evaluation sample is compared to the corresponding feature vector of each class in the training data.



Relative Scale $2^{-1.0}$ Relative Scale 2^{0} Relative Scale $2^{1.0}$

Fig. 2: Adaptive Radius and Adaptive Sampling Areas

2.2. Adaptive Sampling Support Area Dimension

Scaling of a texture leads to a scaled spatial extent of structures. Therefore the number of pixels covering structural information changes. As a consequence, the size of the sampling support area has to be adapted accordingly. Ojala and Mäenpää [3] first used Gaussian filtering to adapt the sampling support area to various LBP scales to compute multi-resolution LBP, generally improving the robustness of the method. By employing low-pass filtering, a pixel at a single spatial location encodes information of its spatial neighborhood. We use the estimated global scale of a texture in relation to a *trained base scale* to adapt the LBP radius as well as the size of the sampling support area to achieve scale-invariance. The radius of the Gaussian filter for a texture at estimated global scale s in relation to the *trained base scale* of texture l is computed as

$$g_r = \frac{\iota(s,l)\pi}{N}, \tag{5}$$

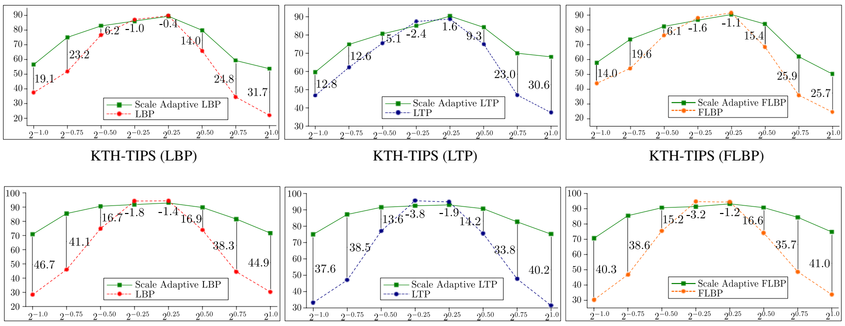
for N being the number of neighbors. The Gaussian filter coefficients are then computed such that P percent of the mass of the Gaussian function is covered

$$\begin{aligned}
\int_{-g_r}^{g_r} e^{-\frac{x^2}{2\sigma_g{}^2}}dx &= P\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma_g{}^2}}dx \\
2\int_{0}^{g_r} e^{-\frac{x^2}{2\sigma_g{}^2}}dx &= P\sigma_g\sqrt{2\pi} \\
\sigma_g &= \frac{g_r}{\sqrt{2}\mathrm{erf}^{-1}(P)}.
\end{aligned} \tag{6}$$

We chose P to be 0.99 which corresponds to 99% of the mass of the Gaussian function. As the sampling of a Gaussian function with very few sampling points leads to a large error we use the error function (erf) to improve the stability of the computation of the one dimensional Gaussian filters centered at 0

$$G(x;\sigma_g) = \frac{-\mathrm{erf}(\frac{x-0.5}{\sigma_g}) - \mathrm{erf}(\frac{x+0.5}{\sigma_g})}{2}, \tag{7}$$

which are then used in a separable convolution. Note that, as a bonus, the computed filter is already normalized, therefore the re-normalization can be avoided. Figure 2 illustrates the relation between estimated scale in scale-space and the adapted LBP scale.



KTH-TIPS (LBP)

KTH-TIPS (LTP)

KTH-TIPS (FLBP)

Kylberg (LBP)

Kylberg (LTP)

Kylberg (FLBP)

Fig. 3: Mean Overall Classification Accuracies

3. EXPERIMENTS AND RESULTS

We constructed a set of experiments to analyze the impact of the proposed scale-adaptive extension to three different, popular types of LBP based methods, which all utilize different types of encoding schemes. The standard LBP method [2], the Local Ternary Patterns (LTP) operator [7] as well as the Fuzzy Local Binary Patterns (FLBP) method [8] are compared with the scale-adapted variation of each of those methods. Please note that the standard methods were used in combination with the multi-resolution Local Binary Patterns extension [3] using three scales and 8 neighbors, the best configuration we were able to find for the given datasets.

The experiments are based on two different texture databases. The KTH-TIPS database [9] exhibits texture images from 10 different materials captured at 9 different scales with 9 samples per material. Sub-images of size 128 × 128 pixels were extracted from the center of each original image. Unfortunately, besides KTH-TIPS there are no other publicly available high quality texture databases with an available ground-truth of scales. We therefore had to resort to a simulation of the scaling of textures. A subset of the Kylberg texture database [10], consisting of 28 materials with 160 unique texture patches per class, captured at a single scale, was used for the simulation. The high resolution of each patch (576 × 576 pixels) allowed us to simulate the scaling without relying on up-sampling, leading to a smaller amount of interpolation artifacts. The simulation of scaling was performed according to the scales of the KTH-TIPS database, interpreting the original image patches as the maximum scale $2^{1.0}$. Image patches of size 128 × 128 pixels were then extracted from the center of the re-scaled patches. Due to the huge number of samples in the Kylberg database we use a subset consisting of 20 unique texture patches per class (5 per image) for experimentation. The experiments were designed to explicitly reflect the scale invariance properties of the studied methods. We chose KTH-TIPS scale 5 and Kylberg scale 2^0 as the training scale. This gives us the opportunity to study the method's capability of adapting to higher as well as lower relative scales.

The classification was performed based on a k-nearest neighbors classifier, utilizing the histogram intersection as similarity metric. The minimum number of neighbors was set to 1 for all experiments while the maximum number of neighbors was set according to the maximum number of samples per texture class (9 for KTH-TIPS and 20 for Kylberg).

Figure 3 shows the mean overall classification accuracy (OCR) over all k-values. The numbers between the results illustrate the absolute difference in mean OCR between the scale-adapted type of a method and the standard version. The horizontal axes denote the relative scale differences as compared to the training set.

4. DISCUSSION AND CONCLUSION

The experimental results show that the scale-adaptive extension improved the scale-invariance of all three flavors of the LBP method. The scale-adapted methods performed slightly worse as compared their standard counterpart at very small relative scale differences. This is caused by some failed scale estimations. Considering large scale differences however, the scale-adapted methods vastly outperform the standard methods, with improvements of over 40 percentage points. Although the scale-adaptive extension comes at the cost of higher computational demand, it can improve the classification accuracy of LBP based methods in a setting with varying scales significantly without changing the encoding or extraction scheme of the standard method, and could therefore be used in combination with a variety of LBP based methods.

5. REFERENCES

[1] S. Hegenbart, A. Uhl, A. Vécsei, and G. Wimmer, "Scale invariant texture descriptors for classifying celiac disease," *Medical Image Analysis*, vol. 17, no. 4, pp. 458 – 474, 2013.

[2] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, January 1996.

[3] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, July 2002.

[4] Z. Li, G. Liu, Y. Yang, and J. You, "Scale- and rotation-invariant local binary pattern using scale-adaptive texton and subuniform-based circular shift," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2130–2140, 2012.

[5] T. Lindeberg, "Scale-space for discrete signals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 3, pp. 234–254, 1990.

[6] T. Lindeberg, *Discrete scale-space theory and the scale-space primal sketch*, Ph.D. thesis, Royal Institute of Technology, 1991.

[7] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," in *Analysis and Modelling of Faces and Gestures*, October 2007, vol. 4778 of *Lecture Notes in Computer Science*, pp. 168–182.

[8] D. Iakovidis, E. Keramidas, and D. Maroulis, "Fuzzy local binary patterns for ultrasound texture characterization," in *ICIAR*. 2008, vol. 5112 of *Lecture Notes in Computer Science*, pp. 750–759, Springer.

[9] E. Hayman, B. Caputo, M. Fritz, and J.-O. Eklundh, "On the significance of real-world conditions for material classification," in *Proceedings of the 8th European Conference on Computer Vision (ECCV)*. 2004, vol. 3024 of *Lecture Notes in Computer Science*, pp. 253–266, Springer.

[10] G. Kylberg, "The kylberg texture dataset v. 1.0," External report (Blue series) 35, Center for Image Analysis, Swedish University of Agricultural Sciences, Uppsala University, Uppsala, Sweden, September 2011.

An Orientation-Adaptive Extension to Scale-Adaptive Local Binary Patterns

Sebastian Hegenbart and Andreas Uhl
University of Salzburg
Department of Computer Sciences
shegen@cosy.sbg.ac.at and uhl@cosy.sbg.ac.at

***Abstract*—Methods based on Local Binary Patterns have been used successfully in a wide range of texture classification tasks. A restriction shared by all methods based on Local Binary Patterns is the high sensitivity to signal scale. In recent work we presented a general framework for scale-adaptive computation of Local Binary Patterns, improving the accuracy in texture classification scenarios involving varying texture-scales highly. In this work, the scale-adaptive methodology is extended by an orientation-adaptive computation of patterns, leading to a scale- and rotation-invariant classification. The results suggest that estimating a global orientation to build orientation-adaptive LBPs is superior to the previously introduced rotation-invariant encodings. The proposed framework allows the use of the highly-discriminative LBPs in less-constrained situations, where both orientation, as well as scale variations, are to be expected.**

I. Introduction

A big challenge in texture classification scenarios in unconstrained environments is dealing with varying scales and orientations. This is especially true in medical image acquisition such as endoscopy [3]. As a result, research focusing on scale- and rotation-invariant feature descriptors has been a hot topic in the recent past.

Methods based on Local Binary Patterns (LBP [9]) have been successfully used in a wide range of texture classification scenarios. A restriction shared by all those methods is their high sensitivity in terms of signal scaling, therefore reducing their applicability to a constrained environment with only minor scale variances among textures. The correct alignment of micro structures in terms of orientation is an essential requirement for the accuracy of the baseline LBP type methods. Ojala et al. [10] alleviate this restriction by using a special type of rotation-invariant pattern encoding, leading to a possibly reduced discriminative power of features. A drawback of that method is the limited angular resolution. As a consequence the rotation-invariant encoding is not very well suited in a scale-adaptive computation. In [2] we proposed a general scale-adaptive methodology that enables the use of the highly-discriminative LBPs in less-constrained situations, where scale variations are to be expected. Experiments have shown that this scale-adaptive framework improved the accuracy of LBP based methods in scenarios with varying scales significantly.

In this work we present an extension to this scale-adaptive framework, alleviating the restriction of correct texture orientation alignment by utilizing a global orientation-estimate. This allows the use of highly-discriminative LBPs in a scenario with varying scales and orientations.

By using multi-scale second moment matrices [7], a global orientation is estimated at dominant local scales, leading to a robust orientation estimation in noisy environments with varying texture scales. By leveraging the already pre-computed scale-spaces, our proposed orientation estimation approach integrates naturally with the scale-adaptive LBP framework at moderate computational cost. Employing the estimated orientation, an orientation-adaptive computation of LBP patterns is performed. Our results suggest that estimating a global orientation to build orientation-adaptive LBPs in a scale-adaptive computation is superior to the previously introduced rotation-invariant encodings.

In Section II we give a review of the general scale-adaptive computation that enables the use of LBPs in scenarios with varying scales. Section III-A describes the orientation-adaptive methodology. The fusion of the orientation- and scale-adaptive computation is covered in Section IV. The experiments conducted to evaluate the proposed methodology are described in Section V, the results presented and discussed in Section VI. Finally Section VII concludes the paper.

II. Scale-Adaptive Computation of LBP

The scale-adaptive computation is based on a global scale estimation combined with a confidence measure for the estimation. Based on the estimated scale, the radius of the LBP as well as the dimension of the sampling area is adapted accordingly. This methodology allows the use of LBP flavored methods in a scenario with varying scales.

A. Scale Estimation

We employ a global scale estimation algorithm which is based on scale-normalized Laplacian of Gaussian extrema in scale-space The scale-space theory was first extensively explored in the field of signal processing by Lindeberg [6]. It presents a framework to analyze signals at different scales. Let $f : \mathbb{R}^2 \mapsto \mathbb{R}$ represent a continuous signal, then the linear scale-space representation $L : \mathbb{R}^2 \times \mathbb{R}_+ \mapsto \mathbb{R}$ is defined by

$$L(\cdot;\sigma) = g(\cdot;\sigma) * f, \tag{1}$$

with initial condition $L(\cdot;0) = f$. Where $\sigma \in \mathbb{R}_+$ is the scale parameter, g is a Gaussian function and "$*$" denotes convolution. The scale-space family L is the solution to the diffusion equation (heat equation):

$$\partial_\sigma L = \sigma \left(\frac{\partial^2 L}{\partial x^2} + \frac{\partial^2 L}{\partial y^2} \right) = \sigma \triangle L. \tag{2}$$

We construct the scale-space and compute the scale-normalized Laplacians ($\sigma^2 \left| \triangle L(\cdot;\sigma) \right|$, denoted as $\overline{\triangle} I(\cdot;\sigma)$) of each image I at each location $x \in \mathbb{N}^2$ at different scales with $\sigma = c\sqrt{2}^k, k \in \{-4, -3.75, \ldots, 7.75, 8\}$ and $c = 2.1214$. Note that the parameter c acts as a scaling factor of the scale-space and was initially chosen such that the center scale equals a 3 pixel radius. We however found during experimentation that the intrinsic scale of natural textures tends to be large. We therefore expanded the scale-space to cover larger scales as well.

Methods based on scale selection employing the scale-space abstraction identify image locations which are simultaneously a local extremum with respect to both the spatial coordinates and the scale-space parameter (3D maxima), a prominent example is the Scale Invariant Feature Transform (SIFT [8]). Experimentation has shown however that the utilization of such locations for a global scale estimation is unreliable. This can be seen in Figure 1, comparing the distribution of the responses of the 3D maxima with the responses of the scale estimation of textures, the extrema are either at various different scales or only a small number of extrema is present, leading to unreliable scale estimations. We therefore use the distribution of responses of the scale normalized Laplacians to estimate a global scale. The scale estimation response function ξ is

$$\xi(t) := \sum_{x,y} \overline{\triangle} I(x, y; t). \tag{3}$$

The global scale is identified by searching for the first local maximum of ξ which is then used as seed point for a least-squares Gaussian fit. By using the first local maximum we are capable of consistently estimating the scale of textures exhibiting more than a single dominant scale. The quality of the estimation is improved by considering only data points within a certain offset from the seed point. In our implementation an offset of ± 5 scale levels is used to fit the Gaussian function. Finally the mean value $\tilde{s}$ of the fitted Gaussian function is interpreted as the dominant level in scale-space. The standard deviation u of the fitted Gaussian is used as uncertainty of the estimation. For a given dominant level in scale-space $\tilde{s_i}$, the spatial scale s_i corresponds to to the scale parameter σ_i at the dominant scale level. Figure 2 illustrates the fitted Gaussian function (dashed line) to the scale estimation response function (solid line) of three textures at different scales.

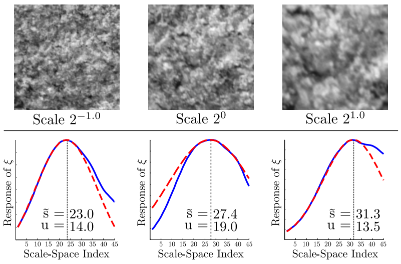


Fig. 2: Scale Estimation of a non-Regular Texture (stone2).

The response of the scale-normalized Laplacian of Gaussian (LoG) attains a maximum if the zeros are aligned with a circular shaped structure. Hence scales estimated, based on the LoG, correlate strongly with the scale of the dominant circular shaped structures of a texture. As a consequence, the estimated scale is highly related to an essential property of each texture, the *intrinsic scale* of a texture.

A texture exhibiting pebbles for example and a texture exhibiting sand, captured at the same distance, might have equal scales in terms of camera-scale, but different scales in terms of the scale-space, a consequence of different intrinsic scales. In contrast, sand and pebbles captured at a different camera-scales, corresponding to the difference of the textures' intrinsic scales, are equal in scale in terms of the scale-space. Scales estimated in the scale-space domain are therefore always a combination of the intrinsic texture scale and the camera-scale.

The identification of an intrinsic scale of a general texture is a non-trivial problem. A requirement on an intrinsic scale estimation method would be scale-invariance, a property that the LoG response in scale-space does not provide. The estimated scale in scale-scape is therefore always a combination of camera-scale and intrinsic texture scale. Please refer to [3] for more details.

III. Orientation-Adaptive Local Binary Patterns

The correct alignment of micro structures in terms of orientation is an essential requirement for the accuracy of the baseline LBP type methods. Ojala et al. [10] alleviate this restriction by using a special type of rotation-invariant pattern encoding. The original pattern is shifted circularly until a minimum with respect to a numeric interpretation of the pattern is found. As a consequence all patterns are implicitly aligned among each other.

A drawback of this approach is the limited angular resolution, depending on the number of used LBP-neighbors. For a standard LBP-neighborhood with 8 neighboring samples, this angular resolution corresponds to 45 degrees. A side-effect of the encoding is the decreased number of individual patterns. The authors propose to use uniform patterns in combination with the rotation-invariant encoding to improve robustness, implicitly improving the angular resolution by considering only special type of micro structures. Uniform patterns are a subset of patterns with a maximum of two transitions between 1 and 0. The proposed combination of rotation-invariant and uniform patterns reduces the number of individual patterns even further. Experiments discussed in Section VI show that the small number of individual patterns leads to a decreased classification accuracy if combined with the scale-adaptive computation. As a consequence we utilize the estimation of a global orientation to build orientation-adaptive LBP. Following literature on LBP, we refer to Local Binary Patterns utilizing the rotation-invariant encoding in combination with uniform pattern as LBP^{riu} from here on.

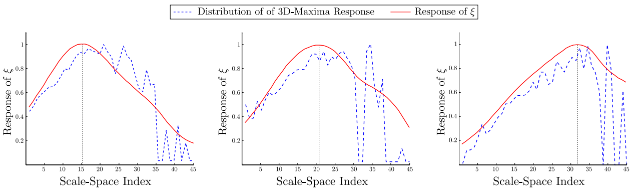


Fig. 1: Distribution of 3D-Maxima compared to the Response of ξ.

A. Orientation Estimation

We utilize the multi-scale second moment matrices [7] of an image, computed at dominant local scales, for a robust orientation estimation in noisy environments with varying texture scales. The second moment matrix (also known as structure tensor) summarizes the predominant directions of the gradient in a specific pixel neighborhood of an image. In contrast to the second moment matrix, the multi-scale second moment matrix is defined over two scale parameters. It allows to estimate the shape of visual structures at their dominant scale, as detected by the scale-estimation algorithm.

The local scale, denoted by t determines the scale in terms of the scale-space a local structure is analyzed at. The integration scale i is used as parameter to a Gaussian function g defining the shape and weights of a specific neighborhood area in the image over which the gradient response is accumulated. We compute the multi-scale second moment matrices at each location $x \in \mathbb{R}^2$ of an image I. The local scale t is selected depending on the estimated texture scale (see Section IV), the integration scale $i = \sqrt{2}t$ depends on the corresponding detection scale. The second moment matrix for an image location x at local scale t is then computed as

$$\mu(x;t,i) = \int_{\xi \in \mathbb{R}^2} (\nabla I)(x-\xi;t)(\nabla I)^T(x-\xi;t)\, g(\xi;i)\, d\xi. \tag{4}$$

We denote $(\nabla I)(x;t)$ as the gradient of the scale-space representation of image I at scale t and position x.

The multi-scale second moment matrix is positive definite, it therefore has two non-negative eigenvalues which correspond to the length of the axes of an ellipse (up to some constant factor). The eigenvectors of the multi-scale second moment matrix correspond to the orientation of the ellipse. By computing the angle between the major axis of the ellipse and the vertical axis we identify the dominant orientation at a specific image position. Due to the ambiguous orientation of the ellipse, all angles are treated modulus 180.

Based on the distributions of all orientations at all pixel locations, a global orientation is estimated for an image. This is done by fitting a Gaussian function to the distribution of orientations. To improve the quality of the estimation, we remove data points with an offset of ±40 degrees from the maximum prior to the fitting process. Finally, the average value of the Gaussian is interpreted as the dominant orientation, the standard deviation of the fitted Gaussian function is interpreted as the uncertainty of the estimation. To avoid using invalid orientation estimations, we reject estimations with an uncertainty above 20. In such a case the estimated orientation is defined as 0 degrees.

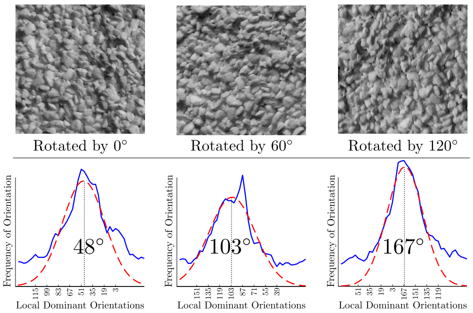


Fig. 3: Orientation Estimation (pearlsugar1).

Figure 3 illustrates the fitting of a Gaussian function (dashed red line) to the distribution of orientations (solid blue line) of three differently rotated images. The numbers centered at each Gaussian function relate to the mean value of the fitted Gaussian function, corresponding to the estimated global orientation of the specific image.

B. Global versus Local Orientation Estimation

As explained in Section III-A a global orientation is computed for a specific image. In theory however, a texture could consist of multiple sub textures with different orientations, leading to potentially worse estimation accuracies. We therefore evaluated the performance of a local orientation estimation on a pixel basis in comparison to the used global orientation estimation. The local orientation estimation is based on the same methodology utilizing multi-scale second moment matrices as described for the global orientation estimation. In contrast to the global orientation however, the estimation is done per pixel instead of fitting a Gaussian function to the distribution of orientations to estimate a global orientation. Figure 4 demonstrates that the accuracy of the local estimation is inferior as compared to the global estimation. The mean absolute error of the estimated orientation (vertical axis) was computed between a reference image without rotation and the same image with a specific rotation, as depicted by the horizontal axis, for all images in the Kylberg database which was also used for experimentation as explained in Section V. The mean absolute error was computed for three relative scales between the reference and the rotated images. We can see that the global estimation is superior to the local estimation in all regards. We assume that this is caused by homogeneous pixel areas which do not allow for a robust estimation of orientation, introducing a large error. The results also indicate that scaling of the textures has only a minor impact to the general accuracy of the orientation estimation method, an important property for using the method in combination with the scale-adaptive methodology.

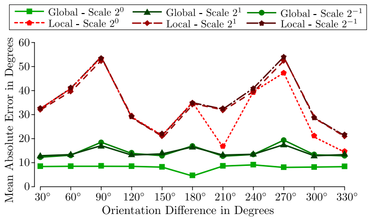


Fig. 4: Global versus Local Orientation Estimation Error.

C. Impact of Signal Noise

Utilizing the multi-scale second moment matrix allows to estimate the orientation of a visual structure at its dominant scale, as a benefiting side effect of utilizing the scale-space data, signal noise is suppressed to some degree. We explicitly constructed an experiment to evaluate this property. The mean absolute error of the orientation estimation is computed for noisy image textures at the same texture scale. Let P be the set of all pixels in image $I \in \mathbb{N}^2$, $\omega = (\omega_p)_{p \in P}$, be a collection of independent identically distributed real-valued random variables following a Gaussian distribution with mean m and standard deviation σ. We simulate thermal noise as additive Gaussian noise with $m = 0$, variance σ for pixel p at position x, y as

$$N(x,y) = I(x,y) + \omega_p, \quad p \in P, \tag{5}$$

with N being the noisy image, for an original image I. Figure 5 illustrates the effects of Gaussian white noise to the global orientation estimation. We see that noise only has a minor impact to the average accuracy of the method, another welcome benefit of using multi-scale second moment matrices for orientation estimation.

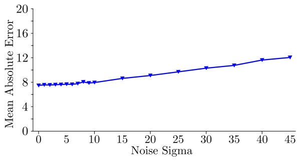


Fig. 5: The Impact of Signal Noise to the Estimation.

IV. Combining the Scale-Adaptive Computation with the Orientation-Adaptive Computation

The orientation-adaptive computational approach integrates very naturally into the scale-adaptive LBP framework. As a consequence of computing the LoG for scale-estimation instead of using the Difference of Gaussians approach, the scale-space data can be re-used for computing the multi-scale second moment matrices used for orientation estimation. Therefore the Gaussian filtering to compute the local scale t can be omitted. We adaptively select the local scale t of the multi-scale second moment matrix, based on the estimated scale of a texture. By doing so, we guarantee a robust orientation estimation across different texture scales. Experimentation has shown that a reasonable value for the local scale t is half of the estimated texture scale. This is explained by the property that the estimated scale at a pixel level highly correlates to the intrinsic scale of a texture, therefore leading to rather large estimated scales. Large local scales however would result in a decreased estimation accuracy. By re-using the scale-space data, the computation of the multi-scale second moment matrices only involves the computation of the first partial derivatives in both image dimensions as well as a convolution with a Gaussian filter to compute the integration scale i. Figure 6 illustrates schematically how the scale- and orientation-adaptive computation is combined. Based on the estimated texture scale, appropriate LBP radii and neighborhood sampling area dimensions are chosen. The ordering of the computation is adaptively chosen depending on the estimated orientation. Please note that due to the ambiguity of the orientation, we compute two patterns at each image location, rotated circularly to accommodate orientations above 180 degrees. This is indicated by the red sampling points which correspond to the respective starting location of the computation.

To compensate for possible errors of the orientation estimation as well as unsuitable alignments on the pixel grid, we compute multiple LBP histograms using a small interval of different orientations depending on the estimated orientation. The size of the interval is chosen depending on the uncertainty measure of the orientation estimation. As a consequence the chosen interval for an unreliable orientation estimation is wider and the error is more likely to be compensated. The interval width chosen for the experiments discussed in Section V was 0.7 times the standard deviation (interpreted in degrees) of the fitted Gaussian. This value was not optimized and might be dependent on the given problem however. For each orientation

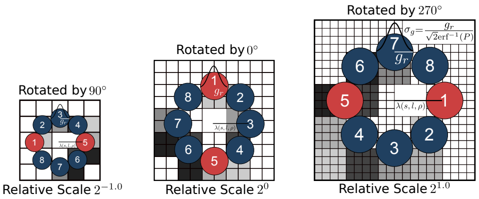


Fig. 6: Illustration of the Scale- and Orientation-Adaptive Computation.

in the interval (in steps of 5 degrees), a separate LBP histogram is computed. Finally, the best alignment of orientations is implicitly chosen during classification by selecting the minimum of all distances computed pairwise between all LBP histograms computed based on the specific orientation within the intervals of two texture images. Note that this does not pose an unfair advantage to the classifier, as no information about class membership is used implicitly or explicitly. This approach is comparable to the cyclic shift of binary iris features used to compensate for small rotational alignment errors in biometric systems for example.

V. Experiments

We constructed a large set of experiments to analyze the performance of the orientation-adaptive extension to the scale-adaptive LPB framework in a scenario with varying scales and varying image rotations. We explicitly compare the accuracy of LBP^{riu} methods employing the scale-adaptive methodology with the accuracy of the standard LBP methods employing the proposed scale- and orientation-adaptive framework. Additionally we analyze the performance of non scale-adaptive methods based on LBP^{riu} in the same scenario.

The used methods are the LBP^{riu} method [10], the Local Ternary Patterns (LTP^{riu}) operator [11] as well as the Fuzzy Local Binary Patterns (FLBP^{riu}) method [4]. Please note that these methods were used in combination with the multi-resolution Local Binary Patterns extension [10] based on three scales and 8 neighbors, the best configuration we were able to find for the given data sets.

The experimentation is based on two independent texture databases. The KTH-TIPS database [1] exhibits texture images from 10 different materials captured at 9 different scales with 9 samples per material. Sub-images of size 128×128 pixels were extracted from the center of each accordingly rotated original image. The rotation was performed using bilinear interpolation. We simulated rotations of 30 degrees, 60 degrees 120 degrees and 180 degrees respectively. Due to the dimension of the original images of material ”cracker“, this material class could not be used for simulating rotation without a large black border within the 128×128 image patches. Unfortunately, besides KTH-TIPS there are no other publicly available high quality texture databases with an available ground-truth of scales. We therefore had to resort to a simulation of the scaling of textures. A subset of the Kylberg texture database [5], consisting of 28 materials with 160 unique texture patches per class, captured at a single scale, was used for the simulation. The image database contains rotated versions of each image at 30 degree steps ranging from 0 to 330 degrees. The high resolution of each patch (576×576 pixels) allowed us to simulate the scaling without relying on up-sampling, leading to a smaller amount of interpolation artifacts. The simulation of scaling was performed according to the scales of the KTH-TIPS database, interpreting the original image patches as the maximum scale $2^{1.0}$. Image patches of size 128×128 were then extracted from the center of the re-scaled patches. Due to the huge number of samples in the Kylberg database we use a subset consisting of 20 unique texture patches per class (5 patches per image) for experimentation.

The experiments were designed to explicitly reflect the properties of the studied methods. The images from the KTH-TIPS database at scale 5 without rotations build the training set for experiments based on the KTH-TIPS database. Respectively the images from the Kylberg database at scale 2^0 without rotation are used as training data for experiments based on the Kylberg database. To evaluate the impact of rotation and scale, the classification was performed on the corresponding scaled and rotated version of the data from each of the databases. The used classification method was a k-nearest neighbors classifier. The maximum value of k was chosen depending on the number of samples per material class. In case of the Kylberg database the maximum value of k was set to 20 while the maximum value of k was 9 in case of the KTH-TIPS database. To allow for an unbiased evaluation, all interpreted results are the mean accuracy over all possible k-values ranging from 1 to the specific maximum.

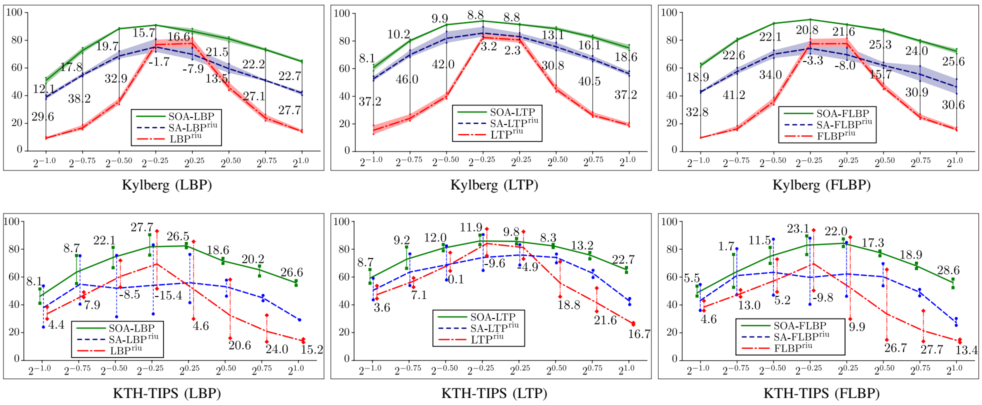
VI. Results

Figure 7 presents the results of the experiments. The horizontal axis denotes the relative scale difference between training data and evaluation data while the vertical axis corresponds to the classification accuracy. The bold lines show the mean classification accuracy over all image rotations (5 different rotations for the KTH-TIPS set and 12 for the Kylberg database). We visualize the minimum and maximum classification accuracy over all rotations with error bars in case of the KTH-TIPS database as well as a smaller error bar with the corresponding area in case of the Kylberg database.

Methods utilizing the proposed scale- and orientation-adaptive methodology are abbreviated as *SOA* and the respective name of the used LBP based method, the scale-adaptive method. Methods utilizing the scale-adaptive methodology in combination with the rotation-invariant encoding are abbreviated as *SA* and the specific method's name. The name of the methods based on LBP are used as known from literature.

The difference in mean classification accuracy between the proposed scale- and orientation-adaptive (SOA) framework and the scale-adaptive (SA) framework using the rotation invariant encoding is reflected by the upper row of numbers. The lower row of numbers label the difference between the scale-adaptive framework based on LBP^{riu} with the respective standard method.

Figure 7 shows that the mean classification accuracy of methods utilizing the scale- and orientation-adaptive methodology (SOA) are superior in terms of classification accuracy and variance as compared to methods utilizing the scale-adaptive (SA) framework with rotation-invariant encoding. Comparing the results with prior experiments in [2], we see that the accuracy of the standard methods decreased due to the rotation invariant encoding. This is reflected by the fact that the maximum results are below the results in [2]. In general we observe a minor degree of variation caused by the different orientations across the results. Interestingly the methods employing the scale-adaptive framework (SA) exhibit the highest degree of variance, a property we expected due to the reduced discriminative power as discussed in Section III. Methods utilizing the proposed scale- and orientation-adaptive (SOA) framework show the smallest degree of variance with respect to orientation. Additionally the mean classification accuracy is clearly above the accuracy of traditional methods as



Kylberg (LBP) Kylberg (LTP) Kylberg (FLBP)

KTH-TIPS (LBP) KTH-TIPS (LTP) KTH-TIPS (FLBP)

Fig. 7: Mean Overall Classification Accuracies over all Rotations.

well as methods utilizing the scale-adaptive (SA) approach in combination with the rotation-invariant encoding. Concerning the results based on the KTH-TIPS database, we can see that the variation caused by rotation is considerably higher across all methods. The smallest variations caused by rotation is again observed for methods utilizing the proposed scale- and orientation-adaptive (SOA) methodology. In parallel to the Kylberg database, methods based on the SOA framework show the highest mean accuracy. We observe the highest amount of variance of methods utilizing the SOA methodology at small relative scales. We assume this is caused by the higher impact of the orientation estimation error for textures at a smaller relative scale. In general, the trends observed for the Kylberg database are confirmed by the results based on the KTH-TIPS database.

VII. Conclusion

In this work, we presented an orientation-adaptive extension to the scale-adaptive LPB framework. By leveraging the already pre-computed scale-spaces, our proposed orientation estimation approach integrates naturally with the scale-adaptive LBP framework at moderate computational cost. In particular, using multi-scale second moment matrices, computed at dominant local scales, leads to 1) robust orientation estimation in noisy environments and 2) scenarios with varying texture scales. Our experiments suggest that estimating a global orientation to build orientation-adaptive LBPs is superior to the previously introduced rotation-invariant encodings; this is reflected by less variance in classification accuracy as well as superior mean accuracy over multiple orientations. In summary, the proposed framework enables the use of the highly-discriminative LBPs in less-constrained situations, where both orientation as well as scale variations are to be expected.

References

[1] E. Hayman, B. Caputo, M. Fritz, and J.-O. Eklundh. On the significance of real-world conditions for material classification. In *Proceedings of the 8th European Conference on Computer Vision (ECCV)*, volume 3024 of *Lecture Notes in Computer Science*, pages 253–266. Springer, 2004.

[2] S. Hegenbart and A. Uhl. A scale-adaptive extension to methods based on lbp using scale-normalized laplacian of gaussian extrema in scale-space. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing ICASSP '14*, pages 4352–4356, 2014.

[3] Sebastian Hegenbart, Andreas Uhl, Andreas Vécsei, and Georg Wimmer. Scale invariant texture descriptors for classifying celiac disease. *Medical Image Analysis*, 17(4):458 – 474, 2013.

[4] D. Iakovidis, E. Keramidas, and D. Maroulis. Fuzzy local binary patterns for ultrasound texture characterization. In *ICIAR*, volume 5112 of *Lecture Notes in Computer Science*, pages 750–759. Springer, 2008.

[5] G. Kylberg. The kylberg texture dataset v. 1.0. External report (Blue series) 35, Center for Image Analysis, Swedish University of Agricultural Sciences, Uppsala University, Uppsala, Sweden, September 2011.

[6] T. Lindeberg. Scale-space for discrete signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(3):234–254, 1990.

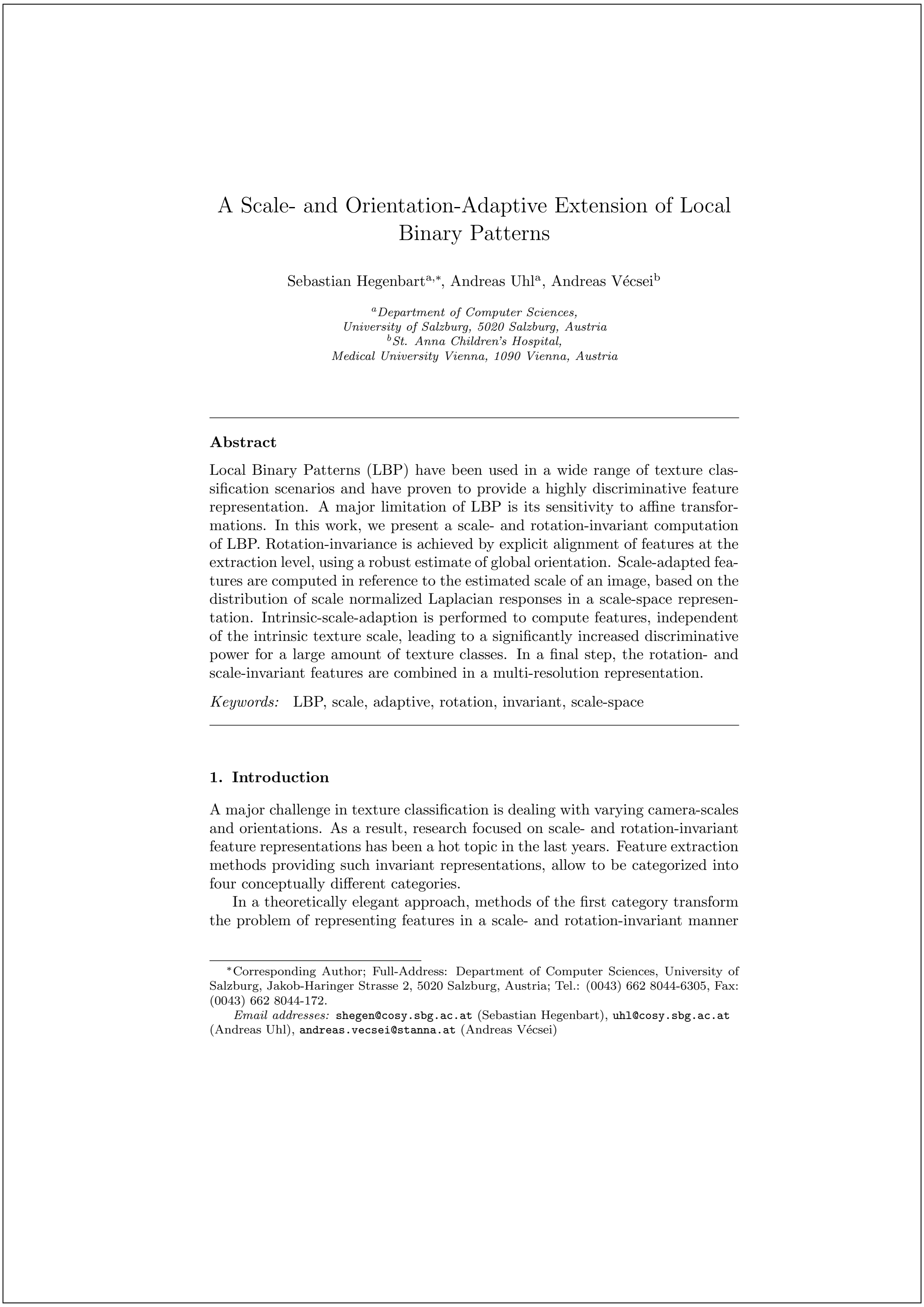
[7] Tony Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, Norwell, MA, USA, 1994.

[8] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

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

[10] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, July 2002.

[11] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. In *Analysis and Modelling of Faces and Gestures*, volume 4778 of *Lecture Notes in Computer Science*, pages 168–182, October 2007.



A Scale- and Orientation-Adaptive Extension of Local Binary Patterns

Sebastian Hegenbart[a,*], Andreas Uhl[a], Andreas Vécsei[b]

[a]Department of Computer Sciences,
University of Salzburg, 5020 Salzburg, Austria
[b]St. Anna Children's Hospital,
Medical University Vienna, 1090 Vienna, Austria

Abstract

Local Binary Patterns (LBP) have been used in a wide range of texture classification scenarios and have proven to provide a highly discriminative feature representation. A major limitation of LBP is its sensitivity to affine transformations. In this work, we present a scale- and rotation-invariant computation of LBP. Rotation-invariance is achieved by explicit alignment of features at the extraction level, using a robust estimate of global orientation. Scale-adapted features are computed in reference to the estimated scale of an image, based on the distribution of scale normalized Laplacian responses in a scale-space representation. Intrinsic-scale-adaption is performed to compute features, independent of the intrinsic texture scale, leading to a significantly increased discriminative power for a large amount of texture classes. In a final step, the rotation- and scale-invariant features are combined in a multi-resolution representation.

Keywords: LBP, scale, adaptive, rotation, invariant, scale-space

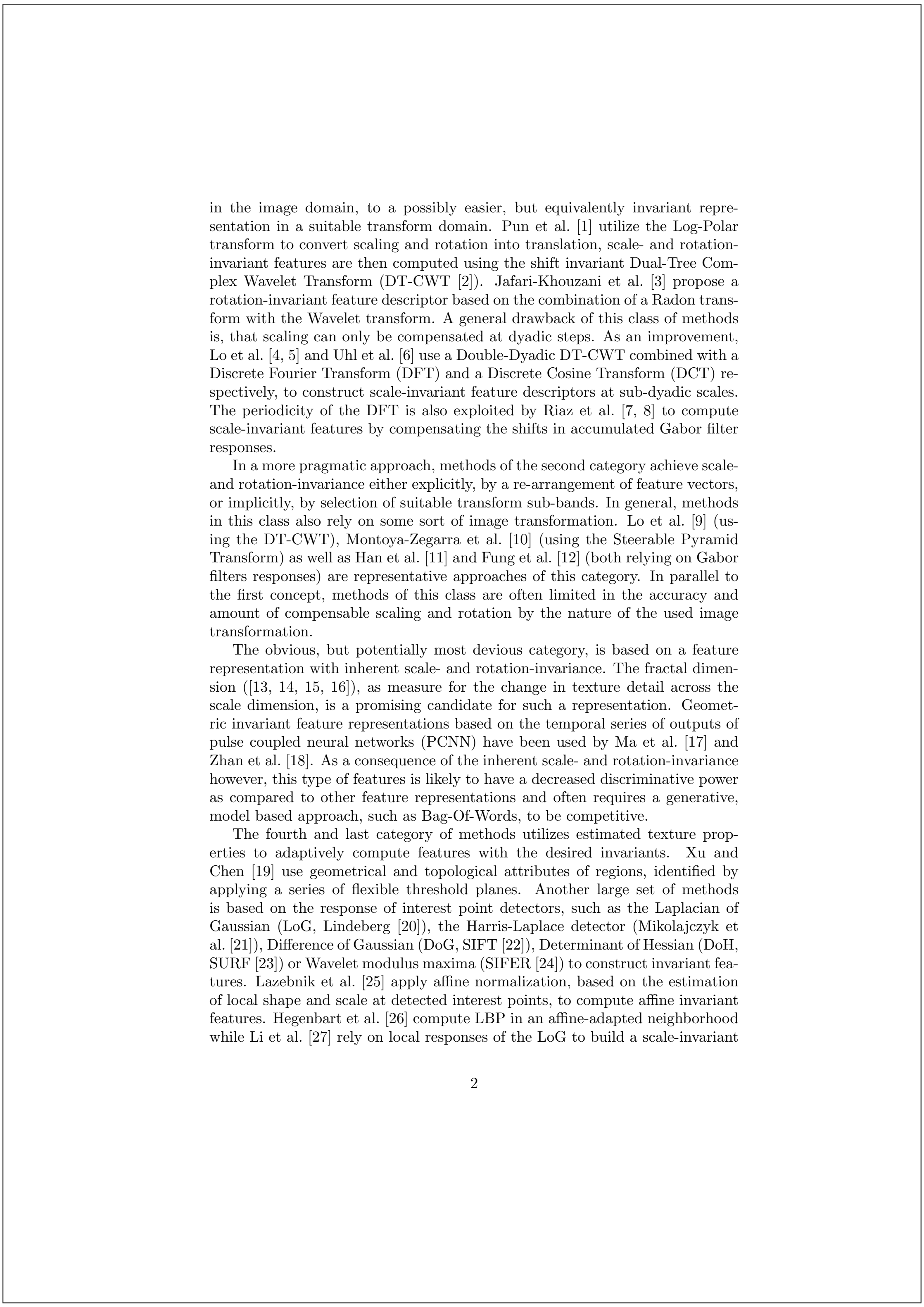
1. Introduction

A major challenge in texture classification is dealing with varying camera-scales and orientations. As a result, research focused on scale- and rotation-invariant feature representations has been a hot topic in the last years. Feature extraction methods providing such invariant representations, allow to be categorized into four conceptually different categories.

In a theoretically elegant approach, methods of the first category transform the problem of representing features in a scale- and rotation-invariant manner

[*]Corresponding Author; Full-Address: Department of Computer Sciences, University of Salzburg, Jakob-Haringer Strasse 2, 5020 Salzburg, Austria; Tel.: (0043) 662 8044-6305, Fax: (0043) 662 8044-172.

Email addresses: shegen@cosy.sbg.ac.at (Sebastian Hegenbart), uhl@cosy.sbg.ac.at (Andreas Uhl), andreas.vecsei@stanna.at (Andreas Vécsei)



in the image domain, to a possibly easier, but equivalently invariant representation in a suitable transform domain. Pun et al. [1] utilize the Log-Polar transform to convert scaling and rotation into translation, scale- and rotation-invariant features are then computed using the shift invariant Dual-Tree Complex Wavelet Transform (DT-CWT [2]). Jafari-Khouzani et al. [3] propose a rotation-invariant feature descriptor based on the combination of a Radon transform with the Wavelet transform. A general drawback of this class of methods is, that scaling can only be compensated at dyadic steps. As an improvement, Lo et al. [4, 5] and Uhl et al. [6] use a Double-Dyadic DT-CWT combined with a Discrete Fourier Transform (DFT) and a Discrete Cosine Transform (DCT) respectively, to construct scale-invariant feature descriptors at sub-dyadic scales. The periodicity of the DFT is also exploited by Riaz et al. [7, 8] to compute scale-invariant features by compensating the shifts in accumulated Gabor filter responses.

In a more pragmatic approach, methods of the second category achieve scale- and rotation-invariance either explicitly, by a re-arrangement of feature vectors, or implicitly, by selection of suitable transform sub-bands. In general, methods in this class also rely on some sort of image transformation. Lo et al. [9] (using the DT-CWT), Montoya-Zegarra et al. [10] (using the Steerable Pyramid Transform) as well as Han et al. [11] and Fung et al. [12] (both relying on Gabor filters responses) are representative approaches of this category. In parallel to the first concept, methods of this class are often limited in the accuracy and amount of compensable scaling and rotation by the nature of the used image transformation.

The obvious, but potentially most devious category, is based on a feature representation with inherent scale- and rotation-invariance. The fractal dimension ([13, 14, 15, 16]), as measure for the change in texture detail across the scale dimension, is a promising candidate for such a representation. Geometric invariant feature representations based on the temporal series of outputs of pulse coupled neural networks (PCNN) have been used by Ma et al. [17] and Zhan et al. [18]. As a consequence of the inherent scale- and rotation-invariance however, this type of features is likely to have a decreased discriminative power as compared to other feature representations and often requires a generative, model based approach, such as Bag-Of-Words, to be competitive.

The fourth and last category of methods utilizes estimated texture properties to adaptively compute features with the desired invariants. Xu and Chen [19] use geometrical and topological attributes of regions, identified by applying a series of flexible threshold planes. Another large set of methods is based on the response of interest point detectors, such as the Laplacian of Gaussian (LoG, Lindeberg [20]), the Harris-Laplace detector (Mikolajczyk et al. [21]), Difference of Gaussian (DoG, SIFT [22]), Determinant of Hessian (DoH, SURF [23]) or Wavelet modulus maxima (SIFER [24]) to construct invariant features. Lazebnik et al. [25] apply affine normalization, based on the estimation of local shape and scale at detected interest points, to compute affine invariant features. Hegenbart et al. [26] compute LBP in an affine-adapted neighborhood while Li et al. [27] rely on local responses of the LoG to build a scale-invariant

LBP representation. Due to the sparse output of interest point detectors and the stability of selected regions, a feature representation derived from interest points, might not be appropriate for all texture classification scenarios however. Even more, the intrinsic-scale of a large number of textures is inappropriate for a directly adapted computation of discriminative features, due to unsuitably large or small scales. As a consequence, the SIFT, SURF and SIFER features descriptors are primarily used for tasks in computer vision apart from texture classification. A variation of these methods without scale-selection, based on local descriptors, computed at a dense grid, is generally used for computing features for the classification of textures.

In this work, we present a methodology which combines ideas from the second (alignment of features) and the last category (scale-adaption) to construct a scale- and rotation-invariant LBP feature representation. The method integrates seamlessly into the general computation of LBP, providing a high angular resolution with a fine grained compensation of scaling. Rotation-invariance is achieved by explicit alignment of features at the extraction level, based on a robust global estimate of orientation, using information provided by multi-scale second moment matrices [28]. The distribution of scale normalized Laplacian responses, in a scale-space representation of an image, allows a reliable estimation of the global image scale, which is used for a scale-adaptive feature computation. Based on the estimation of the global scale, intrinsic-scale-adaption is applied to compute features independent of the intrinsic texture scale. This assures the use of suitable LBP-radii, increasing the discriminative power of the feature representation significantly for a large amount of texture classes. In a final step, the rotation- and scale-invariant features are combined in a multi-resolution representation to further improve the discriminative power.

1.1. Limitations of LBP with Image Scaling and Rotation

The Local Binary Pattern method [29] represents textures as the joint distribution of underlying micro structures, modeled via intensity differences in a pixel neighborhood. Such a neighborhood is defined in relation to a center pixel at position (x, y) as a tuple of n equidistant points on a circle with a fixed radius r. The position of neighbor number k is computed as

$$\eta^{r,n}(k; x, y) \quad = \quad \begin{pmatrix} x + r\cos\left(\frac{2\pi k}{n}\right) \\ y - r\sin\left(\frac{2\pi k}{n}\right) \end{pmatrix}^T_. \tag{1}$$

A weighted sum, representing the pixel neighborhood, is computed and interpreted as binary label, based on a sign function $sg(x)$ mapping to 1 if $x \geq 0$ and 0 else. For a position (x, y) in an image, the standard LBP, based on n neighbors and radius r is computed as

$$\mathbf{LBP}^{r,n}(x, y) \;=\; \sum_{k=0}^{n-1} 2^k \, sg\Big(I\big(\eta^{r,n}(k; x, y)\big) - I(x, y)\Big). \tag{2}$$

Finally, the distribution of patterns is represented by a histogram, which is then used, in conjunction with a meaningful distance function, as an LBP feature.

The LBP feature representation has been used in a wide range of texture classification scenarios and has proven to be highly discriminative. A restriction of LBP however, is its sensitivity to affine transformations. As a consequence of the fixed-scale radius and the fixed sampling area dimension of the pixel neighborhood, the locally computed patterns implicitly encode the underlying micro structures of a texture at a scale directly related to the camera-scale of an image. As a result, the LBP feature representation is unable to compensate for different camera-scales. Even more, a rotation of an image is reflected as a circular shift in the individual patterns, which affects the distribution of patterns in a non-linear fashion. As a consequence, the standard LBP feature representation requires either an implicit or explicit alignment of patterns, which is generally done at the encoding level, to compensate for image rotations.

A widely used rotation-invariant encoding of LBP is based on the work of Ojala and Mäenpää [30]. The authors construct a rotation-invariant representation at the encoding level by implicit alignment of patterns, representing each individual pattern as the minimal decimal interpretation of all possible bitwise circular shifts of that specific pattern. A major limitation of encoding level based approaches is the highly limited angular resolution. As a consequence, Ojala et al. [30] suggest to combine their rotation-invariant encoding with uniform LBP. This combination however, leads to an even smaller number of individual patterns and a possibly decreased discriminative power of the feature representation. In the same work, the authors propose a multi-resolution representation, which improves the discriminative power of the features, by adding the capability of describing underlying micro structures at multiple scales. The multi-resolution representation however lacks a scale-selection mechanism and is therefore unable to compensate for image scaling.

Li et al. [27] were the first to compute scale-adapted LBP, based on the estimation of local texture scale. The authors use a direct mapping from the estimated local texture scale (in terms of the scale-space) to compute scale-adapted LBP-radii. Rotation-invariance is achieved, based on a modification of the rotation-invariant encoding of Ojala and Mäenpää, using bit alignment on the basis of sub-uniform patterns. Unfortunately, using the estimated local image scale as LBP-radius, significantly reduces the reliability of the method. This is a result of computing the features in dependence of the intrinsic texture scale, which is inappropriate for a large number of texture classes (in particular natural textures), due to either very large LBP-radii (low discriminative power) or very tiny LBP-radii (limited possibility of scale-adaption).

The proposed scale- and orientation-adaptive (SOA)-LBP, based on prior work [31, 32], addresses these limitations. The low angular resolution of encoding level based rotation-invariant representations, is significantly improved by alignment of patterns at the extraction level, using a robust estimate of global texture orientation. The reliability of the feature representation is greatly enhanced by the means of intrinsic-scale-adaption, allowing the computation of highly discriminative features, independent of a texture's intrinsic-scale.

2. Scale-Adaptive Local Binary Patterns

We compute a scale-invariant representation of LBP by appropriate selection of LBP-radii (Section 2.2), based on a global estimate for image scale (Section 2.1). To compensate for the changed spatial extent of image structures due to scaling, we perform Gaussian low-pass filtering in reference to the corresponding scale-adapted LBP-radius, to sample neighbors at the correct scale (Section 2.3).

2.1. Estimation of the Global Image Scale

We estimate the global scale of an image utilizing the distribution of scale-normalized Laplacian responses in scale-space. Let $f : \mathbb{R}^2 \mapsto \mathbb{R}$ represent a continuous signal, then the scale-space representation, parametrized in terms of the standard deviation of the Gaussian, $L : \mathbb{R}^2 \times \mathbb{R}_+ \mapsto \mathbb{R}$ is defined by

$$L(\cdot;\sigma) = g(\cdot;\sigma) * f, \tag{3}$$

with initial condition $L(\cdot;0) = f$. We denote $\sigma \in \mathbb{R}_+$ as the scale parameter (the standard deviation of the Gaussian function g) and "$*$" represents a convolution operation. The scale-space family L is the solution to the diffusion equation

$$\partial_\sigma L = \sigma \left(\frac{\partial^2 L}{\partial x^2} + \frac{\partial^2 L}{\partial y^2} \right) = \sigma \triangle L. \tag{4}$$

We construct the scale-space using an exponential spacing of scales $\sigma_i = c\sqrt{2}^{k_i}$, $k_i \in \{-4, -3.75, \ldots, 7.75, 8\}$ and $c = 2.1214$. The value of c acts as a scaling factor and was initially chosen such that the center scale of the representation corresponds to the LBP-radius 3. We later added a set of larger scales to accommodate for the large intrinsic-scales of natural textures. By using an exponential spacing, we provide a fine grained estimation at small scales and still cover a considerable amount of large scales. Note, that as a result of the Gaussian filtering for computing suitable sampling support areas, estimation errors at large scales are not as significant as errors at small scales.

As a consequence of the sparse output of interest point detectors, scale estimation based on such scale-space extrema has shown to be unreliable for a large

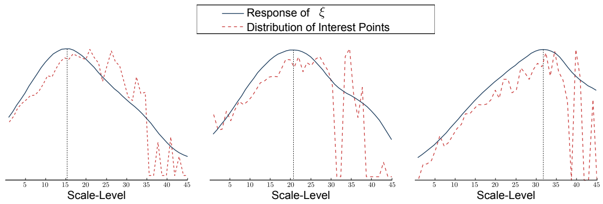


Figure 1: Normalized Response of ξ Compared to the Normalized Response Distribution of Scale-Space Extrema.

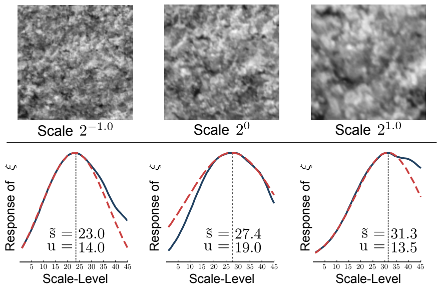


Figure 2: Estimated Scales-Levels $\tilde{s}$ with Uncertainty u for a Texture at Three Camera-Scales.

number texture classes. Figure 1 illustrates this by comparing the response distribution of scale-space extrema with the proposed scale estimation function ξ. It can be observed, that the sparse nature of interest points significantly limits the reliability of the scale estimation.

We therefore use the distribution of the responses of scale-normalized Laplacians in the scale-space representation of an image I, $(\sigma^2\,|\triangle L(\cdot;\sigma)|$, denoted as $\overline{\triangle}I(\cdot;\sigma))$, computed at all scales in the scale-space, to estimate a global image scale. The scale estimation function ξ is

$$\xi(\sigma_i) = \sum_z \overline{\triangle}I(z;\sigma_i), \tag{5}$$

for $z \in \mathbb{R}^2$ corresponding to a Cartesian coordinate on the pixel grid and σ_i denoting a specific scale-level in the scale-space. To determine the global scale of an image, the first local maximum of ξ is searched, which is then used as seed point for a least-squares Gaussian fit. By using the first local maximum we are capable of consistently estimating the scale of textures exhibiting more than a single dominant global scale. The quality of the estimation is improved by using only data points within a certain offset from the seed point. We use 10 percent of the number of scale-levels in the scale-space as positive and negative offset from the estimated first local maximum to fit the Gaussian function. The mean value $\tilde{s}$ of the fitted Gaussian function is interpreted as the dominant level in scale-space. The standard deviation u of the fitted Gaussian is used as uncertainty of the estimation. For a given dominant scale-level in scale-space $\tilde{s_i}$, the spatial scale s_i corresponds to the scale parameter σ_i in $L(\cdot;\sigma_i)$ (the extent of a spatial structure at scale s_i is $\sigma_i\sqrt{2}$). Figure 2 illustrates the determination of a global scale by fitting a Gaussian function (dashed red line) to the scale estimation response function ξ (solid blue line).

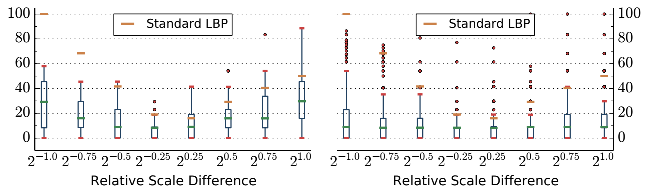
The scale estimation method is reliable for the majority of evaluated images but fails completely for a small fraction (approximately 3%). We identify a failed scale estimation by evaluating the uncertainty u. In our implementation, the scale estimation is considered as failed if u, normalized by the number of scale-levels, is greater than 0.4082 (an empirically found value). In such a case, scale-adapted radii can not be computed reliably. We therefore fall back to a default, computing the standard LBP with a fixed radius. Note that this value (0.4082) was used across all experiments in this work and is assumed to generalize well for a large set of scenarios.

We evaluated the accuracy of the scale estimation for computing scale-adapted LBP-radii, by estimating the global scale of all images in the KTH-TIPS and Kylberg image sets (see Section 5.1) at all 9 scales. Images at the default training scale (2^0) where then used as reference for computing the relative error of scale-adapted LBP-radii compared to the theoretically optimally scale-adapted radius. Figure 3 presents the relative error (in percent) of scale-adapted LBP-radii, compared to the error of a fixed-scale LBP radius.

The results show, that the relative errors of scale-adapted LBP-radii are significantly smaller as compared to the fixed-scale LBP-radius. This indicates that the computation of scale-adapted patterns should improve the scale-invariance of the feature representation. Please note the general asymmetry of the relative error, which can be observed for the fixed-scale LBP radii.

2.2. Intrinsic-Scale-Adaption of the LBP-Radius

Responses of the scale-normalized LoG attain a maximum if its zeros are aligned with a circular shaped image structure. As a consequence, scales estimated based on the LoG, correlate strongly with the scale of the dominant circular shaped structures of a texture. The estimated scale of an image is therefore highly related to an essential property of the texture, the intrinsic-scale. It is critical to realize that the observed scale of an image is always a combination of the intrinsic-scale of the texture and the camera-scale of the image. Scales estimated using the scale-space methodology therefore always represent a combination of the camera- and intrinsic-scale. Estimation of the intrinsic-scale is a highly non-trivial problem. As a consequence of the scale-invariance of the intrinsic-scale, a method for the estimation would be required to be invariant



KTH-TIPS

Kylberg

Figure 3: Relative Error (in Percent) of Scale-Adapted LBP-Radii.

in terms of camera-scale, but sensitive to the intrinsic-scale. Unfortunately we do not know such a method. If so, the intrinsic-scale itself could be a promising candidate to construct a scale-invariant feature representation, similar to the fractal dimension.

An entire category of methods utilizing local texture properties to compute adapted, invariant features (such as Affine Invariant Regions [25] or Li-LBP [27]), are affected negatively by the large variety of intrinsic-scales across texture classes. This is a consequence of using the estimated scale (as combination of the intrinsic- and camera-scale) directly to compute adapted features. Due to unsuitably large or tiny intrinsic-scales for a considerable amount of texture classes, the estimated scales are likely to be inappropriate for computing scale-adapted features. In this work, we propose a method to compute scale-adaptive LBP at suitable and highly discriminative scales by the means of intrinsic-scale-adaption.

We exploit the scale-invariance property of the intrinsic-scale to perform intrinsic-scale-adaption without actual knowledge of the intrinsic-scale. Considering the quotient of two estimated image scales, either the intrinsic-scales cancel each other out (the images are from the same texture class) and the quotient is therefore in terms of the camera-scale, or the intrinsic-scales do not match (images are from different texture classes) and the quotient is basically random. By explicit computation of scale-adapted patterns, based on the quotient between the estimated scale of an image and a trained-base-scale, we are able to adapt for unsuitable intrinsic-scales implicitly.

A trained-base-scale, acting as reference for the computation of intrinsic-scale-adapted patterns, is assigned to each texture class in the training data. In particular, we estimate the scales of each image in the training data and use the median of all estimated scales within a texture class as the trained-base-scale of that class. The scale-adapted LBP-radius used for an image with an estimated scale s, in reference to the trained-base-scale $\bar{s}_l$ of texture class l is then computed as

$$\lambda(s, l, \rho) = \rho \frac{s}{\bar{s}_l}. \tag{6}$$

We define ρ (referred to as base-radius) as the LBP-radius used at the trained-base-scale $\bar{s}_l$. As a trade-off between discriminative power of the representation and the ability of adapting to a large variety of camera-scales, we set $\rho = 3$ as default. Note the linearity of λ as a necessary property for scale-invariance. By computing LBP-radii as a function of the quotient of the estimated image scale and a trained-base-scale, the scale-adaptive representation is independent of the intrinsic-scale of the texture. As a consequence, highly discriminative features at suitable LBP-radii can be computed for a much larger set of texture classes.

Our experiments have shown that scale-adapted LBP computed in reference to a wrong trained-base-scale (the wrong texture class), exhibit appropriately the same intra-class variability as compared to the inter-class variability of features computed at matching trained-base-scales (the correct texture class). This is a direct result of the basically random LBP-radii used to compute scale-adapted patterns in such a case. As a consequence, we distinguish between the computation of training features and evaluation features.

The correct class is obviously known for images in the training data as part of the available ground-truth. We therefore compute training features only in relation to the trained-base-scale of the class of each specific image. Concerning images for evaluation, the class labels are unknown. In this case, features are computed in reference to each texture class, with the corresponding trained-base-scale. During classification, only features computed in reference to the same trained-base-scale are compared (see Section 4).

By using this approach we assure, that features for training will be computed at suitable discriminative scales, close to the base-radius ρ for a majority of images in the training data. Features for evaluation, computed in reference to the correct trained-base-scale (the same class), benefit from intrinsic-scale-adaption, while evaluation features computed in reference to the trained-base-scale of a different texture class are uninformative due to inappropriate (random) LBP-radii and are insignificant for a later classification.

2.3. Adaptive Sampling Support Area Dimension

Scaling of an image changes the spatial extent of textural structures. Therefore the number of pixels covering structural information changes as well. As a consequence, the size of the sampling support area in the LBP neighborhood has to be adapted accordingly. By applying a Gaussian filter, each pixel in the image implicitly encodes information about a circular neighborhood of appropriate spatial scale. The radius of the Gaussian filter for a texture at estimated scale s in relation to a texture class l using base-radius ρ is computed as

$$g_r = \frac{\lambda(s,l,\rho)\,\pi}{n}, \tag{7}$$

for n defining the number of LBP-neighbors. The Gaussian filter coefficients are then computed such that P percent of the mass of the Gaussian function is covered within the interval $[-g_r; g_r]$

$$\begin{aligned}\int_{-g_r}^{g_r} e^{-\frac{x^2}{2{\sigma_g}^2}}\,dx &= P\int_{-\infty}^{\infty} e^{-\frac{x^2}{2{\sigma_g}^2}}\,dx\\ 2\int_{0}^{g_r} e^{-\frac{x^2}{2{\sigma_g}^2}}\,dx &= P\sigma_g\sqrt{2\pi}\\ \sigma_g &= \frac{g_r}{\sqrt{2}\mathrm{erf}^{-1}(P)}.\end{aligned} \tag{8}$$

We chose P to be 0.99 which corresponds to 99% of the mass of the Gaussian function. As the sampling of a Gaussian function with very few sampling points leads to a large error we use the error function (erf) to improve the stability of the computation of the one dimensional Gaussian filters centered at 0

$$G(x;\sigma_g) = \frac{-\mathrm{erf}\left(\frac{x-0.5}{\sigma_g}\right) - \mathrm{erf}\left(\frac{x+0.5}{\sigma_g}\right)}{2}, \tag{9}$$

which are then used in a separable convolution with the analyzed image.

2.4. Computation of Scale-Adapted LBP

The position of LBP-neighbor k, in a scale-adapted computation, in reference to texture class l and an estimated global image scale s, using base-radius ρ with n neighbors is computed as

$$\eta_{l,s}^{\rho,n}(k;x,y) \quad = \quad \begin{pmatrix} x+\lambda(s,l,\rho)\cos\left(\frac{2\pi k}{n}\right) \\ y-\lambda(s,l,\rho)\sin\left(\frac{2\pi k}{n}\right) \end{pmatrix}^T . \tag{10}$$

A Gaussian filter G with the appropriate standard deviation σ_g (see Equation 8) is used to sample neighbors at the correctly adapted spatial scale. Finally, the scale-adapted LBP is computed at position (x,y) with neighborhood $\eta_{l,s}^{\rho,n}$ based on the convolution of image I with G, $(I_g = I * G)$, as

$$\textbf{SA-LBP}_{l,s}^{\rho,n}(x,y) = \sum_{k=0}^{n-1} 2^k \, sg\Big(I_g\big(\eta_{l,s}^{\rho,n}(k;x,y)\big) - I_g(x,y)\Big). \tag{11}$$

The histogram of patterns computed in reference to the trained-base-scale of texture class l is denoted as H_l and added to the SOA-LBP meta-descriptor of the specific image (see Section 4).

3. Orientation-Adaptive LBP

To compensate for the non-linear changes of the LBP distribution caused by a rotation of an image, an explicit or implicit alignment of patterns is required. This is generally performed at the encoding level, leading to a low angular resolution. To improve the angular resolution, we perform pattern alignment at the extraction level, which integrates naturally with the scale-adaptive computation of LBP and is based on an estimate of global image orientation.

3.1. Estimation of the Global Image Orientation

A main requirement on the orientation estimation in the context of scale-adaptive LBP, is robustness to varying image scales. We therefore utilize multi-scale second-moment-matrices (SMM [28]), computed at the global scale of an image, to estimate a global image orientation. The SMM summarizes the predominant directions of the gradient in a specific area of an image. In contrast to the single-scale SMM, the multi-scale SMM is defined over two scale parameters, the local scale σ_i as well as the integration scale i. This allows to estimate the shape of visual structures at appropriate scales, as detected by the scale-estimation algorithm. The integration scale parameter is chosen in relation to the local scale (we use $i = \sqrt{2}\sigma_i$). The local scale parameter is selected as the global scale of the image, using the method described in Section 2.1. The multi-scale SMM of an image at location $z \in \mathbb{R}^2$ is then computed as

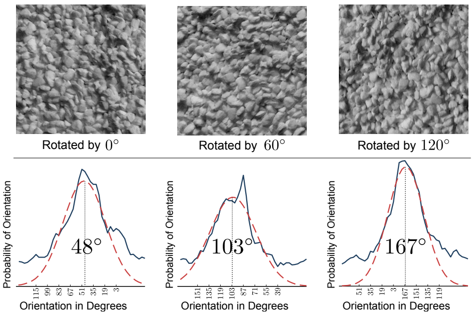


Figure 4: Estimated Orientations for a Texture at Three Orientations.

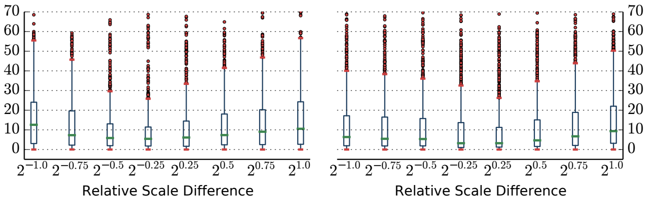
$$\mu(z;\sigma_i,i) = \int_{\xi\in\mathbb{R}^2} \left(\nabla I\right)(z-\xi;\sigma_i)\,\left(\nabla I\right)^T(z-\xi;\sigma_i)\,g(\xi;i)\,d\xi. \tag{12}$$

We denote $\left(\nabla I\right)(z;\sigma_i)$ as the gradient of the scale-space representation of image I at scale σ_i and position z. An important property of SMMs in general, is positive definiteness. The two (non-negative) eigenvalues of an SMM, correspond to the length of the axes of an ellipse (up to some constant factor). The orientation of the eigenvectors correspond to the orientation of the dominant gradient and the orientation perpendicular to the dominant gradient respectively.

To estimate the global orientation of an image I, we compute multi-scale SMMs at a dense grid, corresponding to pixel locations $z \in \mathbb{R}^2$. The orientation at a specific location is determined as the angle between the major axis of the ellipse and the vertical axis of the coordinate system (the axes of the image). Due to the ambiguous orientation of the ellipse, we treat all angles modulus π. Hence, the estimated orientation is unambiguous in $[0;\pi]$. We then estimate the global orientation of an image, based on the distribution of local orientations, computed at all coordinates of the sampled grid.

In parallel to the scale estimation method described in Section 2.1, this is done by fitting a Gaussian function to the distribution of local orientations in a least-squares optimization. To improve the accuracy of the estimation, we remove data points with an offset greater than ±15 degrees from the maximum of the distribution, prior to the fitting process. Finally, the average value of the Gaussian is interpreted as the global orientation, which is used to align the sampling points of the orientation-adaptive LBP.

Figure 4 illustrates the determination of the global orientation from the local orientation distribution. The dashed red line represents the Gaussian function fitted to the distributions of local orientations (solid blue line) of an image at three different orientations. The numbers centered at each figure present the



KTH-TIPS Kylberg

Figure 5: Absolute Errors (in Degrees) of the Orientation Estimation.

estimated global orientation of each image.

To evaluate the accuracy of the orientation estimation method, we computed the absolute error of the estimated orientations (Figure 5) between a reference image at the default training scale (2^0) and the same image at a different scale and random rotation between 30 and 330 degrees in steps of 30 degrees. The error was evaluated from 891 (81*11) random samples at 8 relative scales using the KTH-TIPS as well as the Kylberg image sets (see Section 5.1).

The results indicate that the orientation estimation method is robust in respect of image scaling. We see across all scales, that the medians of the absolute errors are within a range of 5 to 10 degrees. Experiments have shown that the standard multi-resolu¿tion LBP representation can compensate alignment differences of up to 10 degrees, but fails for orientation differences above. In order to improve the orientation-adaptive representation we apply an error compensation technique based on the accumulation of LBP distributions at multiple orientations.

3.2. Orientation Estimation Error Compensation

We found, that a distribution of LBP with a small amount of misaligned patterns (a systematic error) will be dominated by the majority of correctly aligned patterns. As a consequence, we accumulate the distribution of LBP based on multiple orientations within an interval of $\pm\Delta o = 20$ degrees of the estimated global orientation o. Experiments show, that by using this approach an estimated error of up to 20 degrees can be compensated without a significant loss of discriminative power of the feature representation. Figure 6 illustrates this error compensation technique.

To improve the reliability of this scheme, we use thresholding to avoid heavy fluctuation of bits due do interpolation artifacts. The modified sign function $sg(x)$ used in computing the individual patterns therefore requires $x \geq T$ to map to 1. The value of T is selected adaptively based on the Gaussian filtered image I_g, to accommodate for the adapted image properties, as the square root of the standard deviation of all pixel values in I_g.

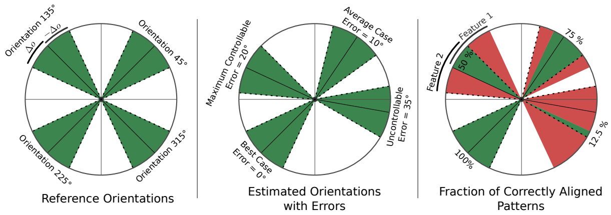


Figure 6: Orientation Estimation Error Compensation using Accumulated Pattern Distributions.

3.3. Computation of Orientation- and Scale-Adaptive LBP (SOA-LBP)

To compute SOA-LBP in reference to a texture class l, estimated global image scale s, global orientation o, base-radius ρ and n neighbors, the position of neighbor k is adapted as

$$\eta^{\rho,n}_{l,s,o}(k;x,y) \quad = \quad \begin{pmatrix} x + \lambda(s,l,\rho)\cos\left(o + \frac{2\pi k}{n}\right) \\ y - \lambda(s,l,\rho)\sin\left(o + \frac{2\pi k}{n}\right) \end{pmatrix}^T . \tag{13}$$

The actual computation of LBP then follows the scheme of the scale-adaptive LBP as depicted in Section 2.4. To accommodate for the ambiguous orientation of multi-scale SMMs, we compute two patterns with initial sample positions at o and $o + \pi$ respectively. Figure 7 illustrates the computation of scale- and orientation-adaptive LBP schematically. The red sampling points indicate the initial sample positions.

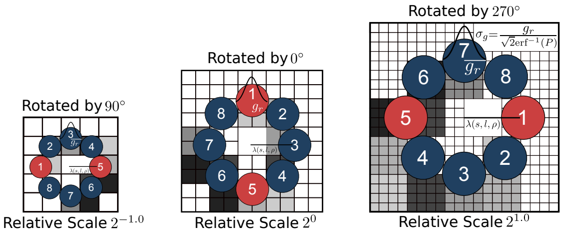


Figure 7: Schematic Computation of Scale- and Orientation-Adaptive LBP.

4. SOA-LBP in a Multi-Resolution Feature Representation

The computation of multiple LBP-features (histograms) per image, each in reference to an individual trained-base-scale, requires the construction of a metafeature-representation for classification. We abstract the set of computed LBP-features per image as a single SOA-LBP meta-descriptor and define a meaningful distance function between a pair of such descriptors. A meaningful distance exists only between LBP-features computed in reference to the same trained-base-scale. As a consequence, we define the distance between LBP-features computed at different trained-base-scales as ∞. Experimentation has shown, that LBP-features computed at incorrectly adapted scales generally yield a significantly higher intra-class variability as compared to LBP-features computed at correctly adapted scales. The distance between two meta-descriptors is therefore defined as the minimum distance between all pairs of LBP-features abstracted by the descriptors. For two SOA-LBP meta-descriptors M_1 and M_2, both representing a set of LBP-features, each computed individually in reference to a texture class in the training data $\{H_1, \dots, H_n\}$, the distance is defined as

$$D(\mathrm{M}_1, \mathrm{M}_2) = \min\{d(H_l, H_k) \mid H_l \in \mathrm{M}_1 \wedge H_k \in \mathrm{M}_2\}, \tag{14}$$

with

$$d(H_l, H_k) \quad = \quad \begin{cases} 1 - \sum_{i=1}^{N} \min\bigl(H_l(i), H_k(i)\bigr), & \text{if } l = k \\ \infty, & \text{if } l \neq k. \end{cases} \tag{15}$$

In our implementation the histogram-intersection is used as a measure for similarity. A notable drawback of using the meta-descriptor abstraction is, that it does not easily integrate with all classification methodologies. We therefore restrict the experimentation in this work to a classification method that allows for a straight forward integration (a standard k-nearest neighbors classifier). Ojala and Mäenpää [30] suggest to compute multiple LBP-features, each at separate fixed LBP-radii, to improve the discriminative power of the feature representation. Multi-resolution LBP-features are then created from a set of standard LBP-features by concatenation.

We combine the rotation- and scale-invariant SOA-LBP in a multi-resolution feature representation, to improve the general discriminative power, by reducing the required amount of low-pass filtering for adapting the sampling area and adding the capability of describing underlying micro structures at multiple scales.

Our experiments have shown, that the discriminative power of the LBP representation starts to decrease at radii greater than 5.44 pixels (this corresponds to LBP-scale 3 in Ojalas multi-resolution approach). We therefore consider radii within the interval $[1; 5.44]$ to be the most discriminative. To compute scale-adaptive patterns at multiple resolutions, we use a set of distinct base-radii for intrinsic-scale adaption $\rho = \{\rho_1, \rho_2, \rho_3\} = \{1.5, 3, 4.5\}$, instead of relying on a single base-radius. Hence, a multi-resolution SOA-LBP representation computed in reference to texture class l consists of the set of SOA-LBP-features computed at each of the base-radii and is denoted as $H_l = \{h_{l,\rho_1}, h_{l,\rho_2}, h_{l,\rho_3}\}$. Considering the small radius $\rho_1 = 1.5$ as well as the large radius $\rho_3 = 4.5$ it is likely that either the lower- or the upper-bound on discriminative LBP-radii

Algorithm 1: Selection of Valid Multi-Resolution Feature Subsets.

```
Data: Let $H_l^1$ and $H_l^2$ be the sets of multi-resolution LBP-features
(histograms) computed in reference to texture class $l$ at the base-radii
$\rho = \{\rho_1, \rho_2, \rho_3\}$ for two images with estimated scales $s_1$ and $s_2$.
$H_l^1 = \{h_{l,\rho_1}^1, h_{l,\rho_2}^1, h_{l,\rho_3}^1\}$ and $H_l^2 = \{h_{l,\rho_1}^2, h_{l,\rho_2}^2, h_{l,\rho_3}^2\}$
Result: Valid subsets $V_1$,$V_2$ of features from $H_l^1$ and $H_l^2$.
$V_1 = H_l^1$ and $V_2 = H_l^2$
foreach $\rho_i \in \rho$ do
    $r_1 = \lambda(s_1, l, \rho_i)$ // intrinsic-scale-adapted LBP-radius of $h_{l,\rho_i}^1$
    $r_2 = \lambda(s_2, l, \rho_i)$ // intrinsic-scale-adapted LBP-radius of $h_{l,\rho_i}^2$
    if $\mathbf{min}(r_1, r_2) < 1$ or
       $\mathbf{max}(r_1, r_2) > 5.44$ or
       $\mathbf{max}(r_1, r_2) \,/\, \mathbf{min}(r_1, r_2) > 3$ then
        $V_1 = V_1 \setminus h_{l,\rho_i}^1$
        $V_2 = V_2 \setminus h_{l,\rho_i}^2$
    end
end
```

is violated for a considerable amount of images, which effectively reduces the discriminative power of the multi-resolution representation. We therefore adaptively select the best subset of SOA-LBP-features for constructing the multi-resolution representation during each computation of the distance between two SOA-LBP meta-descriptors (see Algorithm 1).

Once the best subset of SOA-LBP-features is identified for a pair of meta-descriptors, the final multi-resolution representation is constructed by simple concatenation of the normalized histograms. Note, that as a consequence of considerably different intrinsic-scales, or a failed scale estimation, the possibility of $V_1 = V_2 = \emptyset$ exists. In such a case, it is likely that the two SOA-LBP-features represent different texture classes. We consider such a pair of features as incomparable in a scale-adaptive sense and define the distance as ∞.

5. Experiments

We evaluate the proposed SOA-LBP in reference to a set of scale- and orientation-invariant methods, representative for all categories discussed in Section 1. To assess the reliability of the intrinsic-scale-adaption for a large number of textures, we rely on four different images sets for experimentation. We specifically study the scale-invariance properties (Section 5.4) as well as the effects of combined scaling and rotation (Section 5.5 and 5.6). We finally present a runtime performance analysis of the SOA-LBP (Section 5.7) in relation to the compared methods.

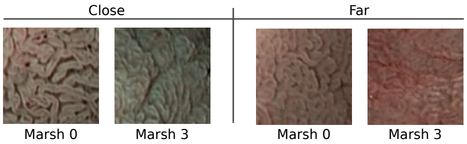


Figure 8: Example Images of the Celiac Set.

5.1. Image Data

We perform the experimentation on four image sets with appropriate characteristics. Table 1 summarizes the most important information about the used data.

Database	**Classes**	**Images per Scale**	**Scales**	**Training Scale**
Celiac	2	102/98	2	Vice Versa
CURET	4	184	2	Mixed
KTH-TIPS	9	81	9	2^0
Kylberg	25	500	9	2^0

Table 1: Information on the Image Sets used for Experimentation.

Celiac. The Celiac image set exhibits duodenal tissue, captured during standard upper endoscopy of patients with indication for celiac disease, using narrow band imaging (NBI [33]), which allows to enhance the contrast of vascular patterns on the mucosal surface. Sub-images of size 128 × 128 pixels, exhibiting regions with particular visual indication for the disease or absence of the disease, were extracted by an expert. As a consequence of the missing scale information, the data was split manually into two distinct sets (near and far), according to camera distance (image scale) by an expert. Due to the nature of endoscopic imagery, the data exhibits a wide variety of different illumination, perspective and scale. The Celiac set represents a two-class classification problem with class Marsh-0 indicating healthy duodenal tissue and class Marsh-3 representing mucosal tissue affected by celiac disease. Figure 8 illustrates representative textures from the Celiac image set.

CURET. The CURET image set contains data with different viewing and illumination conditions. In a four-class classification scenario, textures at two different scales are available as 200 × 200 pixel images. The scale difference of the textures is reported to be approximately 1.7. As a consequence of the significant amount of signal noise in the CURET data, this image set provides an interesting opportunity to evaluate the effects of noise on the proposed method.

KTH-TIPS. The KTH-TIPS [34] image set consists of images from 10 different materials captured at 9 individual relative scales between $2^{-1.0}$ and $2^{1.0}$ with 9 samples per material. Due to the dimension of the original images of material "cracker" (the texture would only fill half of the images at certain scales), we could not use this class for simulating rotations and consequently removed the class in all experiments, leading to a classification scenario with only 9 classes. Sub-images of size 128×128 pixels were extracted from the center of each image to be consistent with the orientation evaluation experiments.

Kylberg. The Kylberg texture set [35] consists of 28 materials captured at a single camera-scale. The data set contains rotated versions of each image at 30 degree steps within a range of 0 to 330 degrees. The large image size (576×576 pixels each) allows to simulate signal scaling without relying on up-sampling, which leads to a reduced amount of unwanted interpolation artifacts. We simulated scaling to match the scales of the KTH-TIPS set such that the scale of the original images is interpreted as the maximum scale $2^{1.0}$ (KTH-TIPS scale 1). Sub-images of size 128×128 pixels were then extracted from the center of the re-scaled images to build the image sets. We created two distinct sets for experimentation, a training set consisting of 20 unique texture patches (types a and b) per material and an evaluation sets comprised of 20 unique texture patches (types c and d) per class. Please note that the texture classes rice1 and rice2 as well as stone1, stone2 and stone3, respectively show minimal visual distinction in textural appearance. As a consequence we removed the texture classes rice2, stone2 and stone3 to improve the interpretability of the experiments, leading to a classification scenario with 25 classes.

5.2. Compared Feature Extraction Methods

We compare the proposed SOA-LBP to a set of methods, representative for the four categories of scale- and rotation-invariant methods, as discussed in Section 1. We believe that the conceptual properties used by these methods will allow us to establish a comprehensive overview. The used methods are
Category I. DT-CWT with Log-Polar Transform (*Log-Polar* [1]).
Category II. Dominant Scale (*Dominant Scale* [10]).
Category III. Fractal Analysis using Filter Banks (*MFS MR8* [13]) and Intersecting Cortical Model (*ICM* [17]).
Category IV. Affine Invariant Regions (*Affine Regions* [25]) and Fisher vector encoding of dense SIFT descriptors (*Dense SIFT* [36]). We also compared the method to a standard, multi-resolution LBP with 3 scales (*LBP* [30]) and the proposed scale-invariant LBP representation of Li et al. (*Li-LBP* [27]).

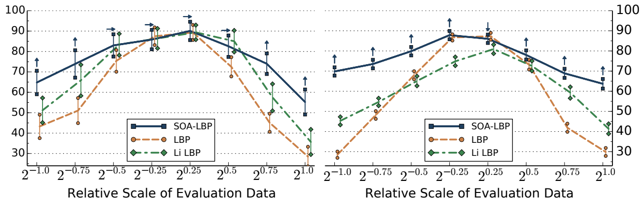
5.3. Evaluation Protocol and Presentation of Results

We implemented the experiments in a scale-constrained cross-validation scheme to accommodate for the rather small size of the Celiac and KTH-TIPS image sets. The scheme is based on two distinct sets for training and evaluation. Images for training were always selected from a fixed scale (the default training scale, see Table 1), while the scales for evaluation varied according to the specific experiment. This approach allows to study the characteristics of each method in reference to signal scaling at various scale differences.

Cross-validation was then performed by an iterated random selection (consistent among all methods) of subsets from the training set (75%) and the evaluation set (25%). A standard k-nearest neighbors classifier was used for classification of features extracted from the specific image subsets. The maximum k-value corresponds to the number of images in each class of the training set (at maximum 20). The reported results represent the mean accuracy over all k-values, averaged in a scale-constrained cross validation with 100 iterations.

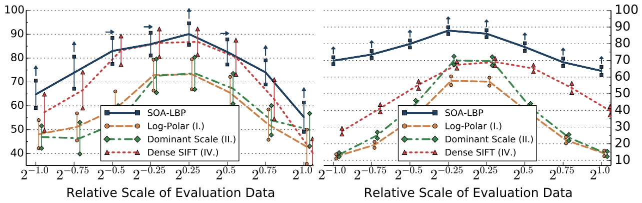
We report statistical significance on a per-figure basis to improve the readability. Two-tailed Wilcoxn rank-sum tests were performed at a significance level $\alpha = 0.001$, to assess the null-hypothesis, that the population median of the cross-validation results obtained with the proposed methodology (SOA-LBP) is equal to the medians of all corresponding methods presented in the specific figure. An arrow pointing upwards ($\uparrow$) indicates, that the null-hypothesis could always be rejected and the SOA-LBP performed significantly better as compared to all corresponding methods in the figure. An arrow pointing to the right ($\rightarrow$) indicates, that the null-hypothesis could not be rejected at least once but no significant difference could be identified. Finally an arrow pointing downwards ($\downarrow$) indicates that at least one method performed significantly better as compared to the proposed method. Note that the markers of each plot are slightly displaced on the x-axis to improve the readability of the error-bars, which represent the standard deviations of the individual cross-validation results.

We present the results based on the CURET and Celiac image sets using asymmetric bar charts (Figures 12 and 15). Each side of a bar represents the classification accuracy of a single experiment. The slope of the bar gives an indication of the scale-invariance of each method. The dashed lines represent the average classification accuracies of both experiments. The arrows indicate statistical significance in relation to the SOA-LBP (e.g. an arrow pointing downwards indicates, that the specific method performed significantly worse as compared to the proposed methodology).



KTH-TIPS Kylberg

Figure 9: Classification Accuracy (y-axis) for Evaluation Scales (Scaling only).



KTH-TIPS

Kylberg

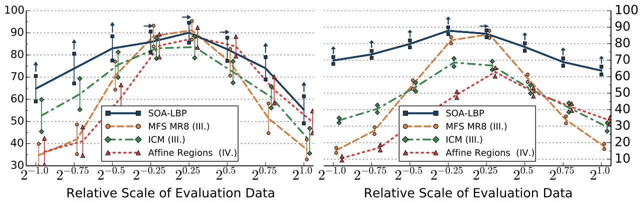
Figure 10: Classification Accuracy (y-axis) for Evaluation Scales (Scaling only).

5.4. Studying the Effects of Image Scaling

The first set of experiments is aimed specifically at studying the characteristics of each evaluated method in regard to image scaling. In these experiments, we only use the scale-invariant representation of methods that allow a selective use of rotation-invariant features. This includes LBP, Li-LBP, SOA-LBP and Dominant Scale. We present the results of the experiments based on the KTH-TIPS, Kylberg and CURET image sets without rotation in Figures 9, 10, 11 and 12. Images at scale 2^0 were used for training, images at all other available scales were used for evaluation (KTH-TIPS and Kylberg).

Based on the CURET data, we follow the experimental setup used by Varma et al. [37]. Two separate training sets were constructed. The first training set consists of textures at both scales, while the second training set is based on textures at a single scale. The evaluation set contains textures at both scales. The difference between the two experiments give an indication for the scale-invariance of each method.

Considering the experiments on the KTH-TIPS image set, we observe that the SOA-LBP performs comparably to the majority of evaluated methods, at evaluation scales close to the training scale. No method performed significantly



KTH-TIPS

Kylberg

Figure 11: Classification Accuracy (y-axis) for Evaluation Scales (Scaling only).

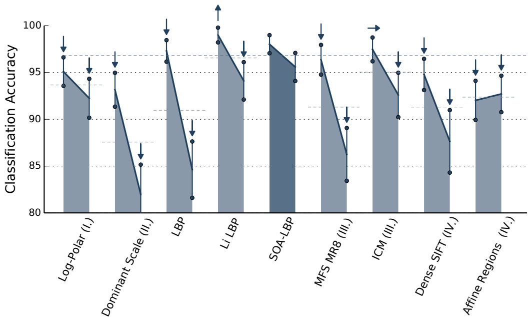


Figure 12: Classification Accuracy of the Experiments on the CURET Data.

better as compared to the proposed methodology however, which indicates that the multi-resolution SOA-LBP feature representation is competitive in scenarios with minimal to no scaling. In case of large scale differences (starting at $2^{0.75}, 2^{-0.75}$) between the training and evaluation data, the SOA-LBP significantly outperforms all evaluated methods.

In parallel to the experiments on the KTH-TIPS data, the SOA-LBPs performance is significantly better as compared to all evaluated methods at large scale differences considering the Kylberg experiments. In contrary to the previous experiments however, this behavior is already recognized at relative scale differences of $2^{0.5}$ and $2^{-0.5}$. The results indicate, that the used multi-resolution representation provides highly discriminative features in the more challenging classification problem provided by the Kylberg set, even at tiny scale differences ($2^{0.25}$, $2^{-0.25}$). The only method that performed significantly better as compared to SOA-LBP was the standard multi-resolution LBP at relative scale $2^{0.25}$, which is caused by a small amount of erroneously estimated image scales of the proposed method. Interestingly, the Li-LBP method performed significantly worse even for small scale differences as compared to the standard LBP as well as the proposed method. We assume this characteristic is caused by the direct mapping from estimated scale to the LBP-radius (the average intrinsic-scale of the Kylberg set is higher as compared to the KTH-TIPS data) in combination with a missing, more powerful, multi-resolution representation.

The experiments on the CURET data indicate a high degree of scale-invariance of the SOA-LBP. Only the Li-LBP method performed significantly better in the experiment without required scale-invariance (mixed training scales). The results on the CURET set show, that the SOA-LBP is suited for classification in noisy scenarios, outperforming the majority of evaluated methods.

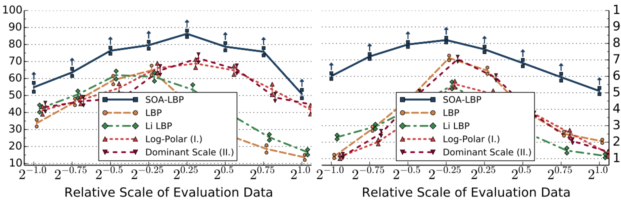
The experiments indicate, that the proposed SOA-LBP provides significantly improved classification accuracies in scenarios with large scale differences. The use of intrinsic-scale-adaption allows the computation of discriminative features for a variety of different textures, while the multi-resolution representation provides highly competitive features even in scenarios with tiny scale differences.

5.5. Studying the Effects of Combined Image Rotation and Scaling

The effects of combined rotation and scaling are studied in the second set of experiments. Feature extraction is based on rotated versions of the Kylberg and the KTH-TIPS image sets. Images at scale 2^0 without rotation were used for training, images at all other available scales were used for evaluation. Subsets of the evaluation sets (KTH-TIPS 891 and Kylberg 1250 images), rotated in steps of 30 degrees, in angles between 30 and 330 degrees, were randomly selected (consistently among all methods) for classification. Only methods providing a scale- and orientation-invariant feature representation where evaluated. LBP was used with the rotation-invariant encoding based on uniform patterns [30]. Li-LBP was used with the proposed sub-uniform patterns [27]. The results are presented in Figures 13 and 14.

We observe, that the rotation of the images decreased the general accuracy of all methods as compared to the previous experiments. The results show the same trends as recognized in the scaling-only experiments however. Again, the proposed SOA-LBP provides significantly improved classification rates at large scale differences between training and evaluation data and performs highly competitive in scenarios with tiny scale differences. The results indicate, that the proposed orientation-adaptive computation is superior as compared to encoding-level based approaches used by LBP and Li-LBP. Interestingly, the Li-LBP method performed worse as compared to the standard LBP method on the Kylberg data even at large scale differences. We assume this is caused by the combination of unsuitable LBP-radii (due to the missing intrinsic-scale-adaption) combined with the less discriminative sub-uniform encoding.

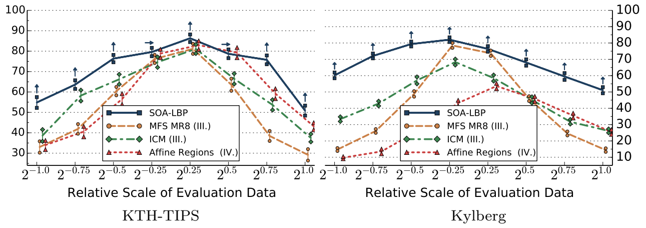
The experiments show that the proposed orientation-adapted computation integrates seamlessly into the scale-adaptive LBP. The results are consistent



KTH-TIPS

Kylberg

Figure 13: Classification Accuracy for Evaluation Scales (Scaling and Rotation).



KTH-TIPS

Kylberg

Figure 14: Classification Accuracy for Evaluation Scales (Scaling and Rotation).

with the previous experiments (scaling only) and indicate that the extraction-level alignment improves the discriminative power of the features.

5.6. Endoscopic Data

We finally study the general capability of all methods to adapt to varying texture scales and orientations, based on a real world problem, the automated diagnosis of celiac disease. The Celiac set exhibits images at two scales with a multiple of different perspectives and orientations. We perform two experiments. The first experiment uses the image set with large camera-scale (close distance to the mucosa) for training and the images with small camera-scale for evaluation. The second experiment is performed vice versa. The results of both experiments, presented in Figure 15 show, that the proposed SOA-LBP again performed reliably in this difficult scenario and was capable of outperforming the majority of methods significantly.

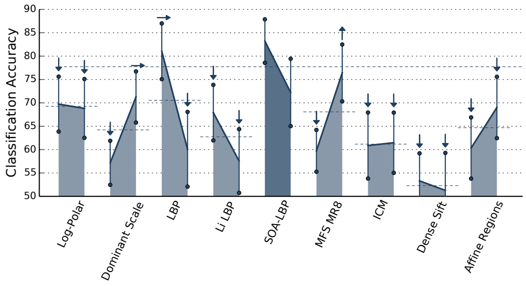


Figure 15: Classification Accuracy of the Experiments on the Celiac Image Set.

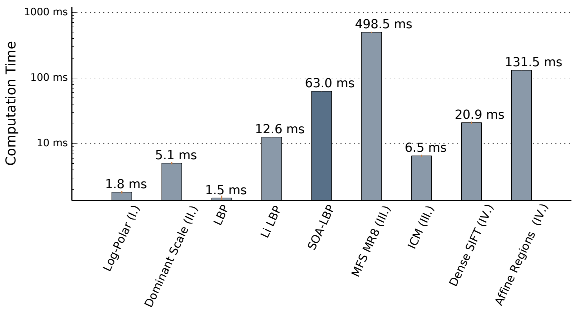


Figure 16: Average Computational Time per Image (KTH-TIPS).

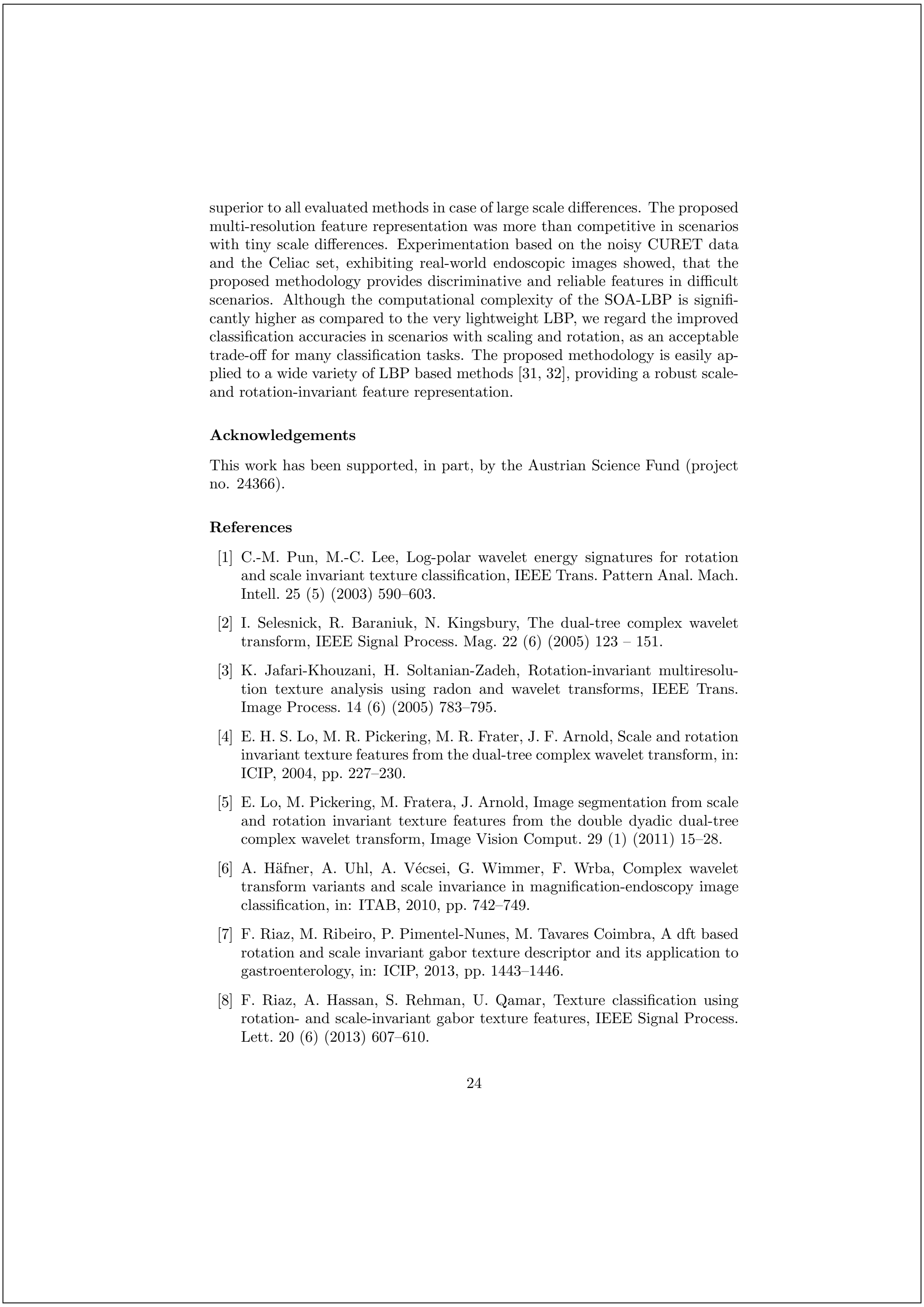
5.7. Runtime Performance Analysis

To study the computational demand of the proposed method, we analyze the required runtime of all considered methods in a multi-threaded Java implementation (JDK 8), running on an Intel i5-2500k processor at 4.29GHz. Due to the nature of the Java programming language (JIT-compilation and garbage collection), we report the computational demand per image as an average of the required computation time for 729 images from the KTH-TIPS data set, in a repeated (20 iterations) experiment (Figure 16). Please note, that the presented performance should not be considered an exact benchmark, as not all methods have undergone equal optimization, but is meant to give the reader an idea of the computational complexity of the proposed methodology.

The results show, that the SOA-LBP is considerably slower as compared to the lightweight LBP or the Li-LBP method, which is caused by the increased demand of computing the scales-space, performing scale- and orientation-estimation and the extra amount of feature computation (performing intrinsic-scale-adaption). Considering the improved classification accuracy in environments with varying scales and orientations however, we think that the average computational demand of 63 ms per image is an adequate trade-off. This is even emphasized as the method ranks in the lower middle range among all methods.

6. Conclusion

We presented a generic methodology to compute a scale- and rotation-invariant feature representation based on LBP, by suitable adaption of the LBP neighborhood. The use of intrinsic-scale-adaption, allowed the computation of features, independent of the intrinsic-scale of textures and increased the reliability of the method significantly. This has been shown in experiments based on four different image sets representing a variety of scenarios. The SOA-LBP was significantly



superior to all evaluated methods in case of large scale differences. The proposed multi-resolution feature representation was more than competitive in scenarios with tiny scale differences. Experimentation based on the noisy CURET data and the Celiac set, exhibiting real-world endoscopic images showed, that the proposed methodology provides discriminative and reliable features in difficult scenarios. Although the computational complexity of the SOA-LBP is significantly higher as compared to the very lightweight LBP, we regard the improved classification accuracies in scenarios with scaling and rotation, as an acceptable trade-off for many classification tasks. The proposed methodology is easily applied to a wide variety of LBP based methods [31, 32], providing a robust scale- and rotation-invariant feature representation.

Acknowledgements

This work has been supported, in part, by the Austrian Science Fund (project no. 24366).

References

[1] C.-M. Pun, M.-C. Lee, Log-polar wavelet energy signatures for rotation and scale invariant texture classification, IEEE Trans. Pattern Anal. Mach. Intell. 25 (5) (2003) 590–603.

[2] I. Selesnick, R. Baraniuk, N. Kingsbury, The dual-tree complex wavelet transform, IEEE Signal Process. Mag. 22 (6) (2005) 123 – 151.

[3] K. Jafari-Khouzani, H. Soltanian-Zadeh, Rotation-invariant multiresolution texture analysis using radon and wavelet transforms, IEEE Trans. Image Process. 14 (6) (2005) 783–795.

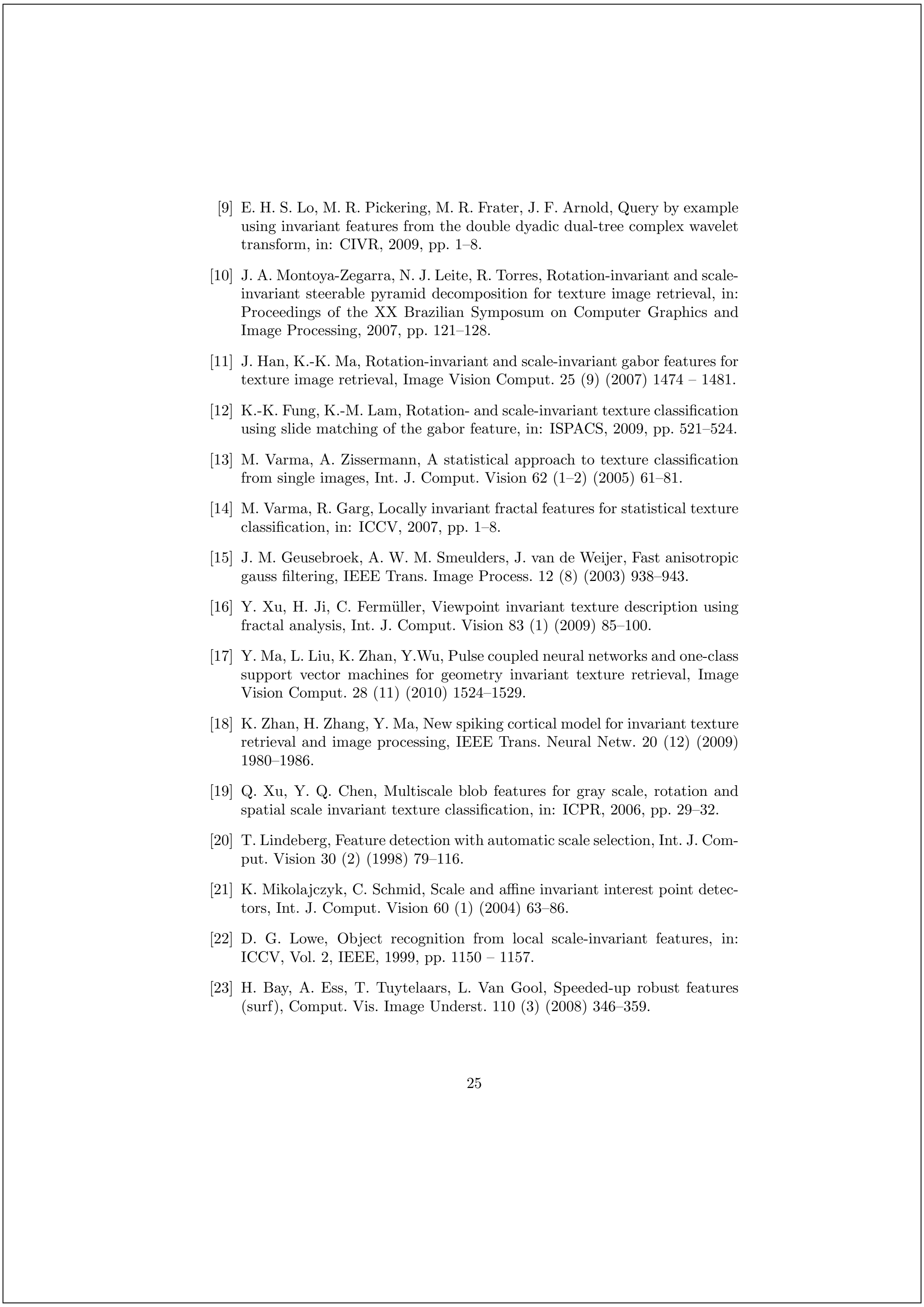
[4] E. H. S. Lo, M. R. Pickering, M. R. Frater, J. F. Arnold, Scale and rotation invariant texture features from the dual-tree complex wavelet transform, in: ICIP, 2004, pp. 227–230.

[5] E. Lo, M. Pickering, M. Fratera, J. Arnold, Image segmentation from scale and rotation invariant texture features from the double dyadic dual-tree complex wavelet transform, Image Vision Comput. 29 (1) (2011) 15–28.

[6] A. Häfner, A. Uhl, A. Vécsei, G. Wimmer, F. Wrba, Complex wavelet transform variants and scale invariance in magnification-endoscopy image classification, in: ITAB, 2010, pp. 742–749.

[7] F. Riaz, M. Ribeiro, P. Pimentel-Nunes, M. Tavares Coimbra, A dft based rotation and scale invariant gabor texture descriptor and its application to gastroenterology, in: ICIP, 2013, pp. 1443–1446.

[8] F. Riaz, A. Hassan, S. Rehman, U. Qamar, Texture classification using rotation- and scale-invariant gabor texture features, IEEE Signal Process. Lett. 20 (6) (2013) 607–610.



[9] E. H. S. Lo, M. R. Pickering, M. R. Frater, J. F. Arnold, Query by example using invariant features from the double dyadic dual-tree complex wavelet transform, in: CIVR, 2009, pp. 1–8.

[10] J. A. Montoya-Zegarra, N. J. Leite, R. Torres, Rotation-invariant and scale-invariant steerable pyramid decomposition for texture image retrieval, in: Proceedings of the XX Brazilian Symposum on Computer Graphics and Image Processing, 2007, pp. 121–128.

[11] J. Han, K.-K. Ma, Rotation-invariant and scale-invariant gabor features for texture image retrieval, Image Vision Comput. 25 (9) (2007) 1474 – 1481.

[12] K.-K. Fung, K.-M. Lam, Rotation- and scale-invariant texture classification using slide matching of the gabor feature, in: ISPACS, 2009, pp. 521–524.

[13] M. Varma, A. Zissermann, A statistical approach to texture classification from single images, Int. J. Comput. Vision 62 (1–2) (2005) 61–81.

[14] M. Varma, R. Garg, Locally invariant fractal features for statistical texture classification, in: ICCV, 2007, pp. 1–8.

[15] J. M. Geusebroek, A. W. M. Smeulders, J. van de Weijer, Fast anisotropic gauss filtering, IEEE Trans. Image Process. 12 (8) (2003) 938–943.

[16] Y. Xu, H. Ji, C. Fermüller, Viewpoint invariant texture description using fractal analysis, Int. J. Comput. Vision 83 (1) (2009) 85–100.

[17] Y. Ma, L. Liu, K. Zhan, Y.Wu, Pulse coupled neural networks and one-class support vector machines for geometry invariant texture retrieval, Image Vision Comput. 28 (11) (2010) 1524–1529.

[18] K. Zhan, H. Zhang, Y. Ma, New spiking cortical model for invariant texture retrieval and image processing, IEEE Trans. Neural Netw. 20 (12) (2009) 1980–1986.

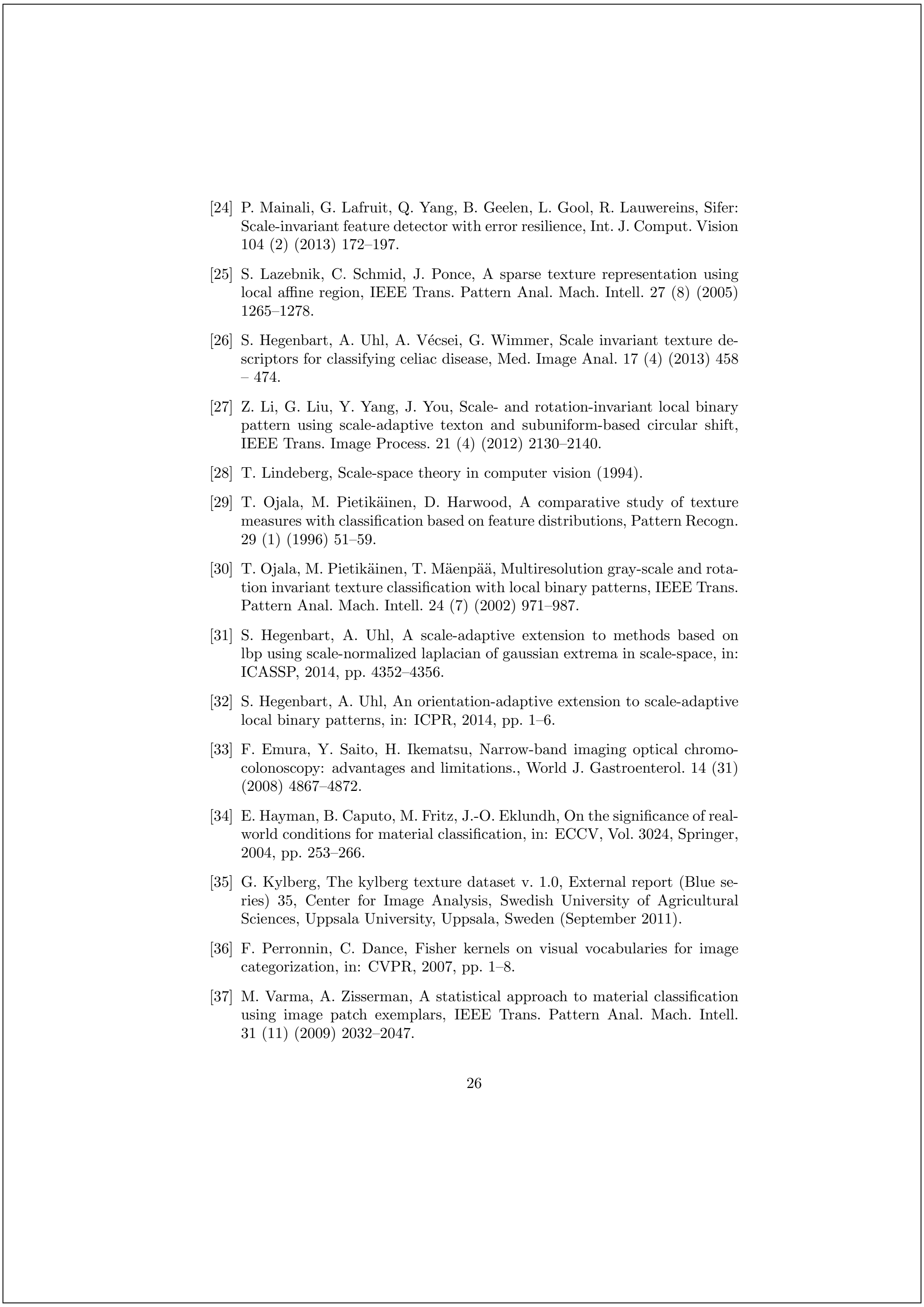
[19] Q. Xu, Y. Q. Chen, Multiscale blob features for gray scale, rotation and spatial scale invariant texture classification, in: ICPR, 2006, pp. 29–32.

[20] T. Lindeberg, Feature detection with automatic scale selection, Int. J. Comput. Vision 30 (2) (1998) 79–116.

[21] K. Mikolajczyk, C. Schmid, Scale and affine invariant interest point detectors, Int. J. Comput. Vision 60 (1) (2004) 63–86.

[22] D. G. Lowe, Object recognition from local scale-invariant features, in: ICCV, Vol. 2, IEEE, 1999, pp. 1150 – 1157.

[23] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), Comput. Vis. Image Underst. 110 (3) (2008) 346–359.



[24] P. Mainali, G. Lafruit, Q. Yang, B. Geelen, L. Gool, R. Lauwereins, Sifer: Scale-invariant feature detector with error resilience, Int. J. Comput. Vision 104 (2) (2013) 172–197.

[25] S. Lazebnik, C. Schmid, J. Ponce, A sparse texture representation using local affine region, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1265–1278.

[26] S. Hegenbart, A. Uhl, A. Vécsei, G. Wimmer, Scale invariant texture descriptors for classifying celiac disease, Med. Image Anal. 17 (4) (2013) 458 – 474.

[27] Z. Li, G. Liu, Y. Yang, J. You, Scale- and rotation-invariant local binary pattern using scale-adaptive texton and subuniform-based circular shift, IEEE Trans. Image Process. 21 (4) (2012) 2130–2140.

[28] T. Lindeberg, Scale-space theory in computer vision (1994).

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

[30] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[31] S. Hegenbart, A. Uhl, A scale-adaptive extension to methods based on lbp using scale-normalized laplacian of gaussian extrema in scale-space, in: ICASSP, 2014, pp. 4352–4356.

[32] S. Hegenbart, A. Uhl, An orientation-adaptive extension to scale-adaptive local binary patterns, in: ICPR, 2014, pp. 1–6.

[33] F. Emura, Y. Saito, H. Ikematsu, Narrow-band imaging optical chromocolonoscopy: advantages and limitations., World J. Gastroenterol. 14 (31) (2008) 4867–4872.

[34] E. Hayman, B. Caputo, M. Fritz, J.-O. Eklundh, On the significance of real-world conditions for material classification, in: ECCV, Vol. 3024, Springer, 2004, pp. 253–266.

[35] G. Kylberg, The kylberg texture dataset v. 1.0, External report (Blue series) 35, Center for Image Analysis, Swedish University of Agricultural Sciences, Uppsala University, Uppsala, Sweden (September 2011).

[36] F. Perronnin, C. Dance, Fisher kernels on visual vocabularies for image categorization, in: CVPR, 2007, pp. 1–8.

[37] M. Varma, A. Zisserman, A statistical approach to material classification using image patch exemplars, IEEE Trans. Pattern Anal. Mach. Intell. 31 (11) (2009) 2032–2047.

Impact of Endoscopic Image Degradations on LBP based Features using One-Class SVM for Classification of Celiac Disease

Sebastian Hegenbart, Andreas Uhl
Department of Computer Sciences
Salzburg University, Austria

Andreas Vécsei
St.Anna Children's Hospital
Vienna, Austria

***Abstract*—The prevalence data of celiac disease have been continuously corrected upwards in the last years. An automated decision support system could improve the diagnosis and safety of the endoscopic procedure. An approach towards such a system is based on a one-class classifier (such as SVM) trained on celiac data only. By doing so, no special treatment of distorted image areas is needed. However, the performance of such a system is highly dependent on the discriminative power of the extracted features within an unconstrained environment such as the human bowel. Towards such a system we evaluate how well methods used in past work perform using a one-class SVM with images exhibiting common endoscopic image degradations such as blur, noise, light reflections and bubbles.**

I. Introduction

Most methods used for texture classification are developed for still images. Modern endoscopes however, transmit an entire stream of frames. For an automated decision support in endoscopic treatments, methods are needed to identify informative frames for classification. The standard approach towards such a system is therefore based on a stage of identification of informative frames followed by segmentation and classification [1]. Hence, the reliability of such as system is based on the quality of the recognition of informative frames.

An alternative approach is based on a one-class classifier (such as SVM) trained on celiac image data. By restricting the classification method to a single class, all frames of an endoscopic image stream can be treated as informative. No method for distinguishing between distorted and informative frames is needed. As a consequence, frames showing either distortions or no celiac specific markers are classified as no celiac and can therefore be ignored for further processing. This approach implicitly combines the informative frame identification with classification.

The accuracy of the second approach is heavily dependent on the discriminative power of the extracted features. The features are now required to be discriminative between the specific classes and to be able to compensate for image degradations caused by endoscopic distortions.

In recent work [2]–[4] we have shown that the automated classification of celiac disease based on endoscopic imagery is feasible using methods based on Local Binary Patterns (LBP) [5]. However, this work has been based on using a constrained image set with high quality [2]. Towards a more realistic scenario, this work is focused on the evaluation of how well methods based on LBP perform using a one-class SVM in an unconstrained environment. We focus on the most prominent types of endoscopic image degradations such as blur, noise, bubbles and specular reflections. In order to be able to assess how the specific methods are affected by certain types and levels of image degradations we simulate the common types of distortions.

In Section I-B we review the common endoscopic image degradations, Section II covers the simulation of the distortions in order to generate a dataset with known ground truth for evaluation. We discuss the details of feature extraction and classification in Section III and the details of experimentation in Section V. Finally, the results are discussed in Section VI while Section VII concludes the paper.

A. Celiac Disease

Celiac disease is a complex autoimmune disorder caused by the introduction of gluten containing materials such as wheat, rye and barley. 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. People with untreated celiac disease, are at risk for developing various associated complications like osteoporosis, infertility and other autoimmune diseases including type 1 diabetes, autoimmune thyroid disease and autoimmune liver disease.

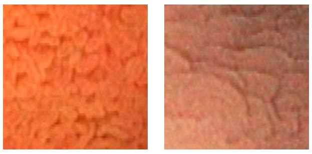
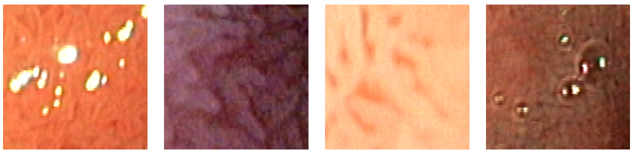

(a) No-Celiac (b) Celiac

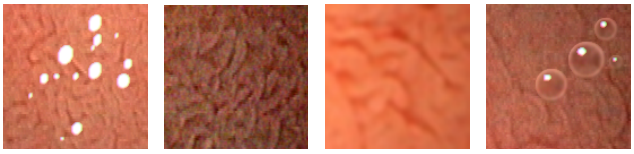
Fig. 1. Examples of Duodenal Image Patches

Figure 1 demonstrates two characteristic images showing healthy mucosal tissue on the left, and the effects of celiac disease on the right. The only treatment is a life long strict



(a) Reflections (b) Noise (c) Blur (d) Bubbles

Fig. 2. Examples of Image Degradations



(a) Reflections (b) Noise (c) Blur (d) Bubbles

Fig. 3. Examples of Simulated Image Degradations

gluten free diet which allows the tissue to heal, leading to a resolution of all symptoms in most cases.

B. Endoscopic Image Degradations

During the endoscopic procedure a small flexible tube (the endoscope) equipped with a camera and a point light source is introduced into the human bowel. The camera has a fixed focus, therefore all areas outside the focal plane appear blurred. Blur is a known problem in all areas of image processing, however in the specific case of classification of celiac disease, blur leads to another associated effect. During the course of the disease the mucosal villi are lost (villous atrophy). The length and form of the villi indicate the severeness of the disease. Depending on the strength of the blur, a healthy mucosa might be misinterpreted as being affected by celiac disease.

The bowel is illuminated using a point light source on the tip of the endoscope. Due to the geometric properties of the bowel a correct exposure can not always be guaranteed. Underexposure leads to an increase of amplifier noise (which is mainly based on thermal noise) within the digital image.

Finally, the third and fourth forms of common degradations are specular reflections and bubbles respectively, visible on the mucosal tissue. Light is reflected by moist tissue, while bubbles build up due to the instillation of water and insufflation of air into the bowel. Figure 2 shows examples of the most common degradations found in endoscopic imagery.

Another type of distortion is the strong lens distortion. The impact of this type of distortion on the automated classification of celiac disease was analyzed in a previous work [6].

II. Simulation of Image Degradations

In order to be able to quantitatively assess the impact of image degradations we need a known ground truth for the level of image degradations. Therefore the four aforementioned types of image degradations are simulated. We perform this simulation on the evaluation set of images as denoted in Table I.

TABLE I
Distribution of Image Data

	Class_0	Class_1	Total
		Images	
Training Set	-	157	157
Evaluation Set	151	149	300

Table I shows the distribution of the used images. Class_0 consist of images showing no villous atrophy (Marsh-0 type), while Class_1 is comprised of images showing mucosal tissue affected by celiac disease (Marsh-3 type). Figure 3 shows examples of simulated distortions.

A. Noise

Amplifier noise is primarily caused by thermal noise. Due to signal amplification in dark (or underexposed) areas of an image, thermal noise has a high impact on these areas. Additional sources contribute to the noise in a digital image such as shot noise, quantization noise and others. These additional noise sources however, only make up a negligible part of the noise and are therefore ignored during this work.

Let P be the set of all pixels in image $I \in \mathbb{N}^2$, $\omega = (\omega_p)_{p\in P}$, be a collection of independent identically distributed real-valued random variables following a Gaussian distribution with mean m and variance σ^2. We simulate thermal noise as additive Gaussian noise with $m = 0$, variance σ^2 for pixel p at position x, y as

$$N(x, y) = I(x, y) + \omega_p, \quad p \in P, \tag{1}$$

with N being the noisy image, for an original image I.

Our image data is extracted from an MPEG2 stream. The compression is based on a discrete cosine transform followed by a quantization step. The characteristics of thermal noise are therefore changed due to the compression. We simulate this effect by applying a low-pass filter (Gaussian filter) to the simulated noise prior to adding it to the original image. Because the lowpass filtering used during compression only affects small high frequency components, this effect is neglected in case of the other types of distortions.

B. Blur

Out of focus blur is one of the most frequent distortions in endoscopic images. Blur is mainly caused by a wrong distance of the camera to the mucosa. Another type of blur is motion blur which is either caused by peristaltic or rapid movement of the endoscope. In this work we only consider out of focus blur. We simulate the point spread function of the blur as a Gaussian

$$f(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \tag{2}$$

which is then convolved with the specific image.

C. Reflections

The light emitted by the endoscope is reflected by moist tissue. Reflections are usually seen as bright spots in certain areas of the image. We model the reflections as ellipsoids with maximum brightness and similar orientation. In natural images color distortions are seen along the contours of the reflections. This effect is caused by the arrangement of the RGB color filters on the CCD chips (Bayer pattern). In this arrangement the color of a single pixel is interpolated using a neighborhood of sensors. Therefore reflections lead to high intensity values in single color channels causing these color distortions along the borders. We simulate this effect by shifting the reflection in each color channel by a single pixel. The borders of reflections are not sharp. We therefore apply a blur to the reflection area to slightly smooth the borders. The mean length of the ellipsoids axes is set to 3 and 4 pixels respectively using a random scale factor with standard deviation of 1.2. This size relates to the common type of small spot reflections seen in endoscopic images. We chose to arrange the reflections such that no overlap happens. This was done to avoid a random factor in the degree of image degradation caused by the arrangement of reflections and does not influence the feature extraction directly.

D. Bubbles

We have shown [2] that the modified immersion technique for capturing images is beneficial to computer aided diagnosis. However, due to the instillation of water and the insufflation of air into the bowel bubbles can build up.

We simulate the appearance of bubbles using a template created in an image manipulation software resembling the visual properties of a bubble. As a simplification, bubbles are treated as circles. This was done to avoid a random factor and should have a negligible effect on the feature extraction. Due to soiled water in the bowel, mucosal tissue covered by bubbles is not clearly visible. We therefore apply a blur to the image area covered by a bubble. Bubbles influence the mucosal appearance instead of having a color. We therefore simulate this effect by considering the luminance component of an image (using the CIELAB color space). By interpolation of the image's luminance information using the bubble template, the mucosal color information is retained while the mucosal appearance resembles natural images containing bubbles.

One or more reflections can be observed on bubble surfaces, this effect is simulated using a single reflection positioned accordingly. These reflections are generated as discussed in Section II-C.

III. Feature Extraction and Classification

We use three LBP-based operators, which have shown to be promising in medical image classification in previous work [4]. The operators are LBP (Local Binary Patterns [5]), ELBP (Extended Local Binary Patterns [7]), and a modified version of the ELBP operator, the ELTP (Extended Local Ternary Patterns) operator [4].

For each color channel three scales (with the meaning of [8]) and filter orientations (in case of the extended LBP based operators: horizontal, vertical and diagonal) are used to compute the distribution of patterns. This results in 9 histograms for LBP and 27 histograms for ELBP and ELTP. For each histogram, only a subset of dominant patterns known as the uniform patterns [9], which make up the majority of discriminative patterns, is used

A. Local Binary Patterns (LBP)

For a radius r and the number of considered neighbors p, the LBP operator is defined as

$$LBP_{r,p}(x,y) = \sum_{k=0}^{p-1} 2^k \, s(I_k - I_c), \tag{3}$$

with I_k being the value of neighbor number k and I_c being the value of the corresponding center pixel. The function s acts as sign function, mapping to 1 if the difference is smaller or equal to 0 and mapping to 0 otherwise. The distribution of patterns is then used as feature for classification. Figure 4 demonstrates the calculation of a pattern.

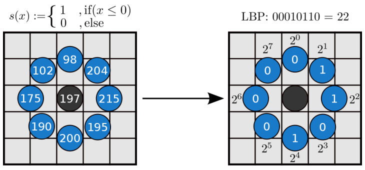
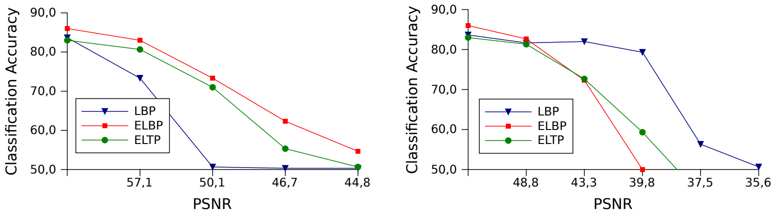


Fig. 4. Example of the Local Binary Pattern Operator

B. Extended Local Binary Patterns (ELBP)

Information extracted by the LBP method from the intensity function of a digital image can only reflect first derivative information. This might not be optimal, therefore Huang et al. [7] suggest using a gradient filtering before feature extraction



(a) Simulated Blur

(b) Simulated Noise

Fig. 5. Impact of Blur and Noise on the Classification Accuracy

and call this operator ELBP or extended LBP. By doing this, the velocity of local variation is described.

C. Extended Local Ternary Patterns (ELTP)

The extended LTP (ELTP) operator is defined consequently in perfect analogy to the ELBP operator. ELTP is based on the LTP (Local Ternary Patterns [10]) operator instead of the LBP operator to suppress unwanted noise in the gradient filtered data. The LTP operator is based on a thresholding mechanism which implicitly improves the robustness against noise. The LTP operator is used to ensure that pixel regions influenced by these kind of distortions do not contribute to the computed histograms. The LTP method is based on a thresholded sign function:

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

The ternary decision leads to two separate histograms, one representing the distribution of the patterns resulting in -1, the other representing the distribution of the patterns resulting in 1. Both histograms are then concatenated and treated as a single histogram.

We apply an adaptive threshold based on the spatial image statistics to 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 which was chosen too high. The calculation is based on an expected value for the standard deviation of the image (β). This value was found based on the training data used during experimentation and represents the average standard deviation of pixel intensity values within all training images. 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. During experimentation we used an α value of 0.05.

$$T_h = \begin{cases} \frac{\beta^{\frac{1}{2}}}{3} + \alpha\sigma, & \text{if } \sigma > \beta \\ \frac{\beta^{\frac{1}{2}}}{3} - \alpha\sigma, & \text{if } \sigma \leq \beta. \end{cases} \tag{5}$$

D. Classification

In this work a one-class Support Vector Machine [11] is used for classification. The classifier was trained using the data within the training set as depicted in Table I. We use parameters found in earlier experimentation. Hence, no further optimization of SVM parameters was performed.

IV. Experiments

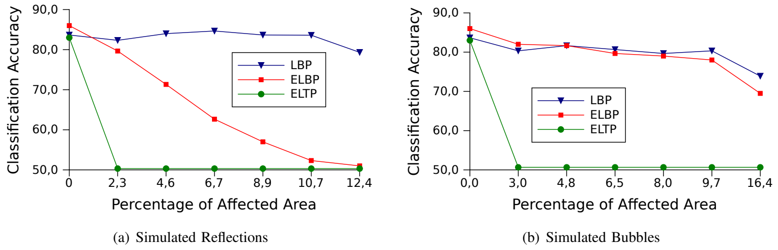
The results of the experiments are presented in Figures 5 and 6. The level of distortion by simulated specular reflections and bubbles is quantified by the percentage of the area of the original image that was affected by the distortion. In case of the reflections, 2.3 percent corresponds to 5 simulated reflections while 12.4 percent correspond to 30 reflections. A single bubble affects approximately 3 percent of the area of an image while 5 bubbles correspond to 16.4 percent. The blur was simulated using Gaussian filters with standard deviations ranging from 0.4 to 0.7. The noise was simulated using standard deviations in steps of 5 ranging from 5 to 25. In order to improve the readability we present the x-axes of Figure 5 labeled with the corresponding PSNR values.

The optimal feature subset for each texture operator was found by using the Sequential Forward Selection (SFS, [12]) algorithm. It must be noted that due to a limited number of image data and the nature of one-class SVM (only a single class to perform cross validation), the SFS algorithm was based on the classification accuracy of the undistorted evaluation set. Therefore the results might be slightly over fitted towards the evaluation data. Overall however, this should not have an impact on the analysis as the same feature set was used for all experiments regarding a specific method.

V. Results

Figure 5 shows how the methods behave when applied to noisy and blurred data. Considering the classification accuracies of the blurred images we already see an effect at a PSNR of 57 which corresponds to a Gaussian filter standard deviation of 0.4. We see that the standard operator (LBP) is most noticeably affected by this type of distortion while both, ELBP and ELTP, are better suited to handle blur.

Considering noise, the standard LBP operator is not significantly influenced by the distortion until a PSNR of 40



(a) Simulated Reflections

(b) Simulated Bubbles

Fig. 6. Impact of Reflections and Bubbles on the Classification Accuracy

(standard deviation of 15). The gradient based operators ELBP and ELTP are able to tolerate noise with standard deviations up to 5 (PSNR of 48.8). ELTP performs slightly better in terms of noise tolerance but both operators fail to achieve reasonable classification accuracies at higher noise levels.

Figure 6 shows the methods' classification accuracies on image data distorted with bubbles and specular reflections. The figures present the classification accuracies in relation to the percentage of the affected image area. Concerning reflections, we see that the accuracy of the standard LBP operator varies insignificantly. The classification rates are stable up to a distorted area of 10 percent. In case of ELBP we see a linear decrease in accuracy considering reflections. For distorted areas larger than 4.6 percent of the original image area the classification accuracy drops below 70 percent. The ELTP method fails completely to classify images distorted with reflections.

In case of simulated bubbles, we see that, in parallel to the specular reflection distortions, LBP is only slightly affected. We do not see classification rates below the 70 percent mark until an affected area of 9.7 percent. It is interesting that in contrary to the simulated reflections, the ELBP operator is as well only mildly affected by this type of distortion. The general behavior is similar to that of the LBP method. In contrast to that, the ELTP again fails completely to classify the distorted images.

VI. Discussion

In general we see from the results that classification in an unconstrained endoscopic environment using LBP based features is feasible. In general blur and noise had the highest impact on the classification accuracy. Especially blur had the most significant impact to classification accuracy. This can in general be explained by the characteristic markers of celiac disease which are lost due to the blur (blurred villi misinterpreted as villous atrophy). We also saw that the specific methods, although all based on LBP, react differently to certain types of image degradations.

A. *Local Binary Patterns (LBP)*

We saw that the LBP operator was heavily affected by blur. LBP considers neighborhoods of pixel intensity values which are all affected by blur. As a consequence, information useful to the method is lost due to this kind of distortion.

The method is suited best to handle noise. The low-frequency part of the noise (caused by quantization of DCT coefficients in the MPEG2 stream) affects entire pixel neighborhoods and therefore does not affect the intra neighborhood intensity values as much as high frequency noise would.

Bubbles and light reflections only had a small impact on the classification accuracy. This can be explained by the small distorted areas that actually affects the LBP operator. In case of reflections only the border of the reflections affect the distribution of patterns (inside the reflections all patterns are 255 which relates to no texture information and is ignored in our implementation). Therefore only a small part of distorted pixel neighborhoods actually affect the LBP method.

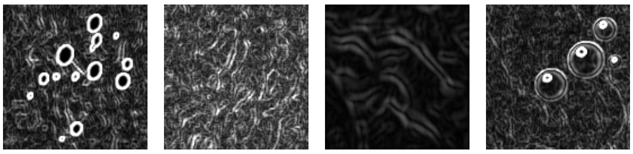
Bubbles had slightly more impact to the classification accuracy as compared to reflections. This is related to the blurred inner part of the bubbles which make up a larger area that negatively affects the operator due to the general information loss.

B. *Extended Local Binary Patterns (ELBP)*

Among all three methods, blur had the least impact on ELBP. Although small scale gradient information is lost due to blur, stronger gradients are retained. Due to the method's invariance in terms of monotonic grayscale changes, blur has a lower impact on ELBP as compared to LBP.

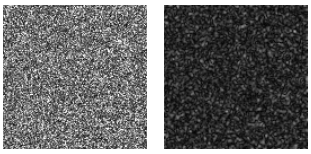
On the other hand, the method was significantly affected by noise. At a noise level of 5, the classification accuracy dropped rapidly. This can be explained by the properties of the low-frequency noise we used. Due to the Gaussian filtering of the noise, entire areas of the degraded image are affected by the same noise level (see Figure 8). Along the boundaries of these areas strong gradients exist which influence the operator's reliability in terms of classification accuracy.

In case of bubbles we see that the decrease in accuracy is almost linear. This is in contrast to the LBP method which was merely affected by this type of distortions and the ELTP operator which failed to classify images with this kind of distortions at the lowest level.



(a) Reflections (b) Noise (c) Blur (d) Bubbles

Fig. 7. Absolute Gradient Values of Distorted Images



(a) High-Frequency (b) Low-Frequency

Fig. 8. Example of Simulated Amplifier Noise

C. Extended Local Ternary Patterns (ELTP)

The ELTP method behaves similar to the ELBP method in case of images distorted with blur. The same also applies to noise with a small improvement as compared to ELBP due to the noise suppression by thresholding. However this improvement is only seen at low classification rates of approximately 70 percent.

In case of reflections and bubbles the method fails to classify images at very small levels of affected areas. This is due to the used thresholding. Considering Figure 7 we see that light reflections and bubbles introduce high-power gradients. In combination with thresholding, the smaller gradients caused by image texture get suppressed to some amount while the high-power gradients introduced by distortions all contribute significantly to the extracted features. This behavior could possibly be relaxed by introducing a second (upper) threshold to eliminate the contribution of high-power gradients.

VII. Conclusion

We evaluated the impact of common endoscopic image distortions using three LBP-based methods for feature extraction. A one-class support vector machines classifier, which was trained using celiac images, was used to classify endoscopic images with simulated distortions. We saw that the specific types of distortions have different effects on the methods.

We saw that distorted images can be accurately classified to some extent. It is interesting to note that bubbles and reflections have lesser impact on the classification rates than expected. Blur, the most common distortion, was best classified using gradient based methods while noise, bubbles and reflections could be handled well by the basic LBP operator.

We conclude, that the unconstrained classification of celiac disease based on LBP using a one-class SVM classifier is feasible to some degree. However, in extreme cases of image distortions an additional step of informative frame identification is unavoidable. By a possible relaxation of the demands on the frame identification method to extreme cases of distortions only, the general reliability could be increased.

We assume that by combining beneficial properties of the evaluated methods a more robust operator could be found to further improve the reliability of classification.

References

[1] M. Liedlgruber and A. Uhl, "A summary of research targeted at computer-aided decision support in endoscopy of the gastrointestinal tract," Department of Computer Sciences, University of Salzburg, Austria, http://www.cosy.sbg.ac.at/research/tr.html, Tech. Rep. 2011-01, 2011.

[2] S. Hegenbart, R. Kwitt, M. Liedlgruber, A. Uhl, and A. Vécsei, "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, Sep. 2009, pp. 718–723.

[3] S. Hegenbart, A. Uhl, and A. Vécsei, "Systematic assessment of performance prediction techniques in medical image classification - a case study on celiac disease," in *Proceedings of the 22nd International Conference on Information Processing in Medical Imaging (IPMI'11)*, Monastery Irsee, Germany, July 2011, accepted.

[4] A. Vécsei, G. Amann, S. Hegenbart, M. Liedlgruber, and A. Uhl, "Automated marsh-like classification of celiac disease in children using an optimized local texture operator," *Computers in Biology and Medicine*, 2011, accepted.

[5] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, January 1996.

[6] M. Liedlgruber, A. Uhl, and A. Vécsei, "Statistical analysis of the impact of distortion (correction) on an automated classification of celiac disease," in *Proceedings of the 17th International Conference on Digital Signal Processing (DSP'11)*, Corfu, Greece, Jul. 2011, accepted.

[7] X. Huang, S. Li, and Y. Wang, "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, 2004, pp. 1–4.

[8] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution Gray-Scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, July 2002.

[9] T. Mäenpää, T. Ojala, M. Pietikäinen, and M. Soriano, "Robust texture classification by subsets of local binary patterns," *Pattern Recognition, International Conference on*, vol. 3, p. 3947, 2000.

[10] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," in *Analysis and Modelling of Faces and Gestures*, ser. LNCS, vol. 4778. Springer, oct 2007, pp. 168–182.

[11] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, p. 2001, 1999.

[12] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 153–158, 1997.

On the Implicit Handling of Varying Distances and Gastrointestinal Regions in Endoscopic Video Sequences with Indication for Celiac Disease

Sebastian Hegenbart, Andreas Uhl
Department of Computer Sciences
Salzburg University, Austria

Andreas Vécsei
St. Anna Children's Hospital
Vienna, Austria

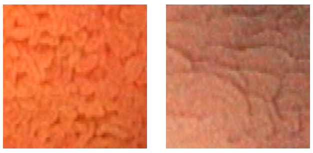
Abstract

We have shown in previous work that problems inherent in the automated diagnosis of standard gastroscopic videos, such as distortion and noise handling, can be handled implicitly, to some extent, by using a one-class support vector machine (SVM) classifier. A video sequence of a standard endoscopic procedure is characterized by rapid changes of perspective towards an inspected area causing various shots at different distances as well as non-predictable transits through gastrointestinal regions. In this work we examine to what extent a one-class support vector machine combined with features based on local binary patterns (LBP) variants can be used to implicitly handle varying camera distances to the mucosa as well as the non-predictable topographical changes during endoscopy.

1. Introduction

The automated diagnosis of endoscopic videos is an emerging area of active research. In the recent past a lot of effort was put into the development of techniques to improve the analysis of sequences captured by using wireless capsule endoscopy. While systems focusing on wireless capsule endoscopy are mainly used to support the analysis of image sequences captured by the capsule, supportive systems (also referred to as computer-aided decision support systems or CADSSs [8]) are focused on assisting a physician during an endoscopic procedure.

We have discussed in previous work [5] that several problems in video classification such as segmentation and distortion handling could implicitly be solved using a one-class support vector machine (SVM). Taking this idea a step further, we investigate two other problems associated with flexible endoscopy. A video of a standard endoscopic procedure is characterized by rapid changes of perspective. In wireless capsule endoscopy, changes in perspectives between consecutive frames are mainly caused by peristaltic combined with a low frame rate. In contrast to that, the rapid changes of perspective during standard endoscopy are caused by the physician maneuvering the flexible endoscope to a desired target within the bowel. On the camera's way to it's desired target a lot of frames at a very far or very close distance to the mucosa as compared to the regular inspection distance are recorded.



(a) No-Celiac (b) Celiac

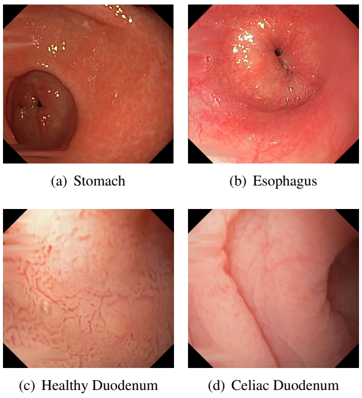
Figure 1. Celiac Endoscopic Images

Another major difference are non-predictable transits through gastrointestinal regions. Although we can expect an endoscopic procedure to start in the esophageal region, following through the stomach into the duodenum, we often see that the endoscope is maneuvered from the duodenum to the stomach and vice versa several times during a single procedure. This complicates topographic segmentation.

Following the idea of [5], we evaluate to what extent those two problems can be handled implicitly by using a one-class support vector machines classifier with LBP based features.

1.1 Celiac Disease

Celiac disease is one of the most common genetically based diseases caused by the introduction of food containing gluten. Prevalence figures for the disease have been constantly corrected upwards in the recent years. A large scale multicenter study by Fasano et al. [3] reports that one in 133 people in the US is affected by the disease. The untreated disease can cause associated complications such as osteoporosis and diabetes. During the course of the dis-


(a) Stomach (b) Esophagus

(c) Healthy Duodenum (d) Celiac Duodenum

Figure 2. Images from Different Gastrointestinal Regions

ease, hyperplasia of the enteric crypts occurs and the mucosa eventually looses its absorptive villi. Once diagnosed, the only treatment is a life long strict gluten free diet which helps the mucosal tissue to heal. Severity of villous atrophy is classified according to the modified Marsh classification in Oberhuber et al. [11] which is based on the scheme proposed by Marsh [10]. In this work we focus on a two-class problem consisting of samples exhibiting healthy mucosal tissue and tissue affected by celiac disease. The severity of the disease of the affected samples ranges from classes Marsh-3A to Marsh-3C. Figure 1 compares healthy tissue with a mucosa affected by celiac disease.

2 Issues with Gastrointestinal Regions and Camera Distance

Besides the handling of endoscopic distortions, the main challenges inherent with standard flexible gastroscopy are rapid changes of perspective and distance to the mucosa as well as multiple non-predictable transits through gastrointestinal regions such as the stomach, the esophagus and the duodenum. Figures 2 and 3 show sample frames from the three gastrointestinal regions that are visible during a gastroscopic treatment as well as a sequence of frames with changing distance to the mucosa.

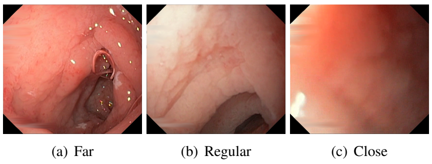

(a) Far (b) Regular (c) Close

Figure 3. Duodenal Sequence with Changing Distance to the Mucosa

2.1 Impact of Camera Distance

The main indications for celiac disease is villous atrophy which reveals itself visible as missing villous structure. Celiac markers besides villous atrophy are a mosaic mucosal texture as well as the visualization of underlying blood vessels. Unfortunately, images recorded from a far distance to the mucosa can be confused with the main indication for the disease. This is due to the bad visualization of structural information from the distance. The same holds for sequences recorded at a close distance to the mucosa. Figures 3 illustrates the problems imposed by bad distances of the endoscopic camera to the tissue.

In previous work [4] we have shown that the modified immersion technique is beneficial to the automated diagnosis of celiac disease. The modified immersion technique described in [2] is based on the instillation of water into the duodenal lumen for better visualization of the villi. The camera is then put into the water to inspect the mucosal tissue. We see in Figure 4 that at the regular distance the camera is put into the instilled water to inspect the tissue. Besides the good visualization of celiac markers, the impact of endoscopic image degradations such as bubbles and reflections is generally reduced by using this technique. Frames exhibiting a farther distance show the mucosa covered with a water film, the camera however is not put into the water. Additionally to the problematic visualization of the tissue at the far distance, endoscopic image degradations are likely to influence the visibility of celiac markers. Finally the close distance faces issues with the visualization of structural changes due to the small field of view combined with a heavily blurred vision due to the suboptimal focus.

2.2 Impact of Upper Gastrointestinal Regions

During a gastroscopic video many frames exhibit tissue from either the esophagus or the stomach. The esophageal tissue usually has a rather smooth texture, sometimes with increased visibility of the underlying blood vessels. In some

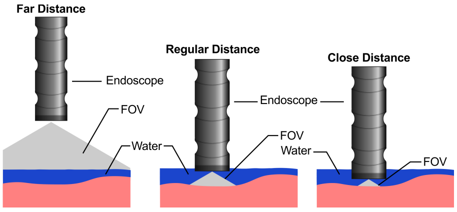


Figure 4. Schematic Figure of Distances

examples we see tissue in the stomach that resembles the nodular tissue of a duodenum affect by the disease. In other cases the stomach tissue is rather flat with a smooth texture. The smooth texture within the esophagus and the duodenum is easily mistaken as villous atrophy and might influence the classification process.

3 Methods

For experimentation we employ feature extraction methods that have been successfully used in previous work [14, 5]. We chose the LBP (local binary patterns) and ELBP (extended local binary patterns) method due to the fact that those methods proved most reliable in terms of noise and distortion handling (bubbles and specular reflections) in [5].

For each RGB color channel three LBP scales (with the meaning of [13]) and filter orientations (in case of the ELBP operator: horizontal, vertical and diagonal) are used to compute the distribution of patterns. This results in 9 histograms for LBP and 27 histograms for ELBP. For each histogram, only a subset of dominant patterns known as the uniform patterns [9], which make up the majority of discriminative patterns, is used.

3.1 Local Binary Patterns

The local binary patterns (**LBP**) [12] operator is used to model a pixel neighborhood in terms of pixel intensity differences. A pixel neighborhood ν is defined in relation to a pixel at (x, y) of the image intensity function f as a sequence of p equidistant points on a circle with radius r around (x, y):

$$\begin{aligned} \varphi_{r,p}(x,y,k) &:= \begin{pmatrix} x + r\cos\left(\frac{2\pi k}{p}\right) \\ y - r\sin\left(\frac{2\pi k}{p}\right) \end{pmatrix}^T \\ (\nu_k) &:= \Big(f\big(\varphi_{r,p}(x,y,k)\big)\Big)_{k\in\{0\,\ldots\,p-1\}} \end{aligned}$$

Based on a sign function a weighted sum is computed and interpreted as binary label according to the specific pixel neighborhood intensity relationship:

$$s(x) \quad := \quad \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}$$

For a position (x, y) the local binary pattern of p neighbors and radius r is computed as:

$$\textbf{LBP} \quad := \quad \sum_{k=0}^{p-1} 2^k\, s(\nu_k - f(x,y)).$$

The joint distributions of these labels are then used to characterize a texture.

3.2 Extended Local Binary Patterns

Information extracted by the LBP-based operators from the intensity function of a digital image can only reflect first derivative information. The extended local binary pattern (**ELBP**, [6]) extends the LBP method and is based on the partial derivatives of the underlying intensity function of a specific image. We utilize different ELBP-neighborhoods (ν) as originally suggested by Huang et al. who used the gradient magnitude:

$$\begin{aligned} (\nu_k^x) &:= \left(\frac{\partial f}{\partial x}(\varphi_{r,p}(x,y,k))\right)_{k\,\in\,\{0\,\ldots\,p-1\}} \\ (\nu_k^y) &:= \left(\frac{\partial f}{\partial y}(\varphi_{r,p}(x,y,k))\right)_{k\,\in\,\{0\,\ldots\,p-1\}} \\ (\nu_k^{xy}) &:= \frac{(\nu_k^x) + (\nu_k^y)}{2} \end{aligned}$$

For a position (x, y) the extended local binary pattern of p neighbors and radius r is computed as:

$$\begin{aligned} \textbf{ELBP}^x &:= \sum_{k=0}^{p-1} 2^k\, s\Big(\nu_k^x - \frac{\partial f}{\partial x}(x,y)\Big) \\ \textbf{ELBP}^y &:= \sum_{k=0}^{p-1} 2^k\, s\Big(\nu_k^y - \frac{\partial f}{\partial y}(x,y)\Big) \\ \textbf{ELBP}^{xy} &:= \sum_{k=0}^{p-1} 2^k\, s\Big(\nu_k^{xy} - \frac{1}{2}\Big(\frac{\partial f}{\partial x}(x,y) + \frac{\partial f}{\partial y}(x,y)\Big)\Big) \end{aligned}$$

The partial derivatives of the signal intensity function are approximated by using Sobel filtering.

3.3 One-Class Support Vector Machine

In this work a one-class support vector machine (**SVM**, [1]) is used for classification. The one-class support vector machine is trained using samples from a single class. The method treats the origin of the feature space as the only initial member of the second class and uses relaxation parameters to separate the image of the trained class from the origin. We use a standard RBF kernel of the form $K(x_i, x_j) = e^{-\gamma||x_i - x_j||^2}$.

	No-Celiac	Celiac	No-Celiac	Celiac
	Images		Patients	
Training-Set 1		157		21
Evaluation-Set 1	151	149	65	19
Training-Set 2		171		40
Regular Evaluation-Set	86	92	74	36
Far Evaluation-Set	61	53	60	22
Close Evaluation-Set	57	50	57	28
	Images		Patients	
Stomach Evaluation-Set	88		88	
Esophagus Evaluation-Set	96		96	

Table 1. Distribution of Image Data

4 Experiments

To perform the experiments we constructed several different data sets. Table 1 shows the number of images and patients used per pathology and specific area of the upper gastrointestinal tract in the experiments. The data sets labeled as “Training-Set 1“ and ”Evaluation-Set 1“ have been used for experimentation in prior work and were created by extracting subimages of size 128×128 from still frames shot during the endoscopic session. Additionally to those data sets we created 6 new data sets to perform the intended experiments in this work.

These data sets were created by extracting suitable frames from video sequences captured during the endoscopic sessions. We also extracted subimages of size 128×128 from those frames. To be able to compare the classification accuracies among the ”Regular“, ”Far“ and ”Close“ sets we searched for suiting sequences showing the same mucosal area at a regular distance as well as either at the far or the close distance or both. The assessment of distance was performed manually based on the visibility of features.

4.1 Interlacing

The endoscopic videos are recorded using an interlaced format. Methods based on local binary pattern are very heavily influenced by interlacing, we therefore apply a deinterlacing filter to the videos prior to extracting frames for feature extraction and classification. We use the Yadif (Yet Another Deinterlacing Filter) as implemented in the FFMPEG [1] software. The algorithm is based on an initial prediction which is obtained by spatial interpolation. Taking the vertical change into consideration as a spatial score the initial prediction is adjusted using intra frame information. Finally the prediction is further refined using temporal information to smooth the pixel intensity variations.

4.2 Parameter Optimization and Feature Subset Selection

The data sets labeled as “Training-Set 1“ and ”Evaluation-Set 1“ were used to optimize certain parameters of both, the classification method as well as the feature extraction methods. This was done to prevent any effects of over-fitting the parameters to the given evaluation data. We chose this approach because performing cross-validation on the training set to optimize parameters is not possible in a one-class support vector machines classification. Please note, that during all experiments the one-class support vector machines were trained on samples from the celiac class exclusively.

The optimized parameters for the support vector machines classification method was the ν parameter which characterizes the fraction of support vectors and outliers used to construct the hyperplane.

The extracted LBP histograms possess a high dimensionality. We therefore use feature subset selection to reduce the feature dimensionality. The applied algorithm for histogram subset selection was the Sequential Forward Selection algorithm (SFS, [7]). The optimization criterion for this algorithm was the overall classification rate (accuracy). The upper bound for the number of selected histograms was 10. To avoid any unwanted over-fitting effects, the SFS algorithm was also applied using ”Evaluation-Set1“ for evaluation with a support vector machine trained on ”Training-Set 1“. The best histogram subset combined with the ν value yielding the best accuracy was then used for further experimentation on the independent data sets.

5 Results

This Section presents the results of the conducted experiments. To discuss the classification results of the experiments considering varying distances, we stick to the standard convention of presenting the sensitivity (true positive rate, the proportion of celiac images which are correctly classified as celiac), the specificity (true negative rate, the proportion of non-celiac images classified correctly as healthy) as well as the accuracy (the overall proportion of correctly classified images). The column labeled as $\boldsymbol{\Delta}$ gives the absolute differences in accuracy between the regular distance with the far and the close distance respectively.

The classification rates of the experiments focused on the impact of gastrointestinal regions are presented as the absolute number of correctly classified images and incorrectly classified images. Additionally we also show the accuracy of correctly classified images in percent.

[1] http://www.ffmpeg.org

5.1 Distance

Table 2 shows the results of the experiments based on the data sets created to evaluate the impact of distance to the mucosal tissue on the classification performance. We see a classification performance of 79.2 and 82.9 percent of the LBP and ELBP method in case of the regular distance. It is noteworthy that the sensitivity of the LBP method is rather low and reached only 68.5 percent.

Compared to the regular distance we observe a considerable drop of accuracy when considering the classification performance of the far distance set. The drop in classification performance of the LBP method is 14.3 percentage points while the ELBP method dropped by 9.8 percentage points. It is interesting that the sensitivity of the ELBP method did not decrease significantly while the sensitivity of the LBP method decreased by 13.8 percentage points. This property holds while the specificity of both methods drop by 16.9 (LBP) and 18.9 (ELBP) percentage points.

Considering the classification results of the close distance set we see that both methods fail to classify the data reliably. The sensitivity is increased to 94 (LBP) and 100 (ELBP) percent respectively while the specificity drops to 14 (LBP) and 3.5 (ELBP) percent.

	Specificity	Sensitivity	Accuracy	Δ
		Regular Distance		
LBP	90.7	68.5	**79.2**	
ELBP	86.1	79.4	**82.6**	
		Far Distance		
LBP	73.8	54.7	**64.9**	-14.3
ELBP	67.2	79.3	**72.8**	-9.8
		Close Distance		
LBP	14.0	94.0	**51.4**	-27.8
ELBP	3.5	100.0	**48.6**	-34.0

Table 2. Classification Results with varying Distances.

5.2 Gastrointestinal Region

Table 3 presents the classification results of the data sets containing images from the stomach and the esophagus. Considering the results of the stomach we see that every second image was misclassified as celiac disease when using LBP features, even more only 33.0 percent of the images were correctly classified when using ELBP based features.

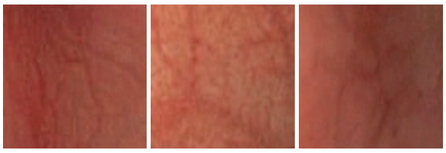
This behavior is also observed for images exhibiting esophageal tissue. The classification accuracy of LBP features drops to 34.4 percent while the accuracy of ELBP based features drops to 16.7 percent.

5.3 Discussion

The results show that the camera distance has a significant impact on the reliability of the classification. The results indicate that the used features are not discriminative enough to distinguish reliably between celiac tissue and tissue recorded at very close distances. The increase in sensitivity is associated with the misinterpretation of heavily blurred areas as villous atrophy by the classification method. Figure 5 compares two patches shot from the regular distance with two patches from a close distance exhibiting a healthy mucosal tissue as well as tissue affected by the disease. We see, that through the very limited field of view and the reduced resolution of the patch, the discriminative power of the textures are lost.

	Correct	Incorrect	Accuracy
		Stomach	
LBP	44	44	**50.0**
ELBP	29	59	**33.0**
		Esophagus	
LBP	33	63	**34.4**
ELBP	16	80	**16.7**

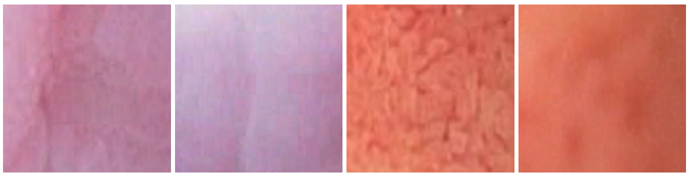
Table 3. Classification Results of Gastrointestinal Regions.



(a) Esophagus (b) Stomach (c) Celiac Duodenum

Figure 6. Comparison of Textures from different Gastrointestinal Regions

Considering shots taken from farther distances we saw that mucosal tissue is classified less reliably as compared to the regular distance. The loss in accuracy however was not as significant as compared to the close distance set. Both the sensitivity and specificity dropped by approximately 15 percentage points for the LBP-based features while the sensitivity of the ELBP-based features stayed constant with a decrease in specificity by approximately 10 percentage points. We therefore consider the classification of far distant shots as feasible to some extent. Referring to the experiments on the stomach and esophageal image set we learn that a high amount of samples is misclassified as celiac disease. In the esophagus this is caused by the missing structure resembling villous atrophy. The same argument holds for the stomach where smooth tissue is misinterpreted as missing villous structure as well. Figure 6 emphasizes this problem by comparing patches from each gastrointestinal region. We clearly see that the visible differences between those three samples is marginal.



(a) Regular Celiac (b) Close Celiac (c) Regular No-Celiac (d) Close No-Celiac

Figure 5. Comparison of Regular Textures with Close Textures

6 Conclusion

We conclude that the used LBP based features are not discriminative enough to describe the properties of villous atrophy in a way that allows to distinguish missing villous structures from villous atrophy. The support vector machine therefore failed to classify images shot from a very close distance. The failure of discrimination indicates that additional effort has to be put into the identification of blurred, close distance shots. Considering the results based on the far image set, we saw that the classification might be feasible under some circumstances. The accuracies of the methods drop significantly however. Mucosal texture from the stomach and the esophagus were very frequently misinterpreted as celiac disease.

Summing up we learned that the problems of varying distances and varying gastrointestinal regions can not be solved implicitly to a satisfactory level by using a one-class support vector machine using features based on local binary patterns. Towards a deployment of such an automated system for diagnosing celiac disease, additional effort has to be put into topographic segmentation as well as the identification of close distance shots.

References

[1] J. S.-T. B. Schölkopf, J.C. Platt and A. Smola. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.

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

[3] A. Fasano, I. Berti, T. Gerarduzzi, T. Not, R. B. Colletti, S. Drago, Y. Elitsur, P. H. R. Green, S. Guandalini, I. D. Hill, M. Pietzak, A. Ventura, M. Thorpe, D. Kryszak, F. Fornaroli, S. S. Wasserman, J. A. Murray, and K. Horvath. 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, February 2003.

[4] S. Hegenbart, R. Kwitt, M. Liedlgruber, A. Uhl, and A. Vécsei. 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)*, pages 718–723, Salzburg, Austria, Sept. 2009.

[5] S. Hegenbart, A. Uhl, and A. Vécsei. Impact of endoscopic image degradations on lbp based features using one-class svm for classification of celiac disease. In *Proceedings of the 7th International Symposium on Image and Signal Processing and Analysis (ISPA'11)*, pages 715–720, Dubrovnik, Croatia, Sept. 2011.

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

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

[8] M. Liedlgruber and A. Uhl. Computer-aided decision support systems for endoscopy in the gastrointestinal tract: A review. *IEEE Reviews in Biomedical Engineering*, 2011. in press.

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

[10] M. Marsh. 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, 1992.

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

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

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

[14] A. Vécsei, G. Amann, S. Hegenbart, M. Liedlgruber, and A. Uhl. Automated marsh-like classification of celiac disease in children using an optimized local texture operator. *Computers in Biology and Medicine*, 41(6):313–325, June 2011.

On the Effects of De-Interlacing on the Classification Accuracy of Interlaced Endoscopic Videos with Indication for Celiac Disease

S. Hegenbart, A. Uhl, G. Wimmer
Department of Computer Sciences
Salzburg University, Austria

A. Vécsei
St. Anna Children's Hospital
Department of Pediatrics
Medical University, Vienna, Austria

Abstract

Interlaced scanning is a technique that has been widely in use to double the perceived frame rate without increasing the used bandwidth. Interlaced scanning is still in use by endoscopic video hardware today. Towards the development of an automated decision support system we focus on the evaluation of the impact of de-interlacing techniques on the accuracy of automated classification of endoscopic video data with indication for celiac disease. In a large experimental setup a variety of de-interlacing methods are evaluated using a set of feature extraction methods from the fields of pattern recognition and medical image analysis.

1. Introduction

Work targeted towards the development and improvement of automated decision support systems is an ongoing research topic. The benefits of such systems are manifold. Besides an improved quality of diagnosis, time and costs can be saved as well. In case of the esophagogastroduodenoscopy (EGD) with indication for celiac disease such a system could improve the targeting of biopsies and therefore both reduce risks and increase the quality of the histological diagnosis.

Interlacing is a method that has been widely in use to double the perceived frame rate without increasing the bandwidth. Cathode ray tube displays (CRT) are natively capable of displaying interlaced formats. Due to the property that interlaced video frames contain data captured at two different times, the development of new display monitor hardware introduced the need for format conversion, in particular the conversion from interlaced to progressive format which is known as de-interlacing. Endoscopic equipment employing interlaced scanning is still in use today. The benefits of applying de-interlacing techniques on interlaced data with regard to classification accuracy has not been investigated so far. This is the main aim of this paper.

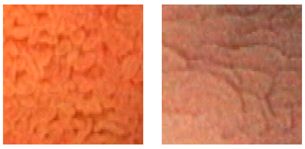

(a) No-Celiac (b) Celiac

Figure 1. Endoscopic Images

The development of de-interlacing methods is focused on visual quality as observed by the viewer of a video sequence often introducing artificial information. More complex de-interlacing techniques incorporate a significant computational effort (such as motion vector estimation), promising better reconstruction quality. It is unclear if the artificially reconstructed data has a positive effect on the classification accuracy at all and if more complex methods have a superior benefit as compared to simpler de-interlacing techniques.

We therefore evaluate these effects by experimentation, using the original interlaced data format as well as the converted progressive format, employing various de-interlacing techniques. To give a comprehensive overview, we use a wide set of feature extraction methods from the field of pattern recognition and medical image analysis. We apply feature extraction methods from the spatial domain with and without special pre-filtering as well as methods based on the wavelet domain.

2. Celiac Disease

Celiac disease is one of the most common genetically based diseases and is caused by the introduction of gluten containing food. Prevalence figures for the disease have been constantly corrected upwards in the recent years. The untreated disease can cause associated complications such as osteoporosis and diabetes. During the course of the disease, hyperplasia of the enteric crypts occurs and the mucosa eventually looses its absorptive villi. Once diagnosed, the only treatment is a life long strict gluten free diet which helps the mucosal tissue to heal. The severity of villous atrophy is classified according to the modified Marsh classification as suggested in Oberhuber et al. [1]. In this work we focus on a two-class problem consisting of samples exhibiting healthy mucosal tissue and tissue affected by celiac disease. The severity of the disease of the affected samples ranges from classes Marsh-3A to Marsh-3C. Figure 1 compares healthy tissue with a mucosa affected by celiac disease.

3. De-Interlacing

De-interlacing methods are used to convert interlaced video signals to progressive video signals. De-interlacing is a trivial task in case of stationary pictures. Most video sequences, and particular endoscopic video sequences, show a wide variety of motion however. In endoscopic videos, motion is caused by the movement of the endoscopic tip by the operating physician as well as peristaltic. De-interlacing methods capable of adapting to motion might therefore be a requirement for successful reconstruction of missing lines. In general de-interlacing methods can be divided into two categories, motion compensated and non-motion compensated techniques. Motion compensated techniques are based on the detection of motion within a set of consecutive frames. Motion compensated techniques use motion vectors computed by motion estimation to improve the de-interlacing results. Motion adaptive techniques can be classified as a sub-category of motion compensated methods. This category of methods usually does not perform any type of motion estimation, instead motion detection is performed. Based on detected motion, the best de-interlacing strategy is then adaptively selected. Non-motion estimated methods, on the other hand, refrain from any sort of motion estimation and use the same set of techniques for all types of scenes (stationary and moving).

Literature in the field of de-interlacing employs the term field to distinguish between the two subfields in a frame. The top-field (or odd-field) is the set of lines with odd numbers. In analogy the bottom-field (or even-field) is the set of lines with even numbers.

3.1 Non-Motion Compensated/Adaptive

- **Bob De-Interlacing**: The bob de-interlacing method is one of the classic methods used for scan line format conversion. It can be described as spatial line doubling, in which lines in each field are doubled. The new line generated can either be just a copy of the previous adjacent line in the field (scan-line duplication mode, denoted as $\mathbf{Bob}_1$ during the Experiments) or computed as an average of the lines above and below the missing line (scan-line interpolation mode, denoted as $\mathbf{Bob}_2$).The bob de-interlacing method retains the horizontal and temporal information at the expense of vertical resolution.
- **Blend De-Interlacing**: The blend method uses temporal information to improve the quality of the reconstructed lines. A temporal frame is build by blending (averaging) the fields of two consecutive frames (we use the current and past frame). Then the bob method is applied to the temporal frame. In parallel to the notation used for the bob method we denote the blend method in scan-line duplication mode as $\mathbf{Blend}_1$ and as $\mathbf{Blend}_2$ in scan-line interpolation mode.

3.2 Motion Adaptive

- **Edge-Based Line Averaging (ELA)**: The ELA method [2] is a motion-adaptive de-interlacing algorithm which is based on an edge-based median filter and an adaptive minimum pixel difference filter. The method utilizes directional correlation between pixels to decide the direction of a linear interpolation. ELA is capable of restoring edges but introduces “salt and pepper” noise if the edge information is interpreted in a wrong way.
- **Hybrid Motion Detection and Edge-Pattern Recognition (HMDEPR)**: The HMDEPR method [3] is a motion-adaptive de-interlacing which employs motion detection to switch between filtering strategies for pixels detected as stationary and pixels detected as part of a motion. The motion detection is based on a comparison of appropriate pixels locations between different consecutive frames. Based on detected motion, two different interpolation strategies are applied. Edge pattern recognition (EPR) is used for intra-field interpolation whereas field insertion is used as inter-field interpolator. The EPR interpolator is especially designed to adapt to textural and edge content and is used to interpolate moving textures.
- **Vertical Temporal Median (VTMedian)**: The three tap vertical temporal median filter [4] implicitly adapts to motion or edges. The interpolated values are found as the median of the vertical neighbors in the same field and the pixel value at the location of the reconstructed pixel in the previous field. The underlying assumption of the method is that in case of a stationary scene, the median is expected to be a value between

the vertical neighbors in the current field, which results in temporal interpolation. In case of motion on the other hand the correlation will be highest between the samples in the current field, therefore intra-field interpolation results.

- **Yet Another De-Interlacing Filter (Yadif)**: The Yadif method is a popular method and implemented in the FFMPEG [1] software. The algorithm is based on an initial prediction which is obtained by spatial interpolation. Taking the vertical change into consideration as a spatial score, the initial prediction is adjusted using intra frame information. Finally the prediction is further refined using temporal information to smooth the pixel intensity variations.
- **Motion Adaptive De-Interlacing with Texture Detection (MATexture)**: The MATexture method [5] classifies missing pixels into four different categories including moving smooth regions, moving texture regions, stationary smooth regions and stationary texture regions. The algorithm performs motion detection by computing pixel differences over three subsequent fields. The texture detection is based on the computation of the variance of the missing pixels and the vertical neighbors. Based on the detection of motion and texture, four de-interlacing methods are selected adaptively which are either a vertical temporal filter for moving smooth regions, a 3D-ELA method for moving texture regions, a median filter for stationary smooth regions and a modified-ELA method for stationary texture regions.

3.3 Motion Compensated

- **Five Field Motion Compensated De-Interlacing Method Based on Vertical Motion (FiveFieldMC)**: The FiveFieldMC method [6] is based on bidirectional motion estimation using two previous and two subsequent fields. Motion compensation is performed using only good lines in the corresponding fields. Depending on the vertical displacement of the motion vector computed for a block of pixels extra information can be reconstructed. The method employs a thresholding mechanism to avoid artifacts due to bad motion vector estimation.

4. Feature Extraction and Classification

To perform the experimental evaluation of de-interlacing methods and their impact on classification accuracy we employ a wide set of feature extraction methods that have been used in medical image analysis and pattern recognition. We assembled a set of methods to cover different aspects of interlaced signal characteristics. We use methods from the wavelet domain, the log-polar domain as well as the spatial domain. Statistical features as well as shape features and pixel based features are used to gain a comprehensive view on the impact of interlacing to the classification accuracy.

4.1 Spatial Domain Methods without pre-Filtering

- **Intersecting Cortical Model (ICM)**: The intersecting cortical model (ICM) [7] is a method derived from Pulse Coupled Neural Networks. Image data from the spatial domain is used as input to the ICM with a series of binary images as output. The entropies of the binary outputs are then used to form a feature vector.
- **Blob Features**: The method [8] uses a series of flexible threshold planes which are computed for an image to construct a set of binary images. Geometrical attributes are then used to describe the image texture. We use the number of identified blobs as well as the shape of the blobs to form a feature vector.
- **Local Binary Patterns (LBP)**: Local Binary Patterns [9] describe a pixel neighborhood in terms of pixel intensity differences. The pixel intensity differences are used to assign a pattern to a neighborhood according to a sign function. The joint distributions of patterns are used as features.
- **Local Ternary Patterns (LTP)**: Local Ternary Patterns [10] are based on Local Binary Patterns and use a modified sign function. The sign function incorporates a thresholding mechanism to improve the invariance to changing illumination conditions.
- **Extended Local Binary Patterns (ELBP)**: Huang et al. [11] suggest using a gradient filtering before feature extraction employing the Local Binary Patterns method. By doing this, the velocity of local variation is described.

4.2 Spatial Domain Methods with pre-Filtering

- **Fractal Analysis**: The fractal dimension gives a measure of how the detail of a pattern changes with the analyzing scale. We compute the local fractal dimension [12] using the Laplacian measure on images prefiltered using the MR8 filter bank [13]. The local fractal dimensions combined with the bag of visual words approach is then used to form the feature vector.

[1] http://www.ffmpeg.org

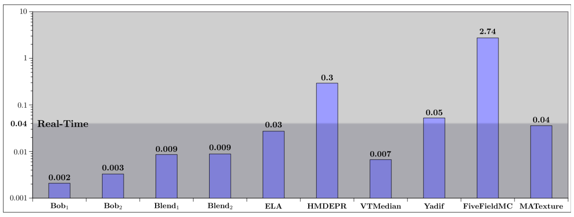


Figure 2. Average Execution Times for a Single Frame

4.3 Statistical Wavelet Domain Methods

- **DT-CWT**: The dual-tree complex wavelet transformation [14] is applied prior to computing statistical features from the resulting subbands. The mean and standard deviations of the absolute subband coefficients are computed and used as features.

- **DT-CWT with DCT**: The method [15] is based on the DT-CWT method. The features are computed by applying the discrete cosine transform DCT across the scale dimension of a feature vector of the DT-CWT.

- **Gabor**: The Gabor Wavelet Transform is used with 5 scales and 6 orientations, the mean and standard deviation of the coefficient magnitudes within a subband are used as features [16].

- **Log-Polar Approach**: The log-polar method [17] transforms the image into the log-polar domain which converts scaling and rotation to translations. The DT-CWT [14] is then applied which is shift invariant. Therefore the DT-CWT features are scale invariant in the log-polar domain. The feature vectors are the subband coefficients' means and standard deviations.

5. Experiments and Results

The data used for experimentation was compiled from 389 interlaced endoscopic video sequences. Subimages with the size of 128×128 pixels exhibiting specific markers were extracted manually from appropriate frames within the video sequences. Only a single subimage was extracted per frame. Table 1 summarizes the distribution of collected subimages.

To perform the experiments we created two distinct sets for training and evaluation. The training set as well as the evaluation set was created by extracting suitable patches from the video streams after conversion to progressive format using the specific de-interlacing technique as indicated in Table 2. We ensured that no images from a patient were within both the training set and the test set. To compare the effects of format conversion to the original interlaced format, we also performed experiments on patches extracted from the original interlaced data, exhibiting lines captured at two different points in time.

The used classifier was a k-nearest neighbor classifier, which was trained on the specific training set. To give a fair and comprehensive overview of the classification results we present the average classification accuracy on the evaluation set with parameter k ranging from 1 to 20.

	No-Celiac	Celiac	No-Celiac	Celiac
	Images		Patients	
Training-Set	122	117	83	26
Evaluation-Set	120	122	82	22

Table 1. Distribution of Image Data

Table 2 lists the average classification accuracies of the conducted experiments. The rows are sorted by feature extraction method. The columns are sorted by the specific de-interlacing method. We present the overall classification accuracy in the column labeled as "Original" which indicates the classification rate of the original interlaced data. We display the increase and decrease of classification accuracy on de-interlaced data as relative percentage points (in relation to the corresponding experiment on the original interlaced data). A positive value indicates an improvement while a negative value indicates a degradation of classification accuracy. Values highlighted in bold indicate that the corresponding classification result is statistically significantly different as compared to the corresponding result based on the interlaced data.We used McNemar's test [18] for evaluating statistical significance, the chosen significance level was $\alpha = 0.01$. The column labeled as "Avg. Δ" lists the average improvement or degradation of the results as compared to the original interlaced data. To put the benefit of de-interlacing into relation we additionally performed an experiment which is based on the interlaced data which was filtered using a Gaussian filter with $\sigma = 1.5$. The results of this experiments are listed in the column labeled as "Gauss" and were not considered for the computation of the

Method	Original	Bob_1	Bob_2	Blend_1	Blend_2	ELA	HMDEPR	VTMedian	Yadif	FiveFieldMC	MATexture	Avg. Δ	Gauss
ICM	79.4	-1.6	**3.9**	**-3.1**	**-6.4**	**4.6**	-1.1	**2.0**	**4.6**	**2.9**	1.7	0.7	**-9.3**
Blob Features	77.0	**7.2**	**6.6**	**3.7**	**4.4**	**7.7**	**-3.2**	**5.3**	**7.3**	**5.8**	**2.6**	4.7	**4.2**
LBP	80.0	**6.5**	**8.3**	1.5	**4.1**	**8.2**	**-9.2**	**6.5**	**7.4**	**9.4**	**4.5**	4.7	**7.1**
LTP	80.3	-1.5	-1.4	**-6.4**	**-11.8**	**-3.0**	**-4.1**	-1.0	**-2.9**	-0.4	**-3.9**	-3.6	**-18.6**
ELBP	75.9	2.3	-0.4	-2.3	**-2.6**	**5.0**	**-14.3**	-1.1	1.7	0.9	1.0	-1.0	**6.5**
DT-CWT	87.5	0.6	0.6	**-3.8**	**-3.9**	0.9	**-2.4**	**1.3**	0.6	0.3	**1.3**	-0.4	**1.5**
DT-CWT DCT	87.3	**1.5**	0.8	**-2.9**	**-4.0**	**1.8**	-1.0	**2.6**	**1.8**	0.6	0.4	0.2	**5.1**
Gabor	86.2	0.3	-0.2	**-3.9**	**-4.1**	-0.3	**-1.4**	-0.1	0.1	-0.1	-0.1	-1.0	**-1.2**
Log-Polar	85.5	-0.5	0.4	**-2.5**	**-3.1**	**1.0**	0.0	0.4	-0.3	0.4	-0.7	-0.5	-0.6
Fractal Analysis	90.4	**-2.1**	-1.3	**-3.8**	**-4.0**	**-1.6**	**-2.3**	-0.8	-0.8	-0.7	-1.1	-1.8	**-3.4**
Average Δ		1.2	1.7	-2.3	-3.1	2.4	-3.9	1.5	2.0	1.9	0.6		-0.9

Table 2. Average Results of the Experimental Evaluation

"Avg. Δ" column.

5.1 Result Discussion: Feature Extraction

It is interesting to see that de-interlacing does not benefit the classification accuracy as much as expected. De-interlacing has a significantly positive effect to only three methods (LBP, Blob Features and ICM). All of these methods are employed in the spatial domain without pre-filtering. Due to this property the interlacing artifacts have a higher impact as compared to other methods. In general, de-interlacing had the most negative effect to LTP however, which is interesting because the method also performs in a pixel based manner in spatial domain. We assume that the thresholding used in LTP, which is computed adaptively based on pixel variance information is causing this behavior. An indication for this is that the blend methods and the Gauss method had the most significant impact on the classification performance using LTP. Both methods lead to a significant decrease of resolution and therefore influence the pixel variance.

Considering the wavelet based methods we see that de-interlacing had a negligible effect in most cases. Interestingly all wavelet methods' accuracies decreased when applied to the de-interlaced data employing the blend methods. The fractal analysis method also could not benefit significantly from de-interlacing. In general the method's performance decreased using de-interlaced data. We assume that the employed MR8 filters have an implicit de-interlacing effect (similar to the Gaussian filter) negating the positive effects of scan line format conversion.

5.2 Result Discussion: De-Interlacing

In general we see that a majority of de-interlacing methods could increase the average classification accuracy slightly. Unfortunately the increase is insignificant. We see that only the ELA method had a significant overall positive effect on the classification accuracy. The only feature extraction methods with decreasing results using ELA were LTP and fractal analysis, both methods did not perform well on the de-interlaced data. On the other hand, by using the blend methods, the classification accuracy significantly decreased for a majority of methods. This behavior can be explained by the decrease of spatial resolution. The HMDEPR method showed a significant negative effect on the LBP based methods. We suppose that the used edge-pattern recognition algorithm and the adaptive interpolation schemes used in HMDEPR have a negative effect on the pixel neighborhoods employed by LBP based methods. We observe that the motion adaptive and motion compensated methods do not have a superior benefit as compared to non-motion compensated/adaptive methods. This might be a result caused by the different environmental constraints posed by endoscopic video sequences as compared to the type of video sequences the methods were developed for. Comparing the effects of applying the de-interlacing methods to the effects of using a pre-filtering with a Gaussian filter we note that the effects are comparable. Interestingly the Gaussian filtering method performs better compared to the blend methods.

5.3 Execution Performance

Figure 2 presents the execution performance of the evaluated de-interlacing methods. Please consider that all methods were implemented in Java with performance in mind. However, no full optimization of the code was performed. Usually de-interlacing methods are implemented in languages compiling into native code with a lot of low level optimization, increasing the performance heavily. Therefore the execution times should only be considered as a reference for the reader to compare the complexity between the methods. We see that a majority of methods are capable of executing in real-time in our Java implementation (we define real-time capability as the ability of handling 25 frames per second). Only the motion compensated method FiveFieldMC, the Yadif method and the HMDEPR method were above that threshold. We see that the motion estimation used by FiveFieldMC is very expensive in terms of computational effort. Considering the effects on the classification accuracy of de-interlacing methods, expensive methods do not provide any additional gain and can be substituted by simpler and faster methods without the general loss of classification accuracy.

6. Conclusion

We saw that de-interlacing does not have a significant positive effect on the classification accuracy of endoscopic data with indication for celiac disease in general. Even more the blend type methods led to a significant decrease of classification accuracy. The effects of de-interlacing are comparable to the effects of applying a Gaussian filter.

We note that de-interlacing only had an overall benefiting effect on methods based in the spatial domain without pre-filtering. The effects of de-interlacing on methods based on the wavelet transform were negligible. Pre-filtering of image data such as applying the MR8 filter bank led to an implicit form of de-interlacing and corresponding methods did not gain any additional benefit from de-interlacing.

We have also shown that complex methods employing motion estimation and motion compensation do not have a superior benefit as compared to simpler de-interlacing methods. A de-interlacing method especially designed for this type of videos might improve the classification accuracy. However, considering the computational complexity of the applied motion compensated methods, we suggest to use faster and simpler methods instead. We suggest to apply de-interlacing on endoscopic data with indication for celiac disease only when using features from the spatial domain without pre-filtering. The blend type methods should be avoided in general.

Our results show that an automated decision support system based on endoscopic video data can operate on interlaced data without significant loss of accuracy.

References

[1] M. Annegarn, T. Doyle, P. Frencken, and D. V. Hees. Video signal processing circuit for processing an interlaced video signal. *Patent*, 1(US 4740842), April 1988.

[2] A. Häfner, A. Uhl, A. Vécsei, G. Wimmer, and F. Wrba. Complex wavelet transform variants and scale invariance in magnification-endoscopy image classification. In *Proceedings of the 10th International Conference on Information Technology and Applications in Biomedicine (ITAB'10)*, Corfu, Greece, Nov. 2010.

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

[4] G. G. Lee, M.-J. Wang, H.-T. Li, and H.-Y. Lin. A motion-adaptive deinterlacer via hybrid motion detection and edge-pattern recognition. *EURASIP Journal on Image and Video Processing*, 2008, Jan. 2008.

[5] S.-G. Lee and D.-H. Lee. A motion-adaptive de-interlacing method using an efficient spatial and temporal interpolation. *IEEE Transactions on Consumer Electronics*, 49(4):1266–1271, Nov. 2003.

[6] Y. Ma, L. Liu, K. Zhan, and Y.Wu. Pulse coupled neural networks and one-class support vector machines for geometry invariant texture retrieval. *Image and Vision Computing*, 28(11):1524–1529, 2010.

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

[8] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2-3):153–157, 1947.

[9] H. M. Mohammadi, P. Langlois, and Y. Savaria. A five-field motion compensated deinterlacing method based on vertical motion. *IEEE Transaction on Consumer Electronics*, 53(3):1117–1124, Aug. 2007.

[10] G. Oberhuber, G. Granditsch, and H. Vogelsang. The histopathology of coeliac disease: time for a standardized report scheme for pathologists. *European Journal of Gastroenterology and Hepatology*, 11:1185–1194, nov 1999.

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

[12] C.-M. Pun and M.-C. Lee. Log-polar wavelet energy signatures for rotation and scale invariant texture classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):590–603, May 2003.

[13] I. Selesnick, R. Baraniuk, and N. Kingsbury. The dual-tree complex wavelet transform. *Signal Processing Magazine, IEEE*, 22(6):123 – 151, Nov. 2005.

[14] Y. Shen, D. Zhang, Y. Zhang, and J. Li. Motion adaptive deinterlacing of video data with texture detection. *IEEE Transaction on Consumer Electronics*, 52(4):1403–1408, Nov. 2006.

[15] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. In *Analysis and Modelling of Faces and Gestures*, volume 4778, pages 168–182, October 2007.

[16] A. Uhl, A. Vécsei, and G. Wimmer. Fractal analysis for the viewpoint invariant classification of celiac disease. In *Proceedings of the 7th International Symposium on Image and Signal Processing (ISPA 2011)*, pages 727 –732, Dubrovnik, Croatia, Sept. 2011.

[17] M. Varma and A. Zissermann. A statistical approach to texture classification from single images. *International Journal of Computer Vision (IJCV)*, 62(1–2):61–81, Oct. 2005.

[18] Q. Xu and Y. Q. Chen. Multiscale blob features for gray scale, rotation and spatial scale invariant texture classification. In *Proceedings of 18th International Conference on Pattern Recognition (ICPR)*, volume 4, pages 29–32, Sept. 2006.

Impact of Histogram Subset Selection on Classification using Multiscale LBP-Operators

Sebastian Hegenbart[1], Andreas Uhl[1], Andreas Vécsei[2]

[1]Department of Computer Sciences, University of Salzburg
[2]St. Anna Children's Hospital, Vienna
shegen@cosy.sbg.ac.at

Abstract. Multiscale Local Binary Pattern based operators are used to extract features from duodenal texture patches with histological ground truth in case of pediatric celiac disease. The multiscale LBP combined with color channels and possibly other filters lead to a high number of computed histograms. The impact of histogram subset selection on the overall classification rates using two feature subset selection algorithms (SFS and SBS) with three LBP-based operators is analyzed and the applicability of these techniques validated.

1 Introduction

Celiac disease is a complex autoimmune disorder in genetically predisposed individuals of all age groups after introduction of gluten containing food. 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. People with untreated celiac disease are at risk for developing various complications like osteoporosis, infertility and other autoimmune diseases including type 1 diabetes. Endoscopy with biopsy is currently considered the gold standard for the diagnosis of celiac disease. 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. [1]. The benefits of an automated support tool for diagnosis are many. Among them are an improved reliability of diagnosis, supported targeting of biopsies and more efficient use of time and manpower.

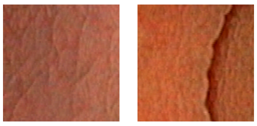
The Local Binary Pattern (LBP) operator is invariant to monotonic intensity variations which is beneficial to texture classification in environments with varying lighting conditions. This property makes the method interesting for classifying endoscopic images. In the context of LBP many modifications and related operators have been suggested over the years. A prominent modification that is often neglected across the literature is the multiscale approach suggested by Mäenpää [2]. This approach is based on low-pass filtering combined with appropriate filter sizes and operator radii to improve the operators' spatial support area. Using this extension, combined with color channels and possibly other filters, the number of computed histograms is considerably higher than the common approach of using a combination of two or three different parametrizations

	Image-Set 1			Image-Set 2		
	Class_0	**Class_1**	**Total**	**Class_0**	**Class_1**	**Total**
Bulbus Duodeni	153	120	273	187	70	257
Pars Descendens	132	164	296	115	58	173

Table 1. Distribution of Image Data

of the operator. It is unclear how a high number of histograms affects the classification. To study the effects we use two feature subset selection schemes to find optimal suitable combinations of histograms. We analyze the impact of the subset selection and validate the applicability of these techniques using two distinct image sets.

2 Materials and Methods



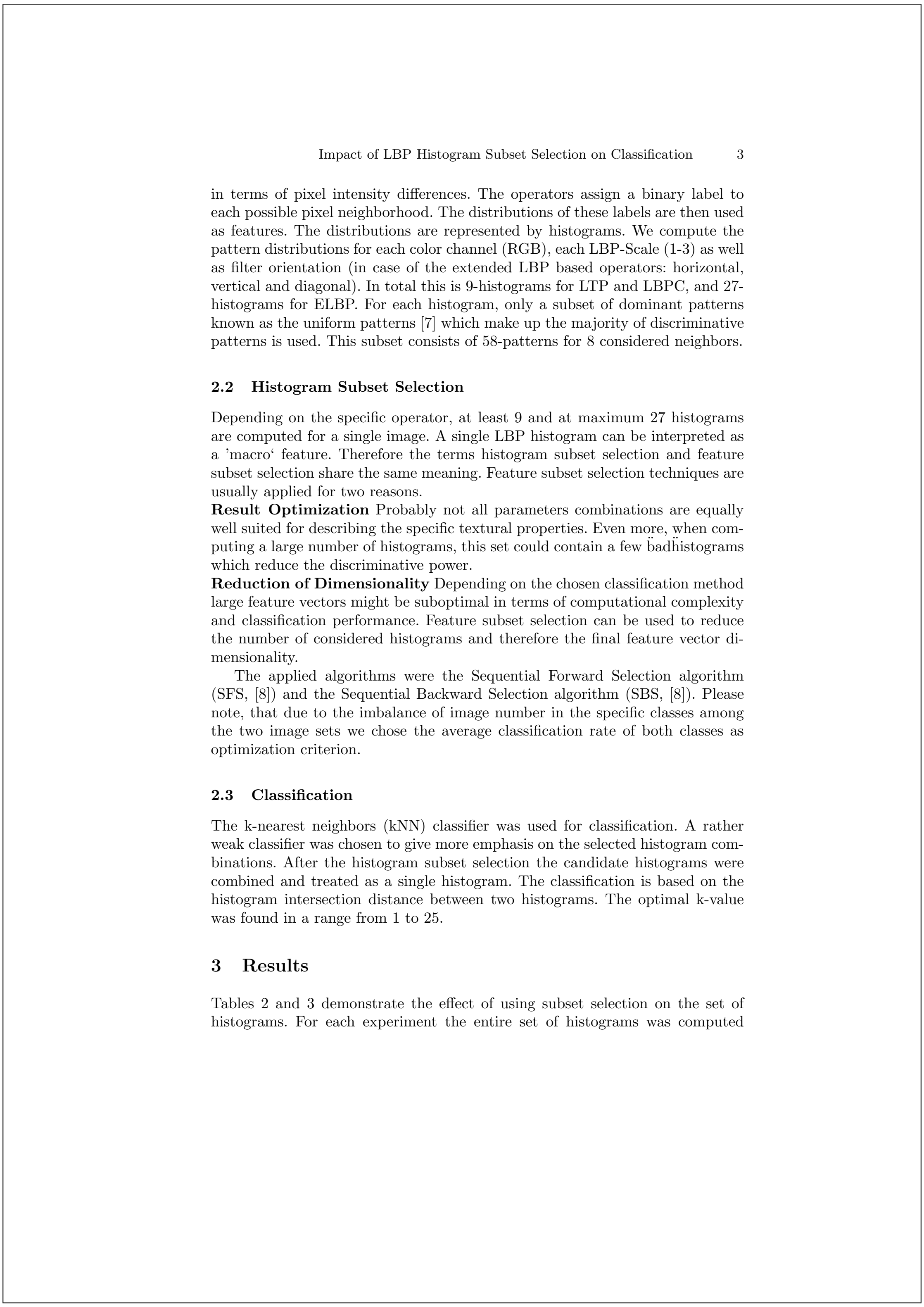
(a) (b)

Fig. 1. Images from the Pars Descendens showing the two perspectives.

The image test set used, contains images taken during duodenoscopies at the St. Anna Children's Hospital using pediatric gastroscopes without magnification. Images were recorded by using the modified immersion technique, which has been shown to be beneficial to automated classification by Hegenbart et al. [3]. There are two duodenal regions with completely different geometric properties, i.e. the duodenal Bulb and the Pars Descendens. Accordingly, we chose to separate the images into two distinct sets. Texture patches with a fixed size of 128 × 128 pixels were extracted from the full sized frames, a size which turned out to be optimally suited in previous experiments [3]. The ground truth for the texture patches used in experimentation was determined by histological examination of biopsies from corresponding regions. In the following, we aim at a two class problem with the classes '$\textbf{Class}_0$' as the class representing healthy tissue and '$\textbf{Class}_1$' representing texture patches showing villous atrophy. Table 1 shows the number of images available per considered class. For evaluation two distinct set of images for both duodenal regions denoted as Set-1 and Set-2 were assembled. This happened at two different points in time, the specific sets reflect the time intervals where the images were captured.

2.1 Feature Extraction

The basic LBP operator was introduced to the community by Ojala et al. in [4]. We use three operators that are based on LBP to conduct our experiments. The operators are LTP (Local Ternary Patterns, [5]), ELBP (extended Local Binary Patterns, [6]), and the LBP operator combined with a contrast measure (LBPC, [4]). The entire family of operators is used to model a pixel neighborhood



in terms of pixel intensity differences. The operators assign a binary label to each possible pixel neighborhood. The distributions of these labels are then used as features. The distributions are represented by histograms. We compute the pattern distributions for each color channel (RGB), each LBP-Scale (1-3) as well as filter orientation (in case of the extended LBP based operators: horizontal, vertical and diagonal). In total this is 9-histograms for LTP and LBPC, and 27-histograms for ELBP. For each histogram, only a subset of dominant patterns known as the uniform patterns [7] which make up the majority of discriminative patterns is used. This subset consists of 58-patterns for 8 considered neighbors.

2.2 Histogram Subset Selection

Depending on the specific operator, at least 9 and at maximum 27 histograms are computed for a single image. A single LBP histogram can be interpreted as a ’macro‘ feature. Therefore the terms histogram subset selection and feature subset selection share the same meaning. Feature subset selection techniques are usually applied for two reasons.
Result Optimization Probably not all parameters combinations are equally well suited for describing the specific textural properties. Even more, when computing a large number of histograms, this set could contain a few b̈adḧistograms which reduce the discriminative power.
Reduction of Dimensionality Depending on the chosen classification method large feature vectors might be suboptimal in terms of computational complexity and classification performance. Feature subset selection can be used to reduce the number of considered histograms and therefore the final feature vector dimensionality.

The applied algorithms were the Sequential Forward Selection algorithm (SFS, [8]) and the Sequential Backward Selection algorithm (SBS, [8]). Please note, that due to the imbalance of image number in the specific classes among the two image sets we chose the average classification rate of both classes as optimization criterion.

2.3 Classification

The k-nearest neighbors (kNN) classifier was used for classification. A rather weak classifier was chosen to give more emphasis on the selected histogram combinations. After the histogram subset selection the candidate histograms were combined and treated as a single histogram. The classification is based on the histogram intersection distance between two histograms. The optimal k-value was found in a range from 1 to 25.

3 Results

Tables 2 and 3 demonstrate the effect of using subset selection on the set of histograms. For each experiment the entire set of histograms was computed

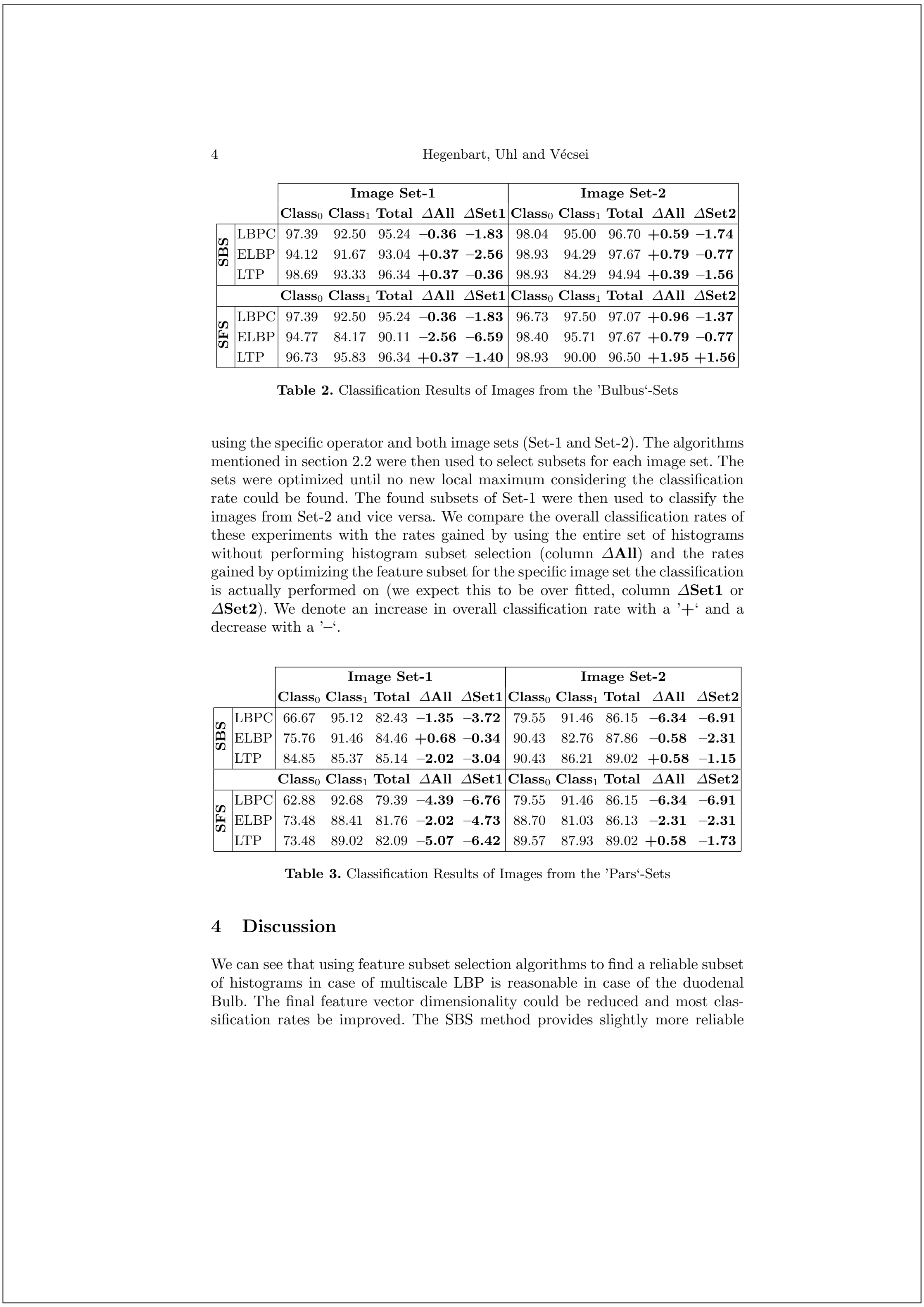


		Image Set-1					**Image Set-2**				
		Class$_0$	**Class$_1$**	**Total**	**ΔAll**	**ΔSet1**	**Class$_0$**	**Class$_1$**	**Total**	**ΔAll**	**ΔSet2**
SBS	LBPC	97.39	92.50	95.24	−**0.36**	−**1.83**	98.04	95.00	96.70	+**0.59**	−**1.74**
	ELBP	94.12	91.67	93.04	+**0.37**	−**2.56**	98.93	94.29	97.67	+**0.79**	−**0.77**
	LTP	98.69	93.33	96.34	+**0.37**	−**0.36**	98.93	84.29	94.94	+**0.39**	−**1.56**
		Class$_0$	**Class$_1$**	**Total**	**ΔAll**	**ΔSet1**	**Class$_0$**	**Class$_1$**	**Total**	**ΔAll**	**ΔSet2**
SFS	LBPC	97.39	92.50	95.24	−**0.36**	−**1.83**	96.73	97.50	97.07	+**0.96**	−**1.37**
	ELBP	94.77	84.17	90.11	−**2.56**	−**6.59**	98.40	95.71	97.67	+**0.79**	−**0.77**
	LTP	96.73	95.83	96.34	+**0.37**	−**1.40**	98.93	90.00	96.50	+**1.95**	+**1.56**

Table 2. Classification Results of Images from the 'Bulbus'-Sets

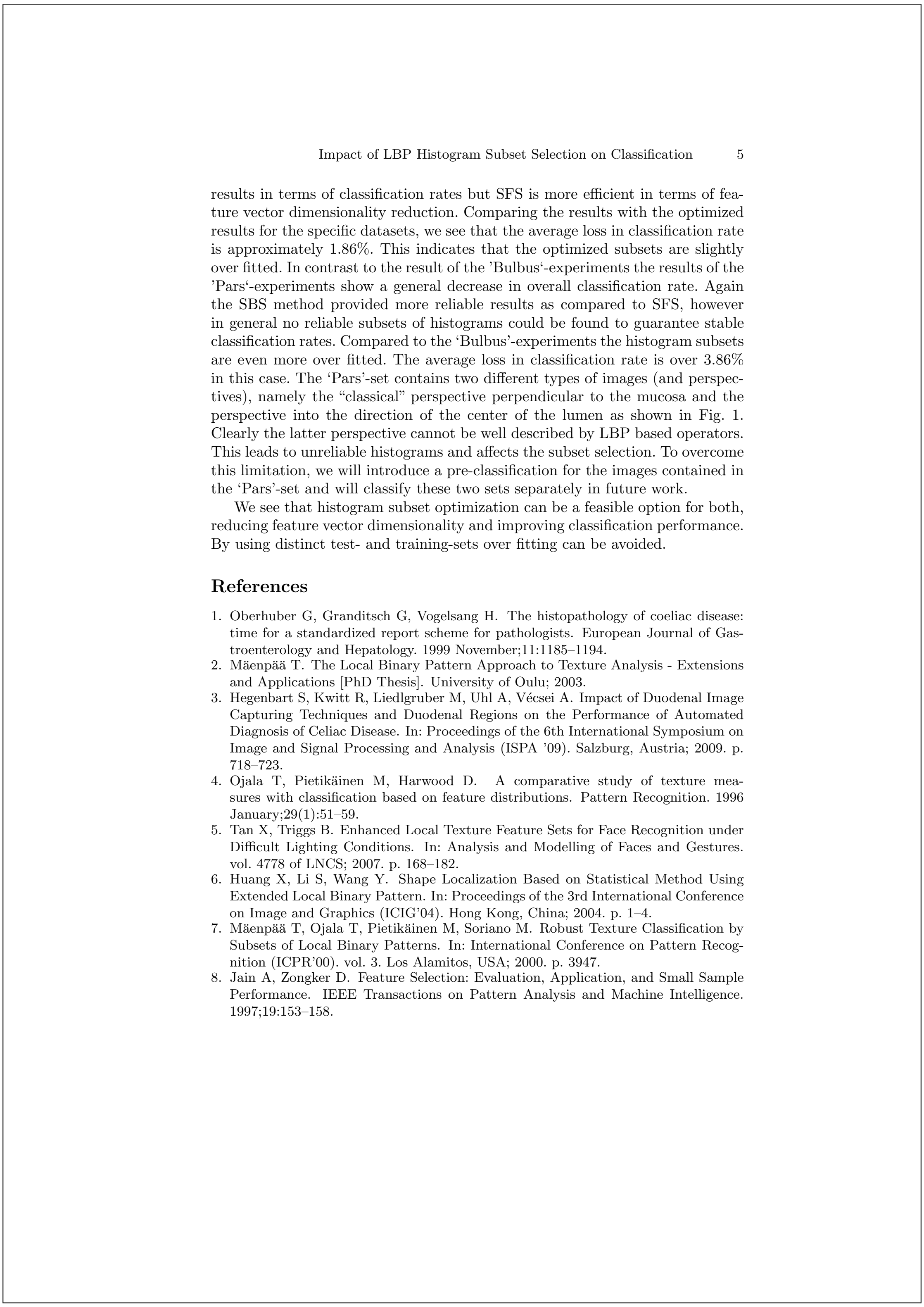
using the specific operator and both image sets (Set-1 and Set-2). The algorithms mentioned in section 2.2 were then used to select subsets for each image set. The sets were optimized until no new local maximum considering the classification rate could be found. The found subsets of Set-1 were then used to classify the images from Set-2 and vice versa. We compare the overall classification rates of these experiments with the rates gained by using the entire set of histograms without performing histogram subset selection (column **ΔAll**) and the rates gained by optimizing the feature subset for the specific image set the classification is actually performed on (we expect this to be over fitted, column **ΔSet1** or **ΔSet2**). We denote an increase in overall classification rate with a '+' and a decrease with a '–'.

		Image Set-1					**Image Set-2**				
		Class$_0$	**Class$_1$**	**Total**	**ΔAll**	**ΔSet1**	**Class$_0$**	**Class$_1$**	**Total**	**ΔAll**	**ΔSet2**
SBS	LBPC	66.67	95.12	82.43	−**1.35**	−**3.72**	79.55	91.46	86.15	−**6.34**	−**6.91**
	ELBP	75.76	91.46	84.46	+**0.68**	−**0.34**	90.43	82.76	87.86	−**0.58**	−**2.31**
	LTP	84.85	85.37	85.14	−**2.02**	−**3.04**	90.43	86.21	89.02	+**0.58**	−**1.15**
		Class$_0$	**Class$_1$**	**Total**	**ΔAll**	**ΔSet1**	**Class$_0$**	**Class$_1$**	**Total**	**ΔAll**	**ΔSet2**
SFS	LBPC	62.88	92.68	79.39	−**4.39**	−**6.76**	79.55	91.46	86.15	−**6.34**	−**6.91**
	ELBP	73.48	88.41	81.76	−**2.02**	−**4.73**	88.70	81.03	86.13	−**2.31**	−**2.31**
	LTP	73.48	89.02	82.09	−**5.07**	−**6.42**	89.57	87.93	89.02	+**0.58**	−**1.73**

Table 3. Classification Results of Images from the 'Pars'-Sets

4 Discussion

We can see that using feature subset selection algorithms to find a reliable subset of histograms in case of multiscale LBP is reasonable in case of the duodenal Bulb. The final feature vector dimensionality could be reduced and most classification rates be improved. The SBS method provides slightly more reliable



results in terms of classification rates but SFS is more efficient in terms of feature vector dimensionality reduction. Comparing the results with the optimized results for the specific datasets, we see that the average loss in classification rate is approximately 1.86%. This indicates that the optimized subsets are slightly over fitted. In contrast to the result of the 'Bulbus'-experiments the results of the 'Pars'-experiments show a general decrease in overall classification rate. Again the SBS method provided more reliable results as compared to SFS, however in general no reliable subsets of histograms could be found to guarantee stable classification rates. Compared to the 'Bulbus'-experiments the histogram subsets are even more over fitted. The average loss in classification rate is over 3.86% in this case. The 'Pars'-set contains two different types of images (and perspectives), namely the "classical" perspective perpendicular to the mucosa and the perspective into the direction of the center of the lumen as shown in Fig. 1. Clearly the latter perspective cannot be well described by LBP based operators. This leads to unreliable histograms and affects the subset selection. To overcome this limitation, we will introduce a pre-classification for the images contained in the 'Pars'-set and will classify these two sets separately in future work.

We see that histogram subset optimization can be a feasible option for both, reducing feature vector dimensionality and improving classification performance. By using distinct test- and training-sets over fitting can be avoided.

References

1. Oberhuber G, Granditsch G, Vogelsang H. The histopathology of coeliac disease: time for a standardized report scheme for pathologists. European Journal of Gastroenterology and Hepatology. 1999 November;11:1185–1194.
2. Mäenpää T. The Local Binary Pattern Approach to Texture Analysis - Extensions and Applications [PhD Thesis]. University of Oulu; 2003.
3. Hegenbart S, Kwitt R, Liedlgruber M, Uhl A, Vécsei A. 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; 2009. p. 718–723.
4. Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on feature distributions. Pattern Recognition. 1996 January;29(1):51–59.
5. Tan X, Triggs B. Enhanced Local Texture Feature Sets for Face Recognition under Difficult Lighting Conditions. In: Analysis and Modelling of Faces and Gestures. vol. 4778 of LNCS; 2007. p. 168–182.
6. Huang X, Li S, Wang Y. 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; 2004. p. 1–4.
7. Mäenpää T, Ojala T, Pietikäinen M, Soriano M. Robust Texture Classification by Subsets of Local Binary Patterns. In: International Conference on Pattern Recognition (ICPR'00). vol. 3. Los Alamitos, USA; 2000. p. 3947.
8. Jain A, Zongker D. Feature Selection: Evaluation, Application, and Small Sample Performance. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1997;19:153–158.

Systematic Assessment of Performance Prediction Techniques in Medical Image Classification A Case Study on Celiac Disease

Sebastian Hegenbart[1], Andreas Uhl[1], and Andreas Vécsei[2]

[1]Department of Computer Sciences, University of Salzburg
[2]St. Anna Children's Hospital, Vienna

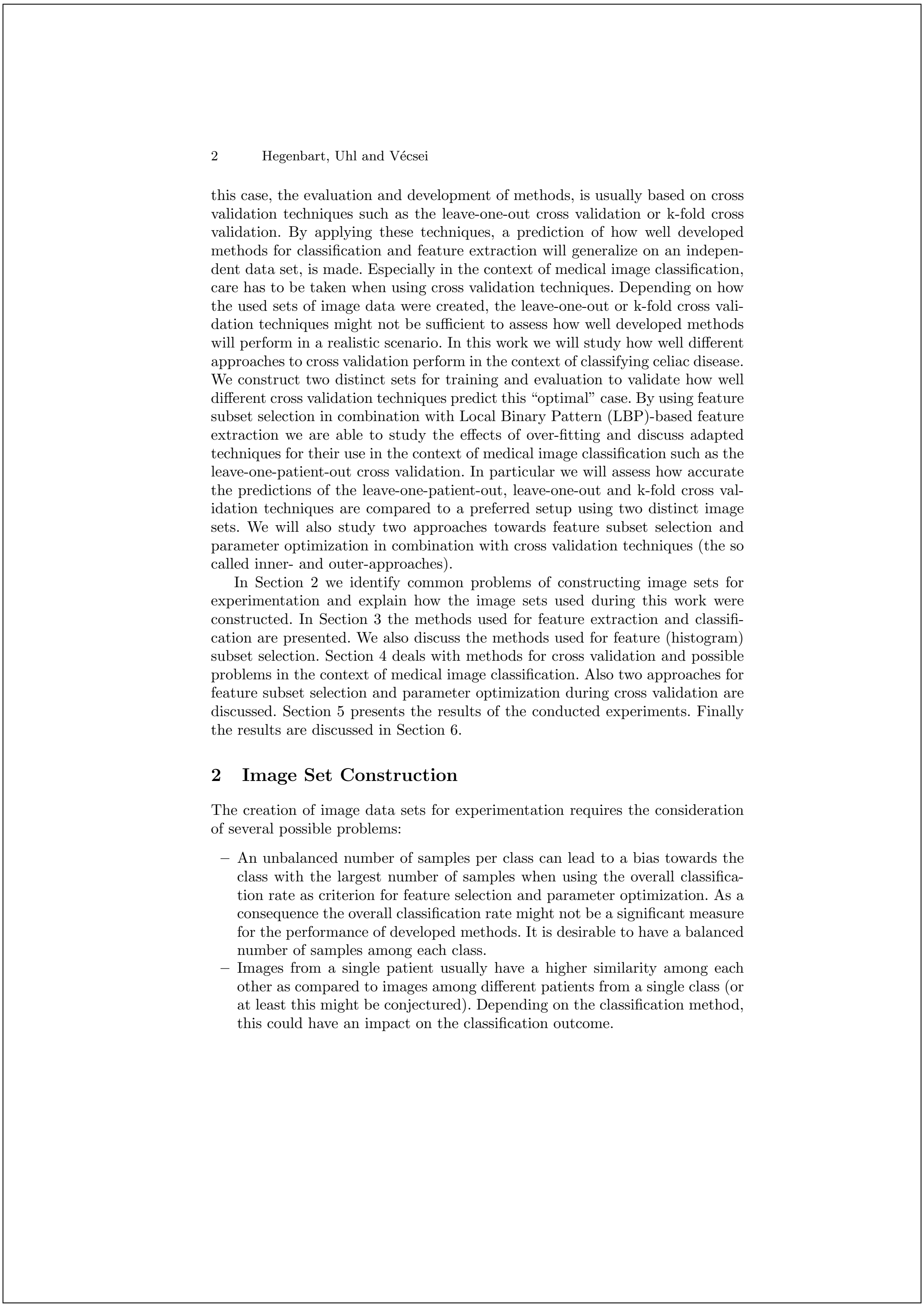
Abstract. In the context of automated classification of medical images, many authors report a lack of available test data. Therefore techniques such as the leave-one-out cross validation or k-fold validation are used to assess how well methods will perform in practice. In case of methods based on feature subset selection, cross validation might provide bad estimations of how well the optimized technique generalizes on an independent data set. In this work, we assess how well cross validation techniques are suited to predict the outcome of a preferred setup of distinct test- and training data sets. This is accomplished by creating two distinct sets of images, used separately as training- and test-data. The experiments are conducted using a set of Local Binary Pattern based operators for feature extraction which are using histogram subset selection to improve the feature discrimination. Common problems such as the effects of over fitting data during cross validation as well as using biased image sets due to multiple images from a single patient are considered.

Key words: celiac disease, classification, cross validation, over fitting, LOPO

1 Introduction

A desirable data setup for experimentation within the field of medical image classification consists of two distinct sets of image samples with a balanced number of images and patients among the specific classes. In this case one set is used for training a classifier as well as performing feature selection and parameter optimization. A method's classification accuracy is then evaluated by using the trained classification method with it's specific parameters on the other set of data samples. In the context of automated classification of medical images however, the available amount of test data is often very limited. Often it is not possible to build distinct data sets for training and evaluation. This can be due to a limited number of patients (e.g. a low prevalence of the specific disease), a limited number of usable images caused by qualitative problems or a high number of classes used to categorize the pathological changes in relation to the available images. In

Systematic Assessment of Performance Prediction Techniques in Medical Image Classification - A Case Study on Celiac Disease.



this case, the evaluation and development of methods, is usually based on cross validation techniques such as the leave-one-out cross validation or k-fold cross validation. By applying these techniques, a prediction of how well developed methods for classification and feature extraction will generalize on an independent data set, is made. Especially in the context of medical image classification, care has to be taken when using cross validation techniques. Depending on how the used sets of image data were created, the leave-one-out or k-fold cross validation techniques might not be sufficient to assess how well developed methods will perform in a realistic scenario. In this work we will study how well different approaches to cross validation perform in the context of classifying celiac disease. We construct two distinct sets for training and evaluation to validate how well different cross validation techniques predict this “optimal” case. By using feature subset selection in combination with Local Binary Pattern (LBP)-based feature extraction we are able to study the effects of over-fitting and discuss adapted techniques for their use in the context of medical image classification such as the leave-one-patient-out cross validation. In particular we will assess how accurate the predictions of the leave-one-patient-out, leave-one-out and k-fold cross validation techniques are compared to a preferred setup using two distinct image sets. We will also study two approaches towards feature subset selection and parameter optimization in combination with cross validation techniques (the so called inner- and outer-approaches).

In Section 2 we identify common problems of constructing image sets for experimentation and explain how the image sets used during this work were constructed. In Section 3 the methods used for feature extraction and classification are presented. We also discuss the methods used for feature (histogram) subset selection. Section 4 deals with methods for cross validation and possible problems in the context of medical image classification. Also two approaches for feature subset selection and parameter optimization during cross validation are discussed. Section 5 presents the results of the conducted experiments. Finally the results are discussed in Section 6.

2 Image Set Construction

The creation of image data sets for experimentation requires the consideration of several possible problems:

- An unbalanced number of samples per class can lead to a bias towards the class with the largest number of samples when using the overall classification rate as criterion for feature selection and parameter optimization. As a consequence the overall classification rate might not be a significant measure for the performance of developed methods. It is desirable to have a balanced number of samples among each class.
- Images from a single patient usually have a higher similarity among each other as compared to images among different patients from a single class (or at least this might be conjectured). Depending on the classification method, this could have an impact on the classification outcome.

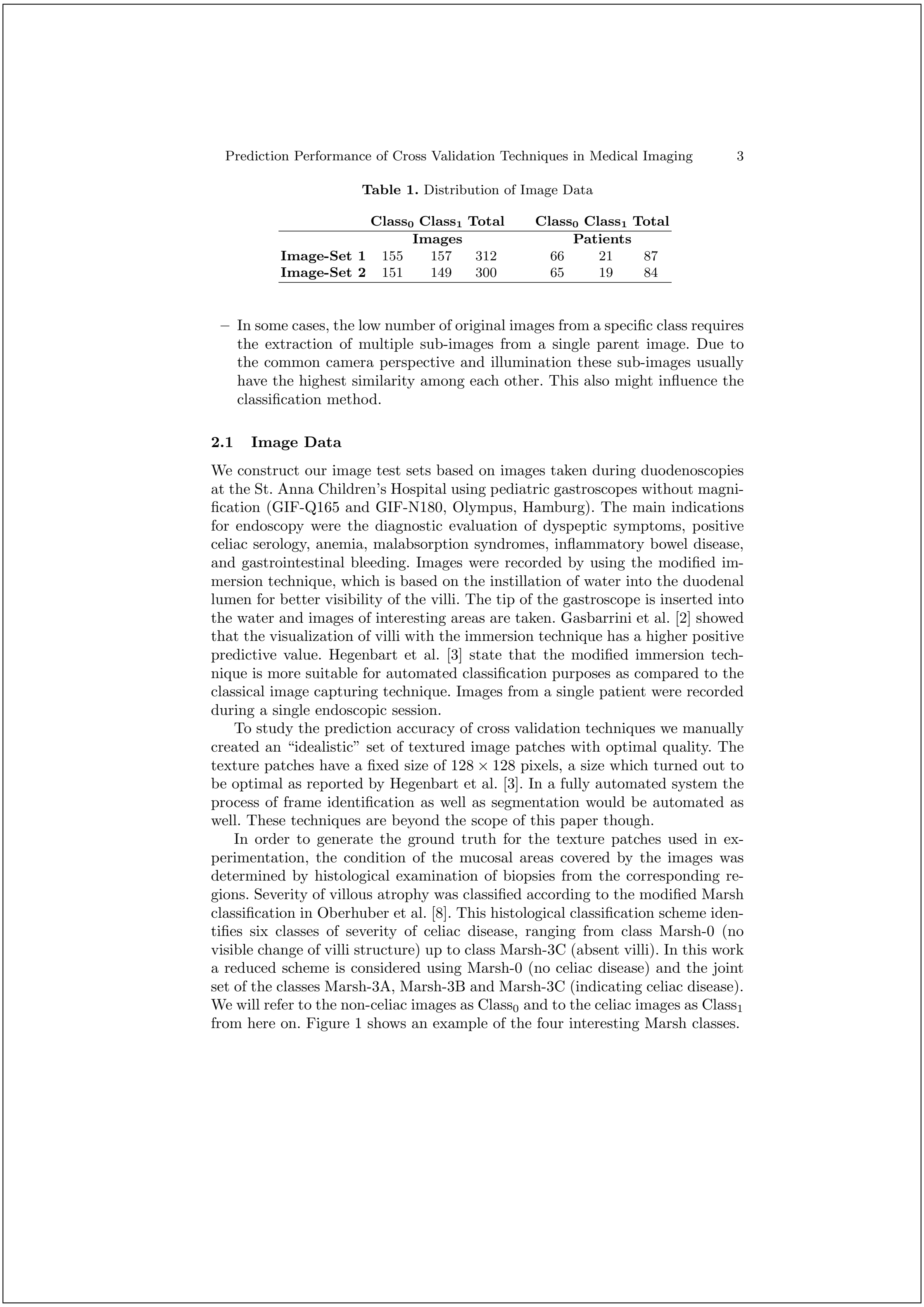


Table 1. Distribution of Image Data

	Class$_0$	**Class$_1$**	**Total**	**Class$_0$**	**Class$_1$**	**Total**
	Images			**Patients**		
Image-Set 1	155	157	312	66	21	87
Image-Set 2	151	149	300	65	19	84

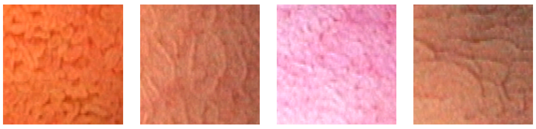
– In some cases, the low number of original images from a specific class requires the extraction of multiple sub-images from a single parent image. Due to the common camera perspective and illumination these sub-images usually have the highest similarity among each other. This also might influence the classification method.

2.1 Image Data

We construct our image test sets based on 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. [2] showed that the visualization of villi with the immersion technique has a higher positive predictive value. Hegenbart et al. [3] state 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.

To study the prediction accuracy of cross validation techniques we manually created an "idealistic" set of textured image patches with optimal quality. The texture patches have a fixed size of 128 × 128 pixels, a size which turned out to be optimal as reported by Hegenbart et al. [3]. In a fully automated system the process of frame identification as well as segmentation would be automated as well. These techniques are beyond the scope of this paper though.

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. [8]. 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). In this work a reduced scheme is considered using Marsh-0 (no celiac disease) and the joint set of the classes Marsh-3A, Marsh-3B and Marsh-3C (indicating celiac disease). We will refer to the non-celiac images as Class$_0$ and to the celiac images as Class$_1$ from here on. Figure 1 shows an example of the four interesting Marsh classes.



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

Fig. 1. Examples of Duodenal Image-Patches used for Experimentation.

2.2 Construction of Distinct Data Sets

The constructed image sets originate from 171 patients (131 control patients and 40 patients with diagnosed celiac disease). In order to guarantee an image set of reasonable size, more than a single texture patch was extracted for each patient from the original images. 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 generator based on a Gaussian distribution to avoid any bias within the data sets. Table 1 shows the distribution of images and patients per class.

3 Feature Extraction and Classification

The basic LBP operator was introduced to the community by Ojala et al. [9]. We use three operators that are based on LBP to conduct our experiments. The operators are LBP (Local Binary Patterns, [11]), ELBP (extended Local Binary Patterns, [4]), and a modified version of the ELBP operator that is introduced in this work, the ELTP (extended Local Ternary Patterns) operator. The entire family of operators is used to model a pixel neighborhood in terms of pixel intensity differences. The operators assign a binary label to each possible pixel neighborhood. The distributions of these labels are then used as features, which are represented by histograms. We compute the pattern distributions for each color channel (RGB), each LBP-Scale (1-3) (see Section 3.1) as well as filter orientation (in case of the extended LBP based operators: horizontal, vertical and diagonal). In total we result in 9-histograms for LBP and 27-histograms for ELBP and ELTP. For each histogram, only a subset of dominant patterns known as the uniform patterns [7], which make up the majority of discriminative patterns, is used. In case of the LBP and ELBP operator this subset consists of 58-patterns for 8 considered neighbors. In case of the ELTP operator two histograms with 58-bins are concatenated, therefore the dimensionality of the ELTP histograms is 116 bins.

3.1 Local Binary Patterns

For the radius r and the number of considered neighbors p, the LBP operator is 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 s function acts as sign function, mapping to 1 if the difference is smaller or equal to 0 and mapping to 0 else. 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. in [10] 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 in [6].

3.2 Extended Local Binary Patterns and Extended Local Ternary Patterns with adaptive Threshold

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. [4] suggest using a gradient filtering before feature extraction and call this operator ELBP or extended LBP. By doing this the velocity of local variation is described by the pixel neighborhoods.

We introduce the extended LTP (ELTP) operator consequently 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. The Local Ternary Pattern operator (LTP) was introduced by Tan and Triggs [11]. 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 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 is based on a thresholded sign function:

$$s(x) = \begin{cases} 1, & \text{if } x \geq T_h \\ 0, & \text{if } |x| < T_h \\ -1, & \text{if } x \leq -T_h. \end{cases} \tag{2}$$

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.

$$H_{I,lower}(i) = \sum_{x,y} (LBP_{r,p}(x,y) = -i) \qquad i = 0, \cdots, 2^p - 1 \tag{3}$$

$$H_{I,upper}(i) = \sum_{x,y}(LBP_{r,p}(x,y) = i) \qquad i = 0, \cdots, 2^p - 1 \tag{4}$$

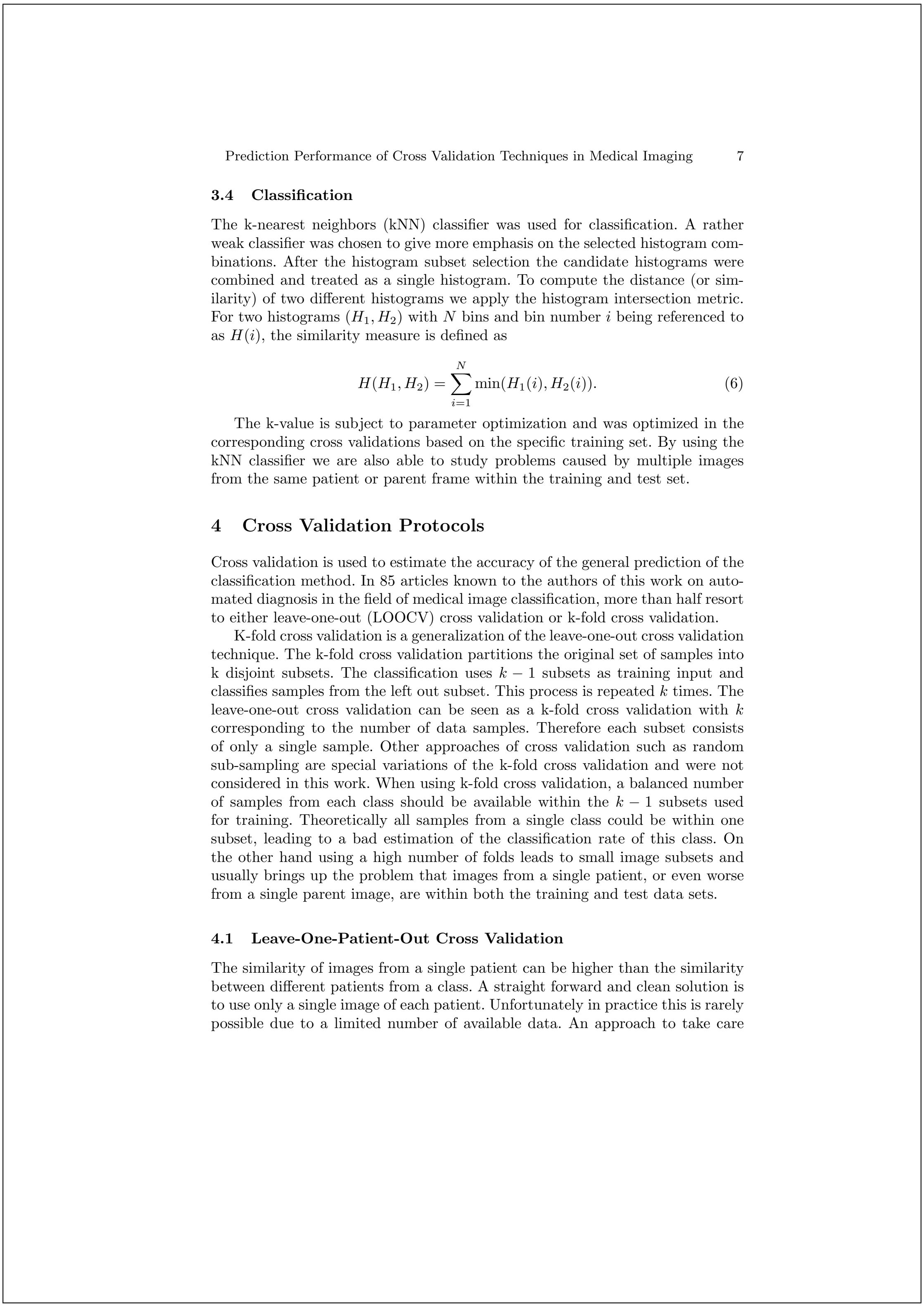
The two computed histograms are concatenated and then treated like a single histogram. Please note that in analogy to the LBP operator, only the uniform subset of patterns was used in this case. The actual optimal values to use for thresholding are unknown a priori. We apply an adaptive threshold based on the spatial image statistics to 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. The calculation is based on an expected value for the standard deviation of the image (β). This value was found based on the training data used during experimentation and represents the average standard deviation of pixel intensity values within all images. 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.

$$T_h = \begin{cases} \beta^{\frac{1}{2}} + \alpha\sigma, & \text{if } \sigma > \beta \\ \beta^{\frac{1}{2}} - \alpha\sigma, & \text{if } \sigma \leq \beta. \end{cases} \tag{5}$$

3.3 Histogram Subset Selection

Depending on the specific operator, at least 9 (LBP) and at maximum 27 (ELBP and ELTP) histograms are computed for a single image. A single LBP histogram can be interpreted as a “macro” feature. Therefore the terms histogram subset selection and feature subset selection share the same meaning. Feature subset selection techniques are usually applied for two reasons.
Result Optimization Probably not all parameters combinations are equally well suited for describing the specific textural properties. Even more, when computing a large number of histograms, this set could contain a few “bad” histograms which reduce the discriminative power.
Reduction of Dimensionality Depending on the chosen classification method large feature vectors might be suboptimal in terms of computational complexity and classification performance. Feature subset selection can be used to reduce the number of considered histograms and therefore the final feature vector dimensionality.

The applied algorithm for histogram subset selection was the Sequential Forward Selection algorithm (SFS, [5]). The optimization criterion for this algorithm was the overall classification rate. The upper bound set on the number of selected histograms was 10. This technique of optimizing the feature subset might be subject to over fitting. We expect the operators computing a larger number of histograms (ELBP and ELTP) to be at higher risk of being over fitted when using “outer” optimization (see section 4.2 for a comparison of approaches for optimization).



3.4 Classification

The k-nearest neighbors (kNN) classifier was used for classification. A rather weak classifier was chosen to give more emphasis on the selected histogram combinations. After the histogram subset selection the candidate histograms were combined and treated as a single histogram. To compute the distance (or similarity) of two different histograms we apply the histogram intersection metric. For two histograms (H_1, H_2) with N bins and bin number i being referenced to as $H(i)$, the similarity measure is defined as

$$H(H_1, H_2) = \sum_{i=1}^{N} \min(H_1(i), H_2(i)). \tag{6}$$

The k-value is subject to parameter optimization and was optimized in the corresponding cross validations based on the specific training set. By using the kNN classifier we are also able to study problems caused by multiple images from the same patient or parent frame within the training and test set.

4 Cross Validation Protocols

Cross validation is used to estimate the accuracy of the general prediction of the classification method. In 85 articles known to the authors of this work on automated diagnosis in the field of medical image classification, more than half resort to either leave-one-out (LOOCV) cross validation or k-fold cross validation.

K-fold cross validation is a generalization of the leave-one-out cross validation technique. The k-fold cross validation partitions the original set of samples into k disjoint subsets. The classification uses $k - 1$ subsets as training input and classifies samples from the left out subset. This process is repeated k times. The leave-one-out cross validation can be seen as a k-fold cross validation with k corresponding to the number of data samples. Therefore each subset consists of only a single sample. Other approaches of cross validation such as random sub-sampling are special variations of the k-fold cross validation and were not considered in this work. When using k-fold cross validation, a balanced number of samples from each class should be available within the $k - 1$ subsets used for training. Theoretically all samples from a single class could be within one subset, leading to a bad estimation of the classification rate of this class. On the other hand using a high number of folds leads to small image subsets and usually brings up the problem that images from a single patient, or even worse from a single parent image, are within both the training and test data sets.

4.1 Leave-One-Patient-Out Cross Validation

The similarity of images from a single patient can be higher than the similarity between different patients from a class. A straight forward and clean solution is to use only a single image of each patient. Unfortunately in practice this is rarely possible due to a limited number of available data. An approach to take care of this problem is the leave-one-patient-out (LOPO) cross validation technique (also used by André et al. [1]). LOPO cross validation is based on the k-fold cross validation. The partitioning of the original set of samples however is done such that each partition contains only images from a single patient. This approach implies that patient information in some, usually unambiguously anonymized, form is available. A variation that is closely related to the LOPO cross validation method is the leave-one-parent-frame out cross validation. In this technique the partitioning is performed such that each partition consists of all sub-images from a parent image. This approach can usually be used if no patient information is available. However, the LOPO cross validation technique should be preferred over the leave-one-parent-image-out technique whenever possible.

4.2 Feature Optimization Combined with Cross Validation

We distinguish between two approaches to feature subset selection and parameter optimization in combination with cross validation.

- The **outer**-approach optimizes features or parameters based on the results of the cross validation used for predicting the classifier's accuracy. This means that the optimization criterion of the feature subset selection method is based on the estimates of a cross validation on the entire data set. These estimations are also used as classification rates later.
- The **inner**-approach optimizes features or parameters within a separate cross validation based on the $k-1$ partitions used for training within the cross validation used for predicting classification accuracy. This means that the optimization criterion of the feature subset selection method is based on a separate cross validation using the training set ($k-1$ partitions) of the current "outer" cross validation. Therefore, for each partition an new feature subset is selected. The classification rate is the estimation of the "outer" cross validation.

The outer-approach is the classical and easier approach frequently found within the literature. This approach however poses the problem that test data is used for optimizing feature subsets or parameters. This can have an influence on the optimization and therefore an effect on the prediction of how well the feature subset or optimized parameters generalize (the optimization over-fits the model towards the data). By using the inner-approach, the risk of over-fitting is reduced, the major drawback is that the computational power needed for this evaluation is considerably higher as compared to the other technique. This is caused by repeated feature subset selection and parameter optimization which is usually the most time consuming element in the automated classification chain.

5 Results

This Section presents the results of the experiments. Please note, that we use subscripts combined with the method names to indicate the type of optimization.

The inner-approach is indicated by a “I”, while the outer-approach is indicated by a “O”. All results except the distinct set classification was computed using the specific cross validation technique on Image-Set1. The abbreviations “Spec.“ and “Sens.“ refer to the Methods’ specificity and sensitivity.

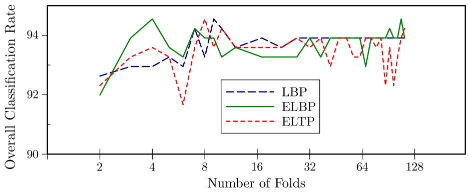


Fig. 2. Overall Classification Rate Estimates of **k-Fold$_{\mathbf{O}}$ (outer)** Cross Validations.

Figures 2 and 3 show the overall classification rates predicted by using k-fold cross validation. Due to computational issues, the values were computed from 2 to 10 in single steps and from 12 to 112 in steps of 5. The mean classification rates are: LBP (93.75%, $\sigma = 0.43$), ELBP (93.48%, $\sigma = 0.5$) as well as ELTP (93.48%, $\sigma = 0.64$) in case of the outer-approach

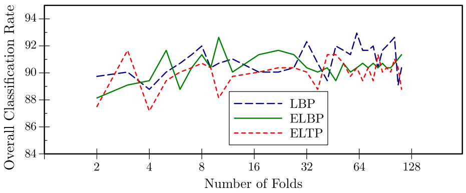


Fig. 3. Overall Classification Rate Estimates of **k-Fold$_{\mathbf{I}}$ (inner)** Cross Validations.

and LBP (90.98%, $\sigma = 1.08$), ELBP (90.48%, $\sigma = 0.91$) as well as ELTP (89.99%, $\sigma = 1.07$) in case of the inner-approach. The columns of Table 2 labeled as Δ list the differences of the predictions of the overall classification rates between the outer- and inner-approach. The experiments based on the inner-approach used a leave-one-out cross validation as the “inner” cross validation method in all cases.

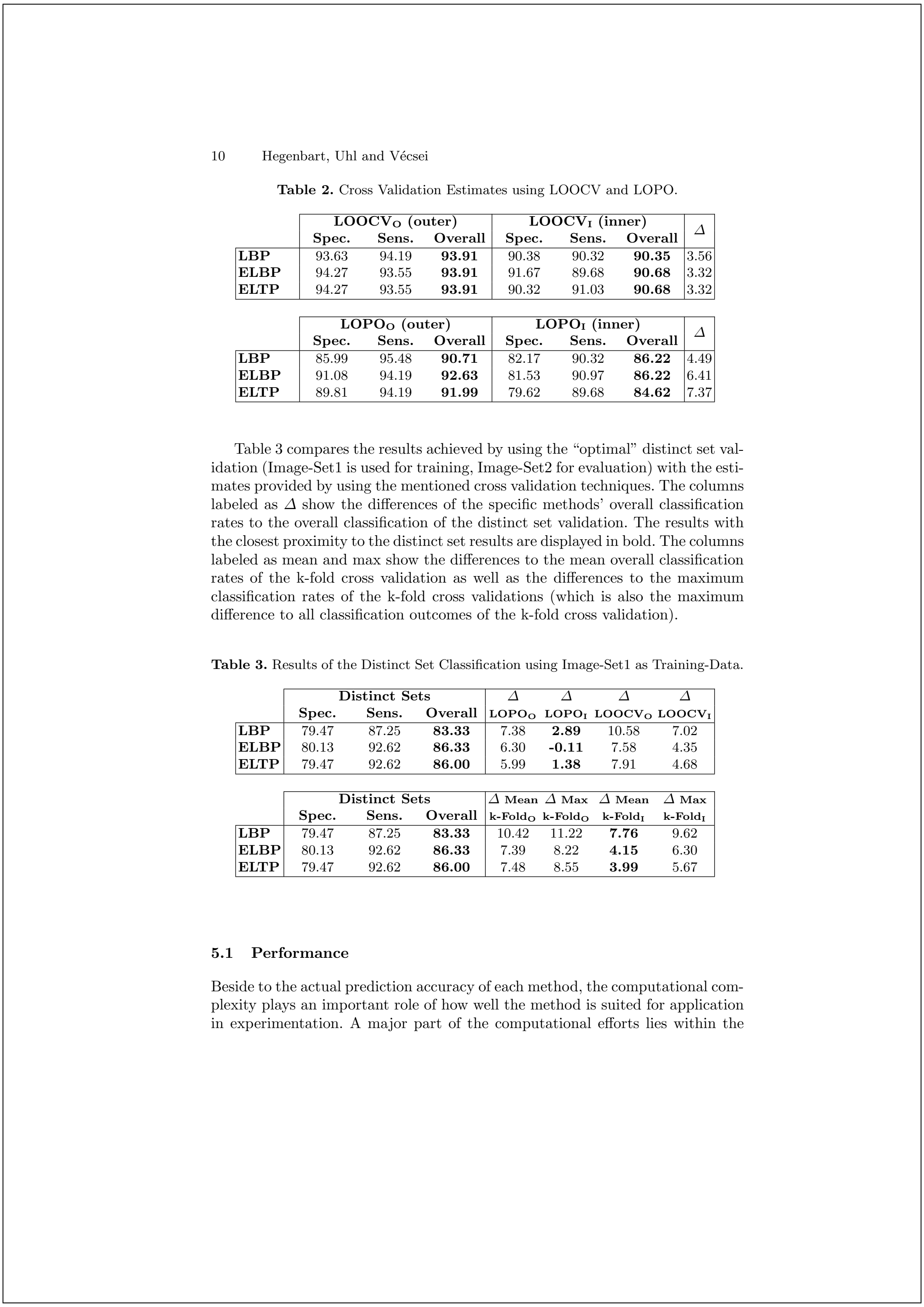


Table 2. Cross Validation Estimates using LOOCV and LOPO.

	LOOCV$_\text{O}$ (outer)			**LOOCV$_\text{I}$ (inner)**			Δ
	Spec.	**Sens.**	**Overall**	**Spec.**	**Sens.**	**Overall**	
LBP	93.63	94.19	**93.91**	90.38	90.32	**90.35**	3.56
ELBP	94.27	93.55	**93.91**	91.67	89.68	**90.68**	3.32
ELTP	94.27	93.55	**93.91**	90.32	91.03	**90.68**	3.32

	LOPO$_\text{O}$ (outer)			**LOPO$_\text{I}$ (inner)**			Δ
	Spec.	**Sens.**	**Overall**	**Spec.**	**Sens.**	**Overall**	
LBP	85.99	95.48	**90.71**	82.17	90.32	**86.22**	4.49
ELBP	91.08	94.19	**92.63**	81.53	90.97	**86.22**	6.41
ELTP	89.81	94.19	**91.99**	79.62	89.68	**84.62**	7.37

Table 3 compares the results achieved by using the "optimal" distinct set validation (Image-Set1 is used for training, Image-Set2 for evaluation) with the estimates provided by using the mentioned cross validation techniques. The columns labeled as Δ show the differences of the specific methods' overall classification rates to the overall classification of the distinct set validation. The results with the closest proximity to the distinct set results are displayed in bold. The columns labeled as mean and max show the differences to the mean overall classification rates of the k-fold cross validation as well as the differences to the maximum classification rates of the k-fold cross validations (which is also the maximum difference to all classification outcomes of the k-fold cross validation).

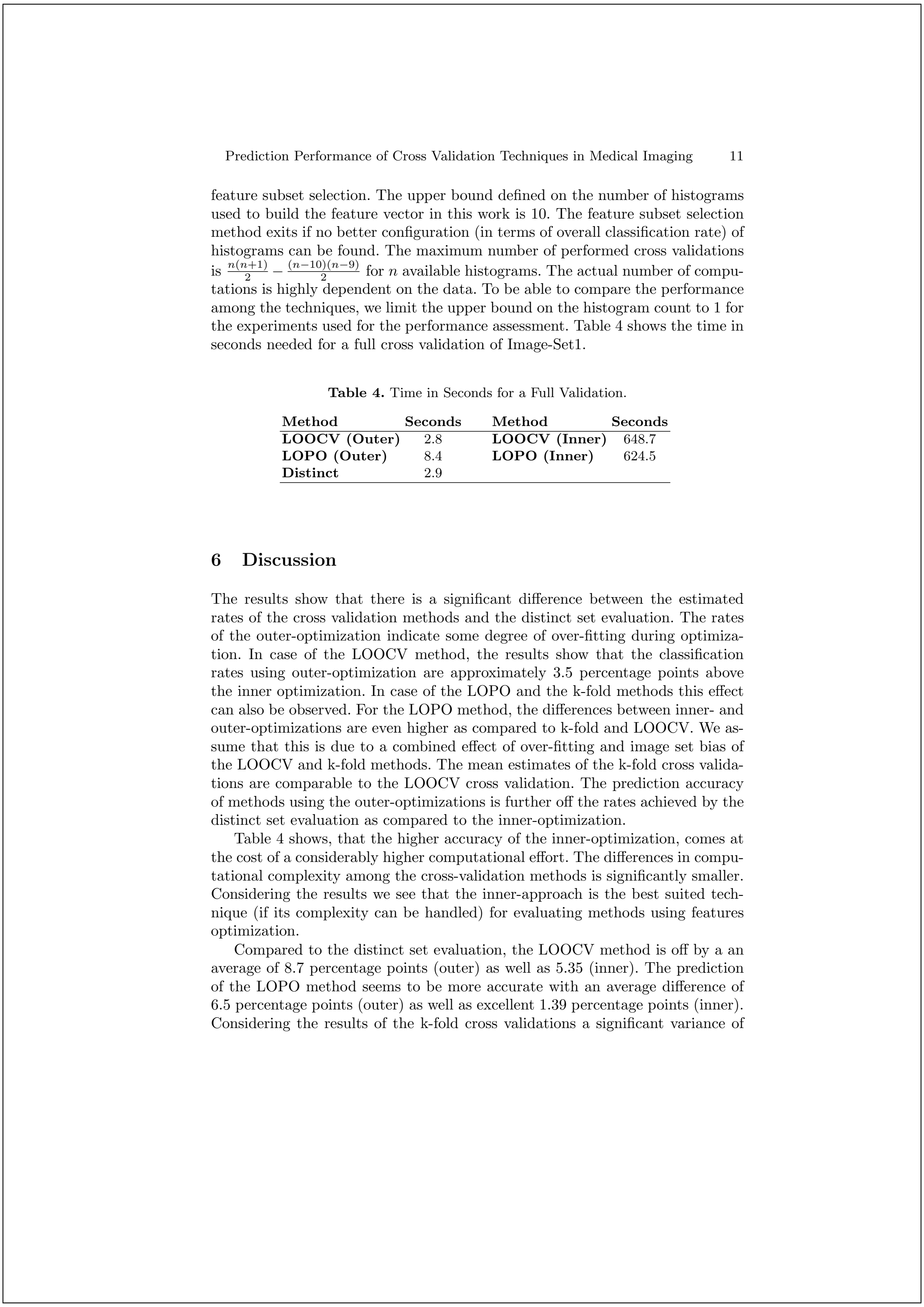
Table 3. Results of the Distinct Set Classification using Image-Set1 as Training-Data.

	Distinct Sets			Δ	Δ	Δ	Δ
	Spec.	**Sens.**	**Overall**	**LOPO$_\text{O}$**	**LOPO$_\text{I}$**	**LOOCV$_\text{O}$**	**LOOCV$_\text{I}$**
LBP	79.47	87.25	**83.33**	7.38	**2.89**	10.58	7.02
ELBP	80.13	92.62	**86.33**	6.30	**-0.11**	7.58	4.35
ELTP	79.47	92.62	**86.00**	5.99	**1.38**	7.91	4.68

	Distinct Sets			Δ **Mean**	Δ **Max**	Δ **Mean**	Δ **Max**
	Spec.	**Sens.**	**Overall**	**k-Fold$_\text{O}$**	**k-Fold$_\text{O}$**	**k-Fold$_\text{I}$**	**k-Fold$_\text{I}$**
LBP	79.47	87.25	**83.33**	10.42	11.22	**7.76**	9.62
ELBP	80.13	92.62	**86.33**	7.39	8.22	**4.15**	6.30
ELTP	79.47	92.62	**86.00**	7.48	8.55	**3.99**	5.67

5.1 Performance

Beside to the actual prediction accuracy of each method, the computational complexity plays an important role of how well the method is suited for application in experimentation. A major part of the computational efforts lies within the



feature subset selection. The upper bound defined on the number of histograms used to build the feature vector in this work is 10. The feature subset selection method exits if no better configuration (in terms of overall classification rate) of histograms can be found. The maximum number of performed cross validations is $\frac{n(n+1)}{2} - \frac{(n-10)(n-9)}{2}$ for n available histograms. The actual number of computations is highly dependent on the data. To be able to compare the performance among the techniques, we limit the upper bound on the histogram count to 1 for the experiments used for the performance assessment. Table 4 shows the time in seconds needed for a full cross validation of Image-Set1.

Table 4. Time in Seconds for a Full Validation.

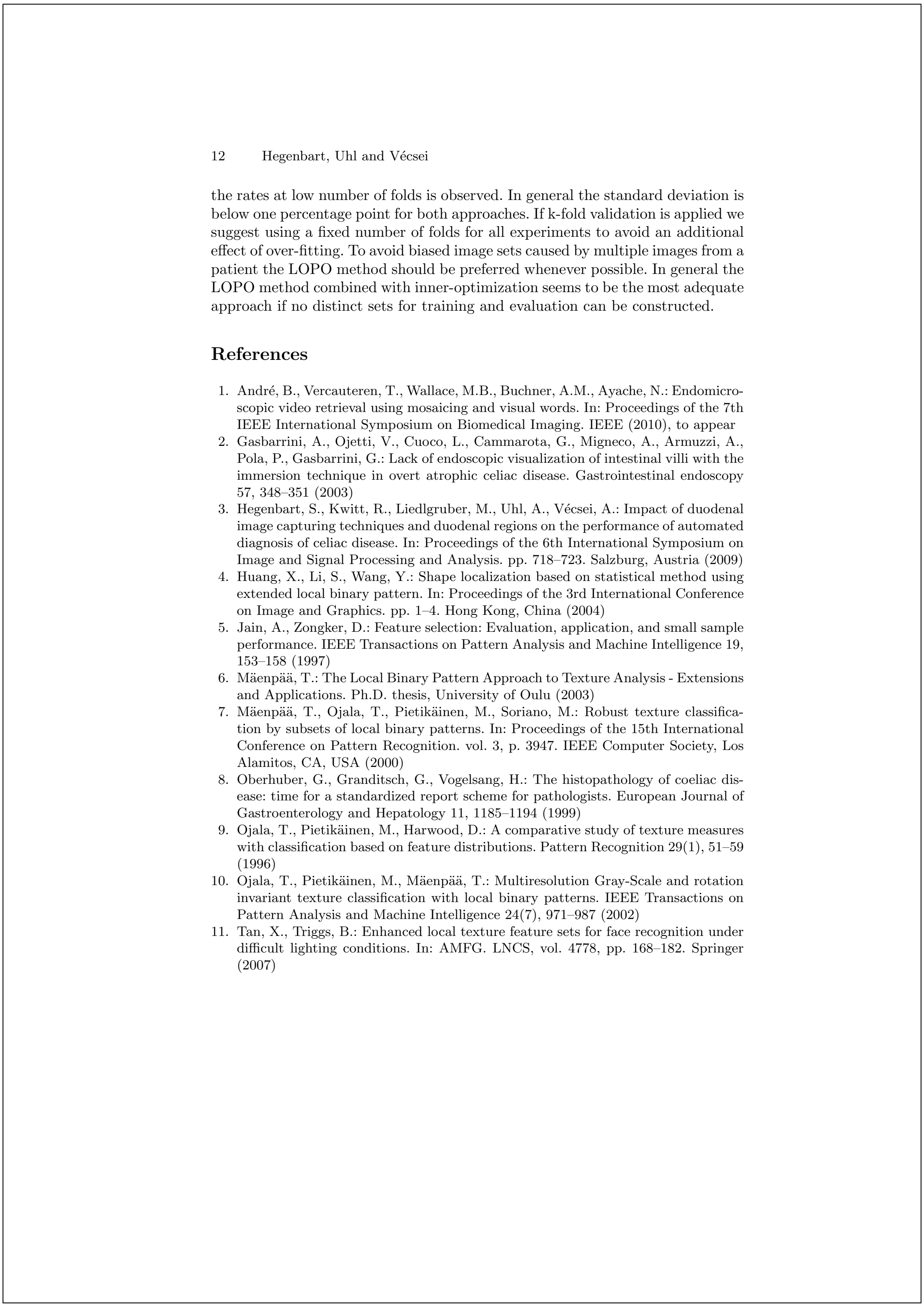
Method	Seconds	Method	Seconds
LOOCV (Outer)	2.8	**LOOCV (Inner)**	648.7
LOPO (Outer)	8.4	**LOPO (Inner)**	624.5
Distinct	2.9		

6 Discussion

The results show that there is a significant difference between the estimated rates of the cross validation methods and the distinct set evaluation. The rates of the outer-optimization indicate some degree of over-fitting during optimization. In case of the LOOCV method, the results show that the classification rates using outer-optimization are approximately 3.5 percentage points above the inner optimization. In case of the LOPO and the k-fold methods this effect can also be observed. For the LOPO method, the differences between inner- and outer-optimizations are even higher as compared to k-fold and LOOCV. We assume that this is due to a combined effect of over-fitting and image set bias of the LOOCV and k-fold methods. The mean estimates of the k-fold cross validations are comparable to the LOOCV cross validation. The prediction accuracy of methods using the outer-optimizations is further off the rates achieved by the distinct set evaluation as compared to the inner-optimization.

Table 4 shows, that the higher accuracy of the inner-optimization, comes at the cost of a considerably higher computational effort. The differences in computational complexity among the cross-validation methods is significantly smaller. Considering the results we see that the inner-approach is the best suited technique (if its complexity can be handled) for evaluating methods using features optimization.

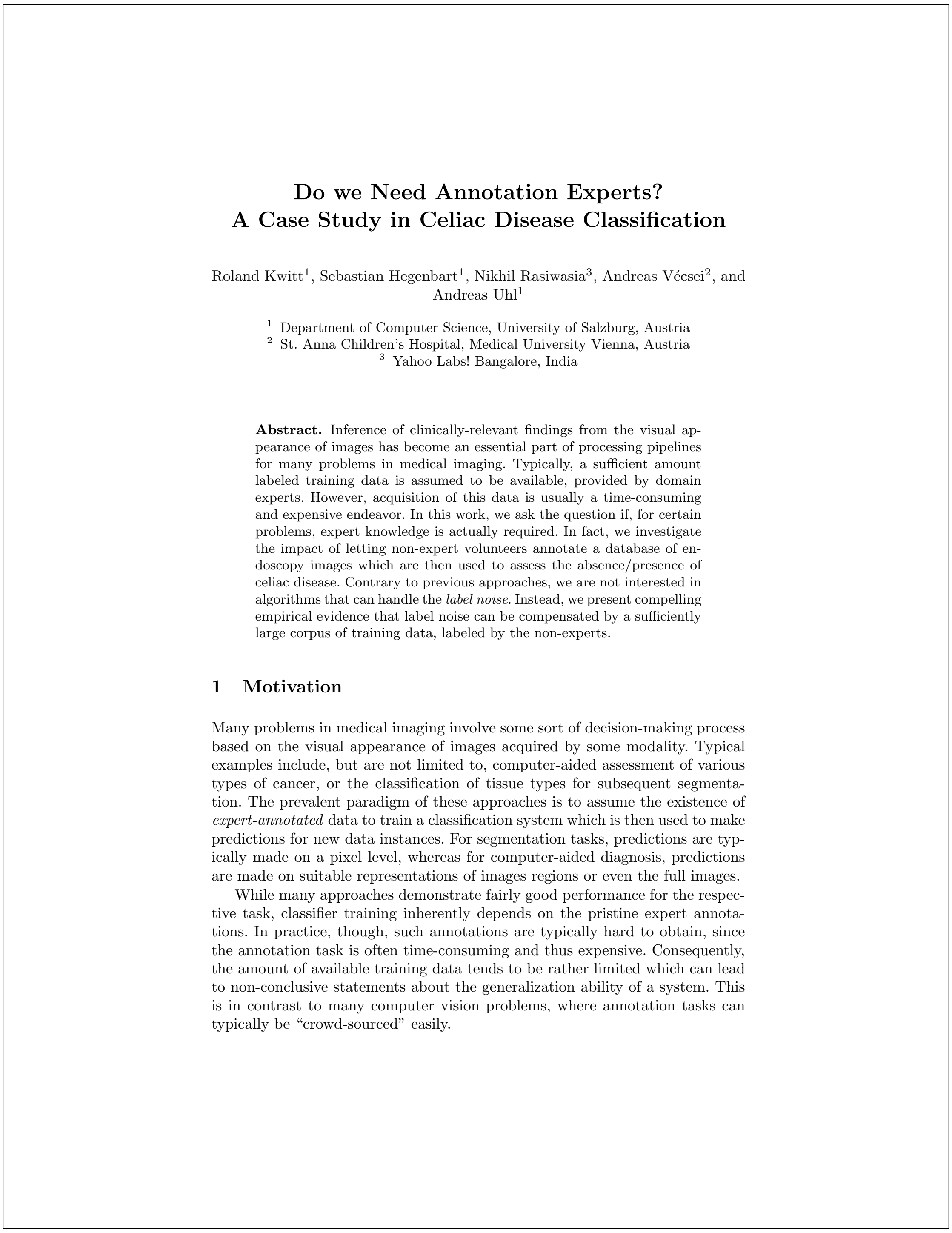
Compared to the distinct set evaluation, the LOOCV method is off by a an average of 8.7 percentage points (outer) as well as 5.35 (inner). The prediction of the LOPO method seems to be more accurate with an average difference of 6.5 percentage points (outer) as well as excellent 1.39 percentage points (inner). Considering the results of the k-fold cross validations a significant variance of



the rates at low number of folds is observed. In general the standard deviation is below one percentage point for both approaches. If k-fold validation is applied we suggest using a fixed number of folds for all experiments to avoid an additional effect of over-fitting. To avoid biased image sets caused by multiple images from a patient the LOPO method should be preferred whenever possible. In general the LOPO method combined with inner-optimization seems to be the most adequate approach if no distinct sets for training and evaluation can be constructed.

References

1. André, B., Vercauteren, T., Wallace, M.B., Buchner, A.M., Ayache, N.: Endomicroscopic video retrieval using mosaicing and visual words. In: Proceedings of the 7th IEEE International Symposium on Biomedical Imaging. IEEE (2010), to appear
2. Gasbarrini, A., Ojetti, V., Cuoco, L., Cammarota, G., Migneco, A., Armuzzi, A., Pola, P., Gasbarrini, G.: Lack of endoscopic visualization of intestinal villi with the immersion technique in overt atrophic celiac disease. Gastrointestinal endoscopy 57, 348–351 (2003)
3. Hegenbart, S., Kwitt, R., Liedlgruber, M., Uhl, A., Vécsei, A.: 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. pp. 718–723. Salzburg, Austria (2009)
4. Huang, X., Li, S., Wang, Y.: Shape localization based on statistical method using extended local binary pattern. In: Proceedings of the 3rd International Conference on Image and Graphics. pp. 1–4. Hong Kong, China (2004)
5. Jain, A., Zongker, D.: Feature selection: Evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 153–158 (1997)
6. Mäenpää, T.: The Local Binary Pattern Approach to Texture Analysis - Extensions and Applications. Ph.D. thesis, University of Oulu (2003)
7. Mäenpää, T., Ojala, T., Pietikäinen, M., Soriano, M.: Robust texture classification by subsets of local binary patterns. In: Proceedings of the 15th International Conference on Pattern Recognition. vol. 3, p. 3947. IEEE Computer Society, Los Alamitos, CA, USA (2000)
8. Oberhuber, G., Granditsch, G., Vogelsang, H.: The histopathology of coeliac disease: time for a standardized report scheme for pathologists. European Journal of Gastroenterology and Hepatology 11, 1185–1194 (1999)
9. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on feature distributions. Pattern Recognition 29(1), 51–59 (1996)
10. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution Gray-Scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002)
11. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. In: AMFG. LNCS, vol. 4778, pp. 168–182. Springer (2007)



Do we Need Annotation Experts? A Case Study in Celiac Disease Classification

Roland Kwitt[1], Sebastian Hegenbart[1], Nikhil Rasiwasia[3], Andreas Vécsei[2], and Andreas Uhl[1]

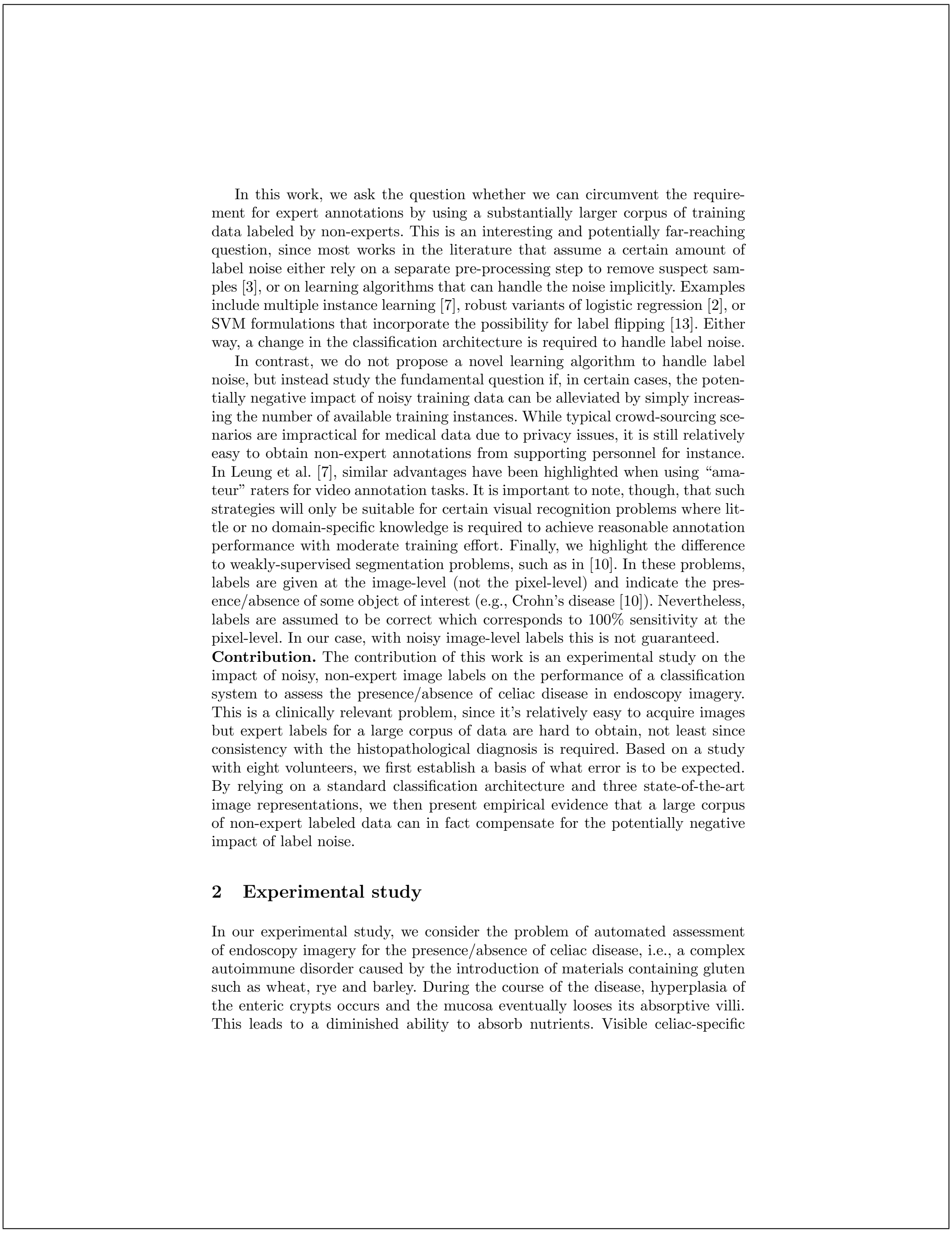
[1] Department of Computer Science, University of Salzburg, Austria
[2] St. Anna Children's Hospital, Medical University Vienna, Austria
[3] Yahoo Labs! Bangalore, India

Abstract. Inference of clinically-relevant findings from the visual appearance of images has become an essential part of processing pipelines for many problems in medical imaging. Typically, a sufficient amount labeled training data is assumed to be available, provided by domain experts. However, acquisition of this data is usually a time-consuming and expensive endeavor. In this work, we ask the question if, for certain problems, expert knowledge is actually required. In fact, we investigate the impact of letting non-expert volunteers annotate a database of endoscopy images which are then used to assess the absence/presence of celiac disease. Contrary to previous approaches, we are not interested in algorithms that can handle the *label noise*. Instead, we present compelling empirical evidence that label noise can be compensated by a sufficiently large corpus of training data, labeled by the non-experts.

1 Motivation

Many problems in medical imaging involve some sort of decision-making process based on the visual appearance of images acquired by some modality. Typical examples include, but are not limited to, computer-aided assessment of various types of cancer, or the classification of tissue types for subsequent segmentation. The prevalent paradigm of these approaches is to assume the existence of *expert-annotated* data to train a classification system which is then used to make predictions for new data instances. For segmentation tasks, predictions are typically made on a pixel level, whereas for computer-aided diagnosis, predictions are made on suitable representations of images regions or even the full images.

While many approaches demonstrate fairly good performance for the respective task, classifier training inherently depends on the pristine expert annotations. In practice, though, such annotations are typically hard to obtain, since the annotation task is often time-consuming and thus expensive. Consequently, the amount of available training data tends to be rather limited which can lead to non-conclusive statements about the generalization ability of a system. This is in contrast to many computer vision problems, where annotation tasks can typically be "crowd-sourced" easily.



In this work, we ask the question whether we can circumvent the requirement for expert annotations by using a substantially larger corpus of training data labeled by non-experts. This is an interesting and potentially far-reaching question, since most works in the literature that assume a certain amount of label noise either rely on a separate pre-processing step to remove suspect samples [3], or on learning algorithms that can handle the noise implicitly. Examples include multiple instance learning [7], robust variants of logistic regression [2], or SVM formulations that incorporate the possibility for label flipping [13]. Either way, a change in the classification architecture is required to handle label noise.

In contrast, we do not propose a novel learning algorithm to handle label noise, but instead study the fundamental question if, in certain cases, the potentially negative impact of noisy training data can be alleviated by simply increasing the number of available training instances. While typical crowd-sourcing scenarios are impractical for medical data due to privacy issues, it is still relatively easy to obtain non-expert annotations from supporting personnel for instance. In Leung et al. [7], similar advantages have been highlighted when using "amateur" raters for video annotation tasks. It is important to note, though, that such strategies will only be suitable for certain visual recognition problems where little or no domain-specific knowledge is required to achieve reasonable annotation performance with moderate training effort. Finally, we highlight the difference to weakly-supervised segmentation problems, such as in [10]. In these problems, labels are given at the image-level (not the pixel-level) and indicate the presence/absence of some object of interest (e.g., Crohn's disease [10]). Nevertheless, labels are assumed to be correct which corresponds to 100% sensitivity at the pixel-level. In our case, with noisy image-level labels this is not guaranteed.

Contribution. The contribution of this work is an experimental study on the impact of noisy, non-expert image labels on the performance of a classification system to assess the presence/absence of celiac disease in endoscopy imagery. This is a clinically relevant problem, since it's relatively easy to acquire images but expert labels for a large corpus of data are hard to obtain, not least since consistency with the histopathological diagnosis is required. Based on a study with eight volunteers, we first establish a basis of what error is to be expected. By relying on a standard classification architecture and three state-of-the-art image representations, we then present empirical evidence that a large corpus of non-expert labeled data can in fact compensate for the potentially negative impact of label noise.

2 Experimental study

In our experimental study, we consider the problem of automated assessment of endoscopy imagery for the presence/absence of celiac disease, i.e., a complex autoimmune disorder caused by the introduction of materials containing gluten such as wheat, rye and barley. During the course of the disease, hyperplasia of the enteric crypts occurs and the mucosa eventually looses its absorptive villi. This leads to a diminished ability to absorb nutrients. Visible celiac-specific

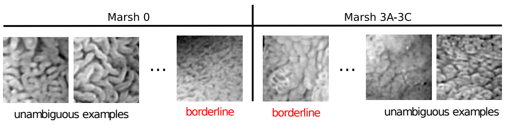


Fig. 1: Typical and borderline examples of images showing non-celiac (Marsh 0) vs. celiac disease (Marsh 3A-3C).

markers that are reported [4] to be characteristic for the pathologic changes of the mucosa include mosaic mucosal patterns, nodular mucosa, scalloping of the duodenal folds, visualization of underlying blood vessels and villous atrophy.

Dataset. Our dataset consists of images acquired during duodenoscopies using pediatric gastroscopes without magnification. 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. 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 [11]. This histological classification scheme identifies six stages of severity of celiac disease, ranging from class Marsh 0 (i.e., no visible change of villi structure) up to class Marsh 3C (i.e., absence of villi). A medical expert assisted in extracting suitable sub-images with a dimension of 128 × 128 pixels from the images captured during endoscopy. Each image shows specific markers for either the absence or presence of celiac disease. While the expert-guided extraction process slightly biases the results, no domain-specific knowledge is needed in practice, since the selection of sub-images is only guided by inspection with respect to certain quality criteria (e.g., sharpness or distortions). Experts were only involved to establish a ground-truth for evaluation purposes.

Our dataset consists of 1050 images from 320 patients with 592 images (240 patients) categorized (consistent with the histopathology) as normal (Marsh 0) and 458 images (80 patients) categorized as containing evidence for celiac disease (Marsh 3A-3C). All images from a single patient have a consistent label. We focus on this, most clinically-relevant binary categorization, as the distinction between all six classes is difficult during endoscopy, even for specialists. This is due to the non-distinct visual appearance of certain classes. A reliable, fine-grained categorization can only be done using histopathology. Some typical and borderline cases for Marsh 0 vs. Marsh 3A-3C cases are shown in Fig. 1.

2.1 Performance of non-expert annotators

To assess the performance of human, non-expert annotators, we randomly selected 100 images with a roughly equal split of *non-celiac* vs. *celiac* instances (60/40). These images were then shown to eight non-expert volunteers, after a 10

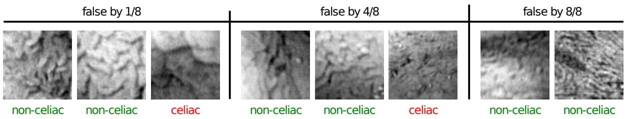


Fig. 2: Labeling erros of human "non-expert" annotators. Images are grouped by errors made by single individuals (left) to errors made by all annotators (right). Images are annotated by their actual ground-truth label.

minute introduction to the annotation problem, where the differences between celiac vs. non-celiac disease were illustrated on an example of 10 typical images per category. This introduction was intended to quickly outline (1) how the disease manifests in architectural changes of the villi and (2) how this affects the visual appearance. Each person then labeled all 100 images individually, *without* knowledge of the class distribution. In addition to the assigned labels, we also recorded the time spent to annotate each image. Interestingly, the mean annotation time per image was only 1.9 seconds with a low standard deviation of ±0.3 seconds. Table 1 lists the accuracy, sensitivity and specificity, averaged over all eight annotators. In our setup, sensitivity corresponds to the percentage of true celiac images actually identified as celiac images by the annotators.

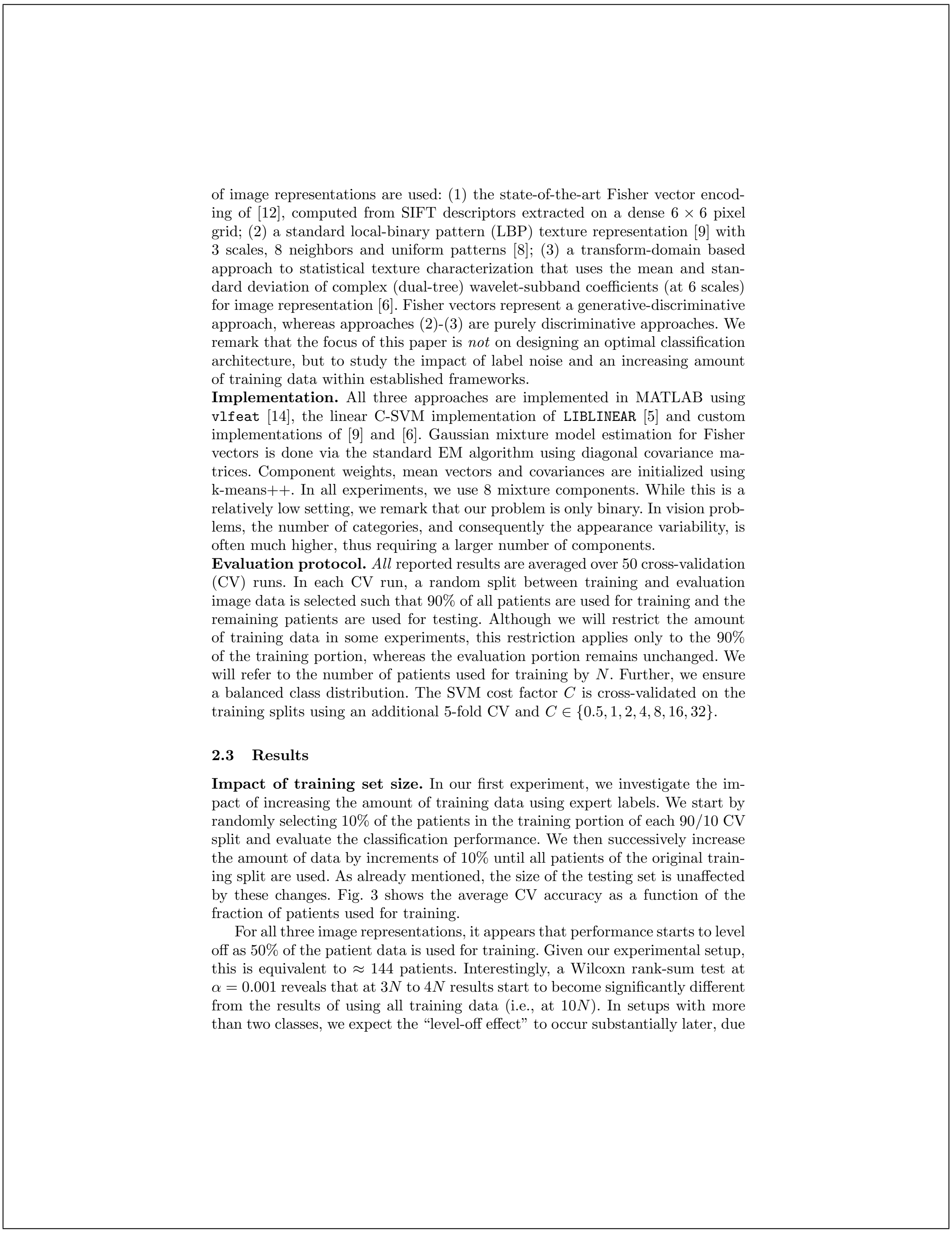
	Accuracy	**Specificity**	**Sensitivity**
AVERAGE	83.8 ± 3.9	81.3 ± 6.7	87.5 ± 9.3
MIN.	79.0	68.3	70.0
MAX.	92.0	90.0	97.5

Table 1: Performance (in %) of eight human "non-expert" annotators.

A closer examination of the annotation errors (with respect to the ground truth) revealed that, out of 100 images, only two images were consistently assigned a false label by all annotators, see Fig. 2 (right). Although, the ground truth label is *non-celiac* in these cases, the presence of the villi is not clearly pronounced making these images hard to categorize without substantial domain knowledge. Some images, falsely labeled by half of the annotators are shown in Fig. 2 (middle). For the *non-celiac* cases, the situation is similar as before in the sense that villi are less pronounced; for the *celiac* image, the non-typical scaling is deceiving and suggests the presence of villi. The left-hand side of Fig. 2 shows images which were falsely labeled only by *single* individuals each. For these images, the appearance is relatively typical for the respective category.

2.2 Classification architecture & Evaluation protocol

We implement a standard classification architecture with a linear support vector machine as a discriminant classifier at the end of the pipeline. Three variants



of image representations are used: (1) the state-of-the-art Fisher vector encoding of [12], computed from SIFT descriptors extracted on a dense 6×6 pixel grid; (2) a standard local-binary pattern (LBP) texture representation [9] with 3 scales, 8 neighbors and uniform patterns [8]; (3) a transform-domain based approach to statistical texture characterization that uses the mean and standard deviation of complex (dual-tree) wavelet-subband coefficients (at 6 scales) for image representation [6]. Fisher vectors represent a generative-discriminative approach, whereas approaches (2)-(3) are purely discriminative approaches. We remark that the focus of this paper is *not* on designing an optimal classification architecture, but to study the impact of label noise and an increasing amount of training data within established frameworks.

Implementation. All three approaches are implemented in MATLAB using `vlfeat` [14], the linear C-SVM implementation of `LIBLINEAR` [5] and custom implementations of [9] and [6]. Gaussian mixture model estimation for Fisher vectors is done via the standard EM algorithm using diagonal covariance matrices. Component weights, mean vectors and covariances are initialized using k-means++. In all experiments, we use 8 mixture components. While this is a relatively low setting, we remark that our problem is only binary. In vision problems, the number of categories, and consequently the appearance variability, is often much higher, thus requiring a larger number of components.

Evaluation protocol. *All* reported results are averaged over 50 cross-validation (CV) runs. In each CV run, a random split between training and evaluation image data is selected such that 90% of all patients are used for training and the remaining patients are used for testing. Although we will restrict the amount of training data in some experiments, this restriction applies only to the 90% of the training portion, whereas the evaluation portion remains unchanged. We will refer to the number of patients used for training by N. Further, we ensure a balanced class distribution. The SVM cost factor C is cross-validated on the training splits using an additional 5-fold CV and $C \in \{0.5, 1, 2, 4, 8, 16, 32\}$.

2.3 Results

Impact of training set size. In our first experiment, we investigate the impact of increasing the amount of training data using expert labels. We start by randomly selecting 10% of the patients in the training portion of each 90/10 CV split and evaluate the classification performance. We then successively increase the amount of data by increments of 10% until all patients of the original training split are used. As already mentioned, the size of the testing set is unaffected by these changes. Fig. 3 shows the average CV accuracy as a function of the fraction of patients used for training.

For all three image representations, it appears that performance starts to level off as 50% of the patient data is used for training. Given our experimental setup, this is equivalent to ≈ 144 patients. Interestingly, a Wilcoxn rank-sum test at $\alpha = 0.001$ reveals that at $3N$ to $4N$ results start to become significantly different from the results of using all training data (i.e., at $10N$). In setups with more than two classes, we expect the "level-off effect" to occur substantially later, due

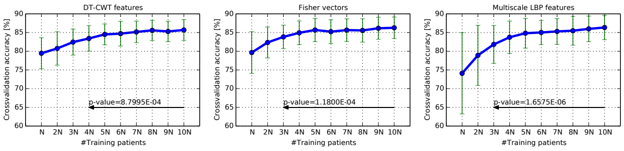


Fig. 3: Impact of increasing the training set size (w.r.t. the # patients), starting from 10% (corresponds to N) of all patients available in the training portion of the data.

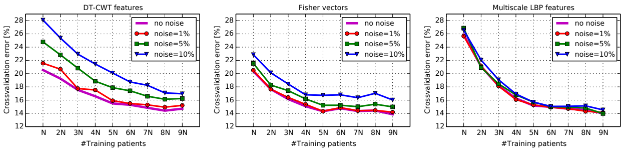


Fig. 4: Impact of different amounts of label noise as a function of the training set size w.r.t. the #patients (best-viewed in color).

to the increased complexity of the problem. In practice, this means that expert labels are required for at least ≈ 480 images (since we have 3 images/patient on avg.) to achieve stable performance.

Impact of artificial label noise. In our second experiment, we study (1) the impact of increasing label noise to simulate non-expert annotators with gradually decreasing performance, and (2) the impact of increasing the number of training samples at the same time. This allows to assess *if*, and to what extent, compensation of label noise can be achieved. We remark, though, that the scenario of *random* label flipping is an unrealistic worst case. In fact, as we have seen in §2.1, labeling errors tend to happen for borderline cases and do not occur totally random. For better illustration, we select the CV error instead of accuracy as an evaluation measure in Fig. 4.

We observe that the positive effect of increasing the training corpus is *not* mitigated by the introduction of label noise. In fact, up to a certain size of pristine training labels, we can achieve equal rates by simple increasing the size of the noisy training corpus. On the example of *DT-CWT features* for instance, $9N$ noisy labels suffice to achieve an error that is comparable to the error achieved by using $5N$ pristine labels (in fact, the null-hypothesis of equal population median cannot be rejected at $\alpha = 0.001$ with a p-value of 0.09). While the magnitude of the impact of label noise seems to be dependent on the image representation, the general behavior remains the same. For the other image representations the compensation effect is even more pronounced.

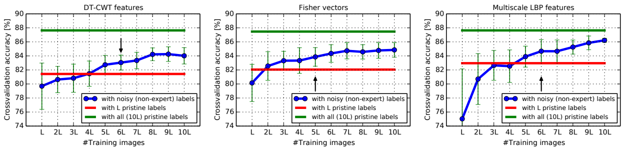


Fig. 5: CV performance as a function of the number of images labeled by non-experts. The baseline result at L is obtained by taking 50/500 images but training is performed using the pristine labels; the performance when *all* 500 pristine training labels are used is indicated by the top line (best-viewed in color).

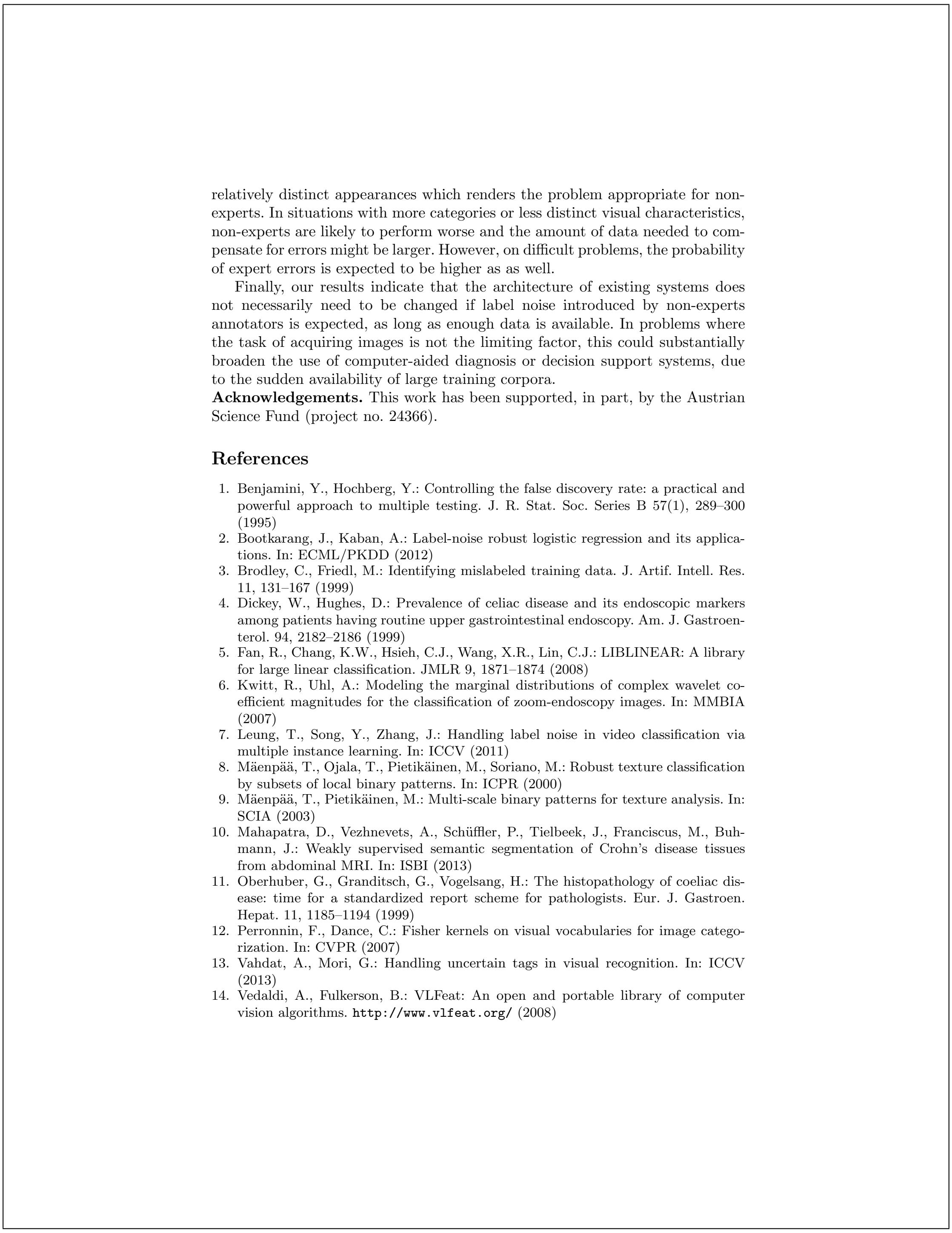
Training with non-expert labels. We consider the actual practical scenario where 500 (randomly chosen) images are labeled by a non-expert, i.e., one volunteer from the experiment in §2.1. The "top annotator" is selected and reaches 89% accuracy on this set. We then compare the classification performance of a system trained with $L = 50$ (10%) of the 500 images using pristine labels vs. systems trained on an increasing number of images with non-expert labels. All remaining $1050 - 500 = 550$ images are used for testing. Results are shown in Fig. 5. We performed a left-tailed Wilcoxn rank-sum test at $\alpha = 0.001$ to assess the null-hypothesis that the population median of the CV results obtained with L pristine labels is *less than* the median of the results obtained with non-expert labels (for each training set size)[4]. The position at which the null-hypothesis cannot be rejected is marked by an arrow. For all three representations this occurs at $6L$ or earlier (with different p-values).

3 Discussion

Given the presented results, several points are worth discussing. First, as we have shown in Fig. 3, a small training corpus with pristine labels does not suffice to achieve stable performance, at least not for the considered problem of celiac disease assessment. In fact, a substantial amount of data is needed until results stabilize and improvements level-off. This effectively shows that limited availability of expert data is an actual problem.

Second, we have presented empirical evidence that a large corpus of non-expert labeled (i.e., noisy) training data can in fact be used to build a classification system that performs equally well as a system trained solely on a limited number of pristine labels. Further, in our particular problem, the relatively good performance of the non-experts reduces the amount of training data required to compensate for the noisy labels. Nevertheless, the question obviously arises how this behavior generalizes to other problems. In our case, visual categories have

[4] We correct for multiple comparisons using the Benjamini-Hochberg [1] procedure to control for FDR.



relatively distinct appearances which renders the problem appropriate for non-experts. In situations with more categories or less distinct visual characteristics, non-experts are likely to perform worse and the amount of data needed to compensate for errors might be larger. However, on difficult problems, the probability of expert errors is expected to be higher as as well.

Finally, our results indicate that the architecture of existing systems does not necessarily need to be changed if label noise introduced by non-experts annotators is expected, as long as enough data is available. In problems where the task of acquiring images is not the limiting factor, this could substantially broaden the use of computer-aided diagnosis or decision support systems, due to the sudden availability of large training corpora.

Acknowledgements. This work has been supported, in part, by the Austrian Science Fund (project no. 24366).

References

1. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B 57(1), 289–300 (1995)
2. Bootkarang, J., Kaban, A.: Label-noise robust logistic regression and its applications. In: ECML/PKDD (2012)
3. Brodley, C., Friedl, M.: Identifying mislabeled training data. J. Artif. Intell. Res. 11, 131–167 (1999)
4. Dickey, W., Hughes, D.: Prevalence of celiac disease and its endoscopic markers among patients having routine upper gastrointestinal endoscopy. Am. J. Gastroenterol. 94, 2182–2186 (1999)
5. Fan, R., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. JMLR 9, 1871–1874 (2008)
6. Kwitt, R., Uhl, A.: Modeling the marginal distributions of complex wavelet coefficient magnitudes for the classification of zoom-endoscopy images. In: MMBIA (2007)
7. Leung, T., Song, Y., Zhang, J.: Handling label noise in video classification via multiple instance learning. In: ICCV (2011)
8. Mäenpää, T., Ojala, T., Pietikäinen, M., Soriano, M.: Robust texture classification by subsets of local binary patterns. In: ICPR (2000)
9. Mäenpää, T., Pietikäinen, M.: Multi-scale binary patterns for texture analysis. In: SCIA (2003)
10. Mahapatra, D., Vezhnevets, A., Schüffler, P., Tielbeek, J., Franciscus, M., Buhmann, J.: Weakly supervised semantic segmentation of Crohn's disease tissues from abdominal MRI. In: ISBI (2013)
11. Oberhuber, G., Granditsch, G., Vogelsang, H.: The histopathology of coeliac disease: time for a standardized report scheme for pathologists. Eur. J. Gastroen. Hepat. 11, 1185–1194 (1999)
12. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: CVPR (2007)
13. Vahdat, A., Mori, G.: Handling uncertain tags in visual recognition. In: ICCV (2013)
14. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/` (2008)

Bibliography

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

[2] Ameling, S., Wirth, S., Paulus, D., Lacey, G., and Vilarino, F. Texture-based polyp detection in colonoscopy. In *Bildverarbeitung für die Medizin 2009*, no. 15 in Informatik Aktuell. Springer Berlin, June 2009, pp. 346 – 350.

[3] Bonamico, M., Mariani, P., Thanasi, E., Ferri, M., Nenna, R., Tiberti, C., Mazzilli, M., and Magliocca, F. Patchy villous atrophy of the duodenum in childhood celiac disease. *J. Pediatr. Gastroenterol. Nutr. 38*, 2 (2004), 204 – 207.

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

[5] Cammarota, G., Cuoco, L., Cesaro, P., et al. 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 39*, 1 (January 2007), 46 – 51.

[6] Cammarota, G., Martino, A., and Pirozzi, G. Direct visualization of intestinal villi by high-resolution magnifying upper endoscopy: a validation study. *Gastrointest. Endosc. 60*, 5 (2004), 732 – 738.

[7] Chand, N., and Mihas, A. A. Celiac disease: Current concepts in diagnosis and treatment. *J. Clin. Gastroenterol 40*, 1 (January 2006), 3 – 14.

[8] Ciaccio, E., C.A., T., G., B., S., L., and P., G. Classification of videocapsule endoscopy image patterns: comparative analysis between patients with celiac disease and normal individuals. *BioMedical Engineering OnLine 9*, 1 (2010).

[9] Ciaccio, E., Tennyson, C., G., B., S.K., L., and Green, P. Robust spectral analysis of videocapsule images acquired from celiac disease patients. *BioMedical Engineering OnLine 10*, 1 (2011).

[10] Ciaccio, E., Tennyson, C. A., Bhagat, G., Lewis, S., and Green, P. Quantitative estimates of motility from videocapsule endoscopy are useful to discern celiac patients from controls. *Digestive Diseases and Sciences 57*, 11 (2012), 2936–2943.

[11] Ciaccio, E., Tennyson, C. A., Bhagat, G., Lewis, S., and Green, P. Use of shape-from-shading to estimate three-dimensional architecture in the small intestinal lumen of celiac and control patients. *Computer Methods and Programs in Biomedicine 111*, 3 (2013), 676 – 684.

[12] CIACCIO, E.J. AND TENNYSON, C. A. AND BHAGAT, G. AND LEWIS, S.K. AND GREEN, P.HR. Transformation of videocapsule images to detect small bowel mucosal differences in celiac versus control patients. *Computer Methods and Programs in Biomedicine 108*, 1 (2012), 28 – 37.

[13] EMURA, F., SAITO, Y., AND IKEMATSU, H. Narrow-band imaging optical chromocolonoscopy: advantages and limitations. *World J. Gastroenterol. 14*, 31 (August 2008), 4867–4872.

[14] ENSARI, A. Gluten-sensitive enteropathy (celiac disease): controversies in diagnosis and classification. *Arch. Pathol. Lab. Med. 134* (2010), 826–836.

[15] 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., AND HORVATH, K. Prevalence of celiac disease in at-risk and not-at-risk groups in the united states: a large multicenter study. *Arch. Intern Med. 163* (2003), 286 – 92.

[16] GADERMAYR, M., LIEDLGRUBER, M., UHL, A., AND VÉCSEI, A. Evaluation of different distortion correction methods and interpolation techniques for an automated classification of celiac disease. *Computer Methods and Programs in Biomedicine 112*, 3 (2013), 694 – 712.

[17] GADERMAYR, M., LIEDLGRUBER, M., UHL, A., AND VÉCSEI, A. Problems in distortion corrected texture classification and the impact of scale and interpolation. In *ICIAP* (2013), vol. 8156 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 513–522.

[18] GADERMAYR, M., LIEDLGRUBER, M., UHL, A., AND VÉCSEI, A. Shape curvature histogram: A shape feature for celiac disease diagnosis. In *Medical Computer Vision. Large Data in Medical Imaging*, Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 175–184.

[19] GADERMAYR, M., UHL, A., AND VÉCSEI, A. Barrel-type distortion compensated fourier feature extraction. In *Advances in Visual Computing*, vol. 8033 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, pp. 50–59.

[20] GADERMAYR, M., UHL, A., AND VÉCSEI, A. Distortion adaptive image classification an alternative to barrel-type distortion correction. In *Advances in Visual Computing* (2013), vol. 8034 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 465–474.

[21] GADERMAYR, M., UHL, A., AND VÉCSEI, A. The effect of endoscopic lens distortion correction on physicians diagnosis performance. In *Bildverarbeitung für die Medizin 2014*, Informatik aktuell. Springer Berlin Heidelberg, 2014, pp. 174–179.

[22] GADERMAYR, M., UHL, A., AND VÉCSEI, A. Feature extraction with intrinsic distortion correction in celiac disease imagery: No need for rasterization. In *Medical Computer Vision. Large Data in Medical Imaging* (2014), Lecture Notes in Computer Science, Springer International Publishing, pp. 196–204.

[23] GADERMAYR, M., UHL, A., AND VÉCSEI, A. Is a precise distortion estimation needed for computer aided celiac disease diagnosis? In *Image and Signal Processing* (2014), vol. 8509 of *Lecture Notes in Computer Science*, Springer International Publishing, pp. 620–628.

[24] GRISAN, E., MIRZAEI, H., AND LEONG, R. Computer-assisted automated image recognition of celiac disease using confocal endomicroscopy. In *ISBI* (April 2014), pp. 121–124.

[25] Gschwandtner, M., Liedlgruber, M., Uhl, A., and Veécsei, A. Experimental study on the impact of endoscope distortion correction on computer-assisted celiac disease diagnosis. In *ITAB* (Nov 2010), pp. 1–6.

[26] Hämmerle-Uhl, J., Höller, Y., Uhl, A., and Vécsei, A. Endoscope distortion correction does not (easily) improve mucosa-based classification of celiac disease. In *Medical Image Computing and Computer-Assisted Intervention MICCAI 2012* (2012), vol. 7512 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 574–581.

[27] Hegenbart, S., and Uhl, A. A Scale-Adaptive Extension to Methods based on LBP using Scale-Normalized Laplacian of Gaussian Extrema in Scale-Space. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '14)* (2014), pp. 4352 – 4356.

[28] Hegenbart, S., and Uhl, A. An Orientation-Adaptive Extension to Scale-Adaptive Local Binary Patterns. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR'14)* (2014), pp. 1120 – 1125.

[29] Hegenbart, S., Uhl, A., and Vécsei, A. Impact of Endoscopic Image Degradations on LBP based Features using One-Class SVM for Classification of Celiac Disease. In *Proceedings of the 7th International Symposium on Image and Signal Processing and Analysis (ISPA'11)* (Dubrovnik, Croatia, 2011), pp. 715 – 720.

[30] Hegenbart, S., Uhl, A., and Vécsei, A. Impact of Histogram Subset Selection on Classification using Multiscale LBP. In *Proceedings of Bildverarbeitung für die Medizin 2011 (BVM'11)* (Lübeck, Germany, 2011), Informatik aktuell, pp. 359 – 363.

[31] Hegenbart, S., Uhl, A., and Vécsei, A. Systematic Assessment of Performance Prediction Techniques in Medical Image Classification - A Case Study on Celiac Disease. In *Proceedings of the 22nd International Conference on Information Processing in Medical Imaging (IPMI'11)* (Monastery Irsee, Germany, 2011), pp. 498 – 508.

[32] Hegenbart, S., Uhl, A., and Vécsei, A. On the Implicit Handling of Varying Distances and Gastrointestinal Regions in Endoscopic Video Sequences with Indication for Celiac Disease. In *Proceedings of the IEEE International Symposium on Computer-Based Medical Systems (CBMS'12)* (2012), pp. 1 – 6.

[33] Hegenbart, S., Uhl, A., and Vécsei, A. A Scale- and Orientation-Adaptive Extension of Local Binary Patterns. Tech. Rep. 2014-05, Department of Computer Sciences, University of Salzburg, Austria, 2014. `http://uni-salzburg.at/index.php?id=38565`; Submitted to Elsevier Journal on Pattern Recognition (October 2014): Under Review.

[34] Hegenbart, S., Uhl, A., Vécsei, A., and Wimmer, G. On the Effects of De-Interlacing on the Classification Accuracy of Interlaced Endoscopic Videos with Indication for Celiac Disease. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems (CBMS'13)* (2013), pp. 137 – 142.

[35] Hegenbart, S., Uhl, A., Vécsei, A., and Wimmer, G. Scale Invariant Texture Descriptors for Classifying Celiac Disease. *Medical Image Analysis 17*, 4 (2013), 458 – 474.

[36] Hopper, A., Cross, S., and Sanders, D. Patchy villous atrophy in adult patients with suspected gluten-sensitive enteropathy: is a multiple duodenal biopsy strategy appropriate? *Endoscopy 40*, 3 (2008), 219 – 224.

[37] Huang, X., Li, S., and Wang, Y. 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, 2004), pp. 1–4.

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

[39] Karkanis, S. A., Iakovidis, D. K., Maroulis, D. E., Karras, D. A., and Tzivras, M. Computer-aided tumor detection in endoscopic video using color wavelet features. *IEEE Transactions on Information Technology in Biomedicine 7*, 3 (Sept. 2003), 141 – 152.

[40] Kwitt, R., Hegenbart, S., Rasiwasia, N., Vécsei, A., and Uhl, A. Do we Need Annotation Experts? A Case Study in Celiac Disease Classification. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'14)* (2014), pp. 454 – 461.

[41] Leong, R., N.Q.Nguyen, Meredith, C., Al-Sohaily, S., Kukic, D., Delaney, P., Murr, E., Yong, J., Merret, N., and Biankin, A. In vivo confocal endomicroscopy in the diagnosis and evaluation of celiac disease. *Gastroenterology 135*, 6 (2008), 1870 – 1876.

[42] Oberhuber, G., Granditsch, G., and Vogelsang, H. The histopathology of coeliac disease: time for a standardized report scheme for pathologists. *European Journal of Gastroenterology and Hepatology 11* (Nov. 1999), 1185–1194.

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

[44] Ojala, T., Pietikäinen, M., and Mäenpää, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 7 (July 2002), 971–987.

[45] Petroniene, R., Dubcenco, E., and Baker, J. Given capsule endoscopy in celiac disease: evaluation of diagnostic accuracy and interobserver agreement. *Am. J. Gastroenterol. 100*, 3 (2005), 685 – 694.

[46] Tan, X., and Triggs, B. Enhanced local texture feature sets for face recognition under difficult lighting conditions. In *Analysis and Modelling of Faces and Gestures* (2007), pp. 168–182.

[47] Uhl, A., Vécsei, A., and Wimmer, G. Complex wavelet transform variants in a scale invariant classification of celiac disease. In *Pattern Recognition and Image Analysis* (2011), vol. 6669 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 742–749.

[48] Valitutti, F., Oliva, S., Iorfida, D., Aloi, M., Gatti, S., Trovato, C. M., Montuori, M., Tiberti, A., Cucchiara, S., and Nardo, G. D. Narrow band imaging combined with water immersion technique in the diagnosis of celiac disease. *Dig. and Liver Dis.*, 0 (2014), –.

[49] Vécsei, A., Amann, G., Hegenbart, S., Liedlgruber, M., and Uhl, A. Automated Marsh-like Classification of Celiac Disease in Children using Optimized Local Texture Operators. *Computers in Biology and Medicine 41*, 6 (2011), 313 – 325.

[50] VÉCSEI, A., FUHRMANN, T., LIEDLGRUBER, M., BRUNAUER, L., PAYER, H., AND UHL, A. Automated classification of duodenal imagery in celiac disease using evolved fourier feature vectors. *Computer Methods and Programs in Biomedicine 95* (2009), 68 – 78.

[51] VÉCSEI, A., FUHRMANN, T., AND UHL, A. Towards automated diagnosis of celiac disease by computer-assisted classification of duodenal imagery. In *MEDSIP* (2008), pp. 1–4. paper no P2.1-009.

[52] WIKIMEDIA-COMMONS. Coeliac disease, 2006. Original uploader: WikipedianProlific at en.wikipedia, later version uploaded by Falcorian at en.wikipedia.

A. Appendix

This section presents a breakdown of the authors' contributions with respect to the papers included in this thesis. Author names are listed in alphabetical order. The explicit contribution of the thesis advisor (Andreas Uhl), medical experts (Andreas Vécsei and Gabriele Amann) as well as a consultant (Nikhil Rasiwasia) can not be stated for a single paper. Although the work would not have been possible without them, I only quantify the contributions of authors directly involved in experimentation or writing each specific publication.

	Contribution (in %)			
Publication	Sebastian Hegenbart	Roland Kwitt	Michael Liedlgruber	Georg Wimmer
Methods for Texture Classification				
VÉCSEI, A., AMANN, G., HEGENBART, S., LIEDLGRUBER, M., AND UHL, A. Automated Marsh-like Classification of Celiac Disease in Children using Optimized Local Texture Operators. *Computers in Biology and Medicine 41*, 6 (2011), 313 – 325	70		30	
HEGENBART, S., UHL, A., VÉCSEI, A., AND WIMMER, G. Scale Invariant Texture Descriptors for Classifying Celiac Disease. *Medical Image Analysis 17*, 4 (2013), 458 – 474	30			70
HEGENBART, S., AND UHL, A. A Scale-Adaptive Extension to Methods based on LBP using Scale-Normalized Laplacian of Gaussian Extrema in Scale-Space. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '14)* (2014), pp. 4352 – 4356	100			
HEGENBART, S., AND UHL, A. An Orientation-Adaptive Extension to Scale-Adaptive Local Binary Patterns. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR'14)* (2014), pp. 1120 – 1125	100			
HEGENBART, S., UHL, A., AND VÉCSEI, A. A Scale- and Orientation-Adaptive Extension of Local Binary Patterns. Tech. Rep. 2014-05, Department of Computer Sciences, University of Salzburg, Austria, 2014. `http://uni-salzburg.at/index.php?id=38565`; Submitted to Elsevier Journal on Pattern Recognition (October 2014): Under Review	100			

	Contribution (in %)			
Publication	Sebastian Hegenbart	Roland Kwitt	Michael Liedlgruber	Georg Wimmer
Characteristic Properties of Duodenal Images and Videos				
Hegenbart, S., Uhl, A., and Vécsei, A. Impact of Endoscopic Image Degradations on LBP based Features using One-Class SVM for Classification of Celiac Disease. In *Proceedings of the 7th International Symposium on Image and Signal Processing and Analysis (ISPA'11)* (Dubrovnik, Croatia, 2011), pp. 715 – 720	100			
Hegenbart, S., Uhl, A., and Vécsei, A. On the Implicit Handling of Varying Distances and Gastrointestinal Regions in Endoscopic Video Sequences with Indication for Celiac Disease. In *Proceedings of the IEEE International Symposium on Computer-Based Medical Systems (CBMS'12)* (2012), pp. 1 – 6	100			
Hegenbart, S., Uhl, A., Vécsei, A., and Wimmer, G. On the Effects of De-Interlacing on the Classification Accuracy of Interlaced Endoscopic Videos with Indication for Celiac Disease. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems (CBMS'13)* (2013), pp. 137 – 142	70			30
Classification and Performance Prediction of Medical Data				
Hegenbart, S., Uhl, A., and Vécsei, A. Impact of Histogram Subset Selection on Classification using Multiscale LBP. In *Proceedings of Bildverarbeitung für die Medizin 2011 (BVM'11)* (Lübeck, Germany, 2011), Informatik aktuell, pp. 359 – 363	100			
Hegenbart, S., Uhl, A., and Vécsei, A. Systematic Assessment of Performance Prediction Techniques in Medical Image Classification - A Case Study on Celiac Disease. In *Proceedings of the 22nd International Conference on Information Processing in Medical Imaging (IPMI'11)* (Monastery Irsee, Germany, 2011), pp. 498 – 508	100			
Kwitt, R., Hegenbart, S., Rasiwasia, N., Vécsei, A., and Uhl, A. Do we Need Annotation Experts? A Case Study in Celiac Disease Classification. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'14)* (2014), pp. 454 – 461	50	50		