Getting One Step Closer to Fully Automatized Celiac Disease Diagnosis

Michael Gadermayr*, Andreas Uhl* and Andreas Vécsei†
* Department of Computer Sciences, University of Salzburg, Austria
e-mail: mgadermayr@cosy.sbg.ac.at, uhl@cosy.sbg.ac.at
† St. Anna Children's Hospital, Department of Pediatrics, Medical University Vienna, Austria

***Abstract*—Up to now, for computer aided celiac disease diagnosis reliable subimages showing discriminative features must be manually extracted by the physicians, prior to the automatized classification. This must be done to get idealistic data which is free from image degradations, in order to enable a reliable computer based classification. However, this interactive stage during medical treatment requires significant time and attention of the physical doctor. Furthermore, an inadequate selection (e.g. of an inexperienced doctor) leads to a decreased classification accuracy. In this work, a method is proposed to select reliable subimages from the original endoscopic images by maximizing a quality measure. Therefore, for the specific problem definition, we introduce five measures which are supposed to be appropriate for reflecting the adequateness of a subimage, with respect to a specific degradation type. Moreover, as none of the single metrics is able to reflect all prevalent degradations, we propose a weighted combination of these metrics. Extensive experiments have been done with five feature extraction techniques, that turned out to be appropriate for celiac disease diagnosis. Finally the best accuracies are achieved by the metric based on the weighted combination.**

***Keywords*—Decision support system, non-interactive, celiac disease, patch selection, quality measurement, feature extraction**

I. Introduction

Celiac disease [1], which is commonly known as gluten intolerance, is a disorder that affects the small intestine after introduction of gluten containing food. The disease leads to an inflammatory reaction in the mucosa of the small bowel caused by a dysregulated immune response triggered by ingested gluten proteins of certain cereals. During the course of celiac disease, the mucosa loses its absorptive villi and hyperplasia of the enteric crypts occurs, leading to a strongly diminished ability to absorb nutrients. According to a large study [2], the overall prevalence of the disease in the USA is 1:133. Figure 1 shows example images, captured during endoscopy.

Up to now, for computer aided celiac disease diagnosis [3], [4], [5], [6], [7] reliable subimages (e.g. patches with a size of 128×128 pixels) showing discriminative features must be manually extracted by the physicians, prior to the automatized classification. This has to be done to get idealistic patches which are free from any image degradations, in order to enable a reliable computer based classification. However, this interactive step prevents the decision support system from being totally automatized. Thereby, significant time and attention of the physical doctor is required and furthermore the classification performance inevitably drops in case of inadequately selected patches [8].

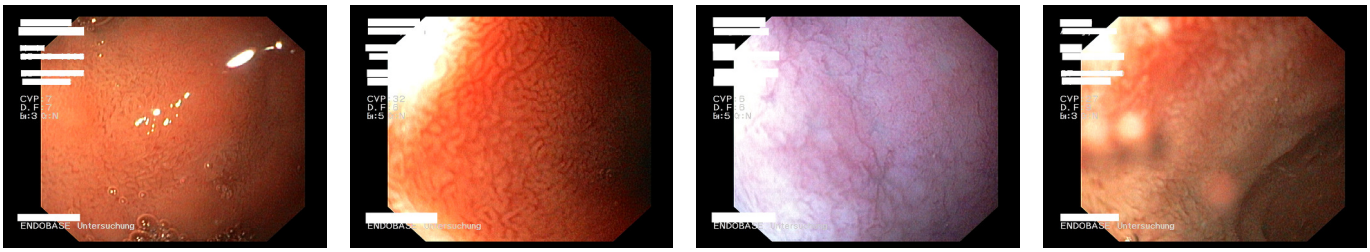
The reason for the decreased classification accuracies in case of randomly or weakly selected patches (or if using the complete images) is the vulnerability of image classification methods to various types of degradations, which are prevalent in endoscopic images [8]. It could have been shown that image degradations definitely affect the feature extraction and consequently lead to a reduced classification accuracy. Such degradations are blur, noise, a lack of contrast and reflections caused by the light of the endoscope.

In other research areas such as computer aided colorectal tumor classification [9] or in melanoma classification [10], image patches are manually extracted in a similar way. However, for these domains, the problem definition is different, as the property to classify is only visible in a certain region which can be effectively determined, whereas in celiac disease diagnosis this property is theoretically visible in large areas, but it is hard to determine reliable regions. As there exists no ground-truth of "good regions" but only a ground truth for the final decision, the adequateness of a patch selection method can only be assessed with respect to a classification method.

Although the current scenario based on idealistic patches might be beneficial for developing feature extraction techniques and classifiers, the manual patch selection stage prevents the decision support system from being totally automatized. Therefore, in this work we investigate if it is possible to select patches for classification in a fully automatized way. To do this, we propose a method to estimate the quality of the image patches which are potentially used in computer based classification. As the quality measure has to deal with many different degradation types, our final metric consists of a weighted sum of simple measures. Each of them is responsible to measure one single degradation type. In the experiments, we evaluate the performance of the proposed combined method as well as the single measures, in order to analyze the impact of the different degradations on the final overall classification accuracy.

The training set in each scenario investigated in this work consists of manually extracted idealistic patches. It should be mentioned that this manual stage can be done beforehand by experts and does not require any interaction during medical treatment. The ground-truth is available for each original image and can be directly taken for all patches extracted from the respective image. The principal aim of this paper is to give insight into the effectiveness of certain measures and not to achieve the best classification rates.

The paper is organized as follows: In Section II, image degradations being prevalent in endoscopic images and image quality measures corresponding to these degradations are in-



(a) (b) (c) (d)

Fig. 1: Example images with varying degradations such as small reflections (1a, center) and even large overexposed sploches (1b, top left), dark regions combined with noise (1b, right and 1d, bottom right) and blurred regions (1c, bottom left and 1d, left).

troduced. In Section III, the experimental results are analyzed and discussed. Finally, Section IV concludes this paper.

II. Patch Selection

As can be seen in Fig. 1, endoscopic images are often affected by significant degradations. In the following we focus on identifying all degradations, prevalent in these images.

- Low illumination combined with low contrast and noise:
  Due to the punctual source of light (caused by a light mounted on the tip of the endoscope), the camera's field of view cannot be illuminated constantly. This is especially the case, if the mucosa is not captured frontally, but in an acute angle (see Fig. 1d, bottom right). Image regions with a low illumination often correspond with a low contrast and with a significantly increased noise level, which is due to the physics of the camera sensor.
- Blurred regions:
  Due to the fixed focus property of the camera, an inappropriate (mostly to small) distance between the mucosa and the lens causes blurred image regions (see Fig. 1b, left and Fig. 1d, top left). Additionally, blur can be introduced by camera motion.
- Reflections and too intensive illuminations:
  Also caused by the punctual source of light, extremely bright splotches (see Fig. 1a, center) or even large overexposed regions (see Fig. 1a, center and Fig. 1b, top left) can be observed in some images.

In [8], the authors showed that the degradations mentioned so far reduce the discriminative content in images, compromise the feature extraction stage and finally lead to reduced classification accuracies. The aim of the image quality methods, introduced in the following is to identify patches with high discriminative content. Therefore, we define measures q, which have to be maximized, when selecting the coordinates for patch selection in the original image.

- The first measure addresses the problem of a too low illumination. As such a weak illumination generally corresponds to images with a low average gray value, we propose a quality measure being based on the mean of the pixel intensities

$$q_A(P) = \frac{1}{|Z|} \cdot \sum_{z \in Z} P(z) \;, \tag{1}$$

  where Z comprises the coordinates of the image patch P.
- The next measure is utilized to detect image regions lacking from any significant gray value differences. Such image patches can be identified by measuring the contrast which is defined by

$$q_C(P) = \sum_{i,j \in K} |i-j| \cdot p(i,j) \;, \tag{2}$$

  where K comprises all gray values in P and $p(i,j)$ stands for the probability of these two gray values to be present in a certain image neighborhood in P. In order to focus on real contrast rather than on noise, for this neighborhood we use a quite large offset of four pixels in vertical and in horizontal direction and average these two values.
- The next measure is based on a blur metric b [11]. For computing this metric, first in one direction, the edges ($e \in E$) are identified by extracting all local minima and maxima. Finally the ratio between the lengths (l) and the pixel differences (δ) of the edges is computed, which indicates the blur level. Formally written, blur is defined by

$$b = \frac{1}{|E|} \sum_{e \in E} \frac{l(e)}{\delta(e)} \;. \tag{3}$$

  As all of our images suffer from more or less significant sensor noise, the patches are previously denoised using a Gaussian filter G_2 with $\sigma = 2$. Finally for blur measurement, we use

$$q_B(P) = -b(P * G_2) \;, \tag{4}$$

  which has a negative sign as we strive for non-blurred images.
- To detect noisy image patches, we sum up the differences between the original image and a denoised version of the same image

$$q_N(P) = \sum_{z \in Z} |P - G_1 * P| \;. \tag{5}$$

  The denoised image is achieved by filtering the original image with a Gaussian G_1 with $\sigma = 1$.
- Finally, we need a measure to address the problem of reflections and extremely high illuminations. These

regions can be detected quite easily by considering the maximum gray values.

$$q_I(P) = \begin{cases} 1, & \text{if } max(P) < T \\ 0, & \text{otherwise .} \end{cases} \quad (6)$$

T is set to 245 (where zero refers to black and 255 refers to white), which turned out to be appropriate for separating extremely bright regions (by manual inspection of a set of training images).

As each of the quality measures copes with single image degradations, but none of the measures is able to cope with all of them, we furthermore fuse them to achieve the image patches with highest discriminative powers. Therefore, we normalize (n) the output of each measure (to be within the interval $[0, 1]$) and sum them up:

$$q_F = \sum_i w_i \cdot n(q_i) \; . \quad (7)$$

The parameters w_i are evaluated by using a separate database.

In the experiments, these measures are compared with the manual patch selection (indicated by q_M) on the one hand, and the random measure q_R on the other hand.

III. Experiments

A. Experimental Setup

The image test set used contains images of the duodenal bulb and the pars descendens taken during duodenoscopies at the St. Anna Children's hospital using pediatric gastroscopes (with a resolution of 768×576 (Olympus GIF Q165) and 528×522 pixels (GIF N180), respectively).

The patch selection stage extracts sub-images with a fixed size of 128×128 pixels, in order to be able to compare the results with the manual extraction, which is also based on equally sized patches and is done by a highly experienced endoscopist. For automatized patch selection, the measures q are applied to 16 potential patch candidates per image. The patch with the highest measure is finally selected for feature extraction and classification. In case of a tie, one patch is randomly selected. The weights w_i are evaluated during exhaustive search, based on a separate dataset. The search space for each w_i is between 0.0 and 1.0 with a step-size of 0.2. In our experiments, for feature extraction the patches are converted to gray value images.

To generate the ground-truth, 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 as proposed in [1]. Although it is possible to distinguish between the several stages of the disease, we only aim in distinguishing between images of patients with (Marsh-3) and without the disease (Marsh-0), because this 2-classes case is more relevant in practice. Another incentive for preferring the 2-classes case is that the distinction between the different stages of the disease is considerably subjective even as far as the histological examination is concerned [12]. Thereby, the ground-truth and furthermore the evaluation in a multi-classes case would be less reliable.

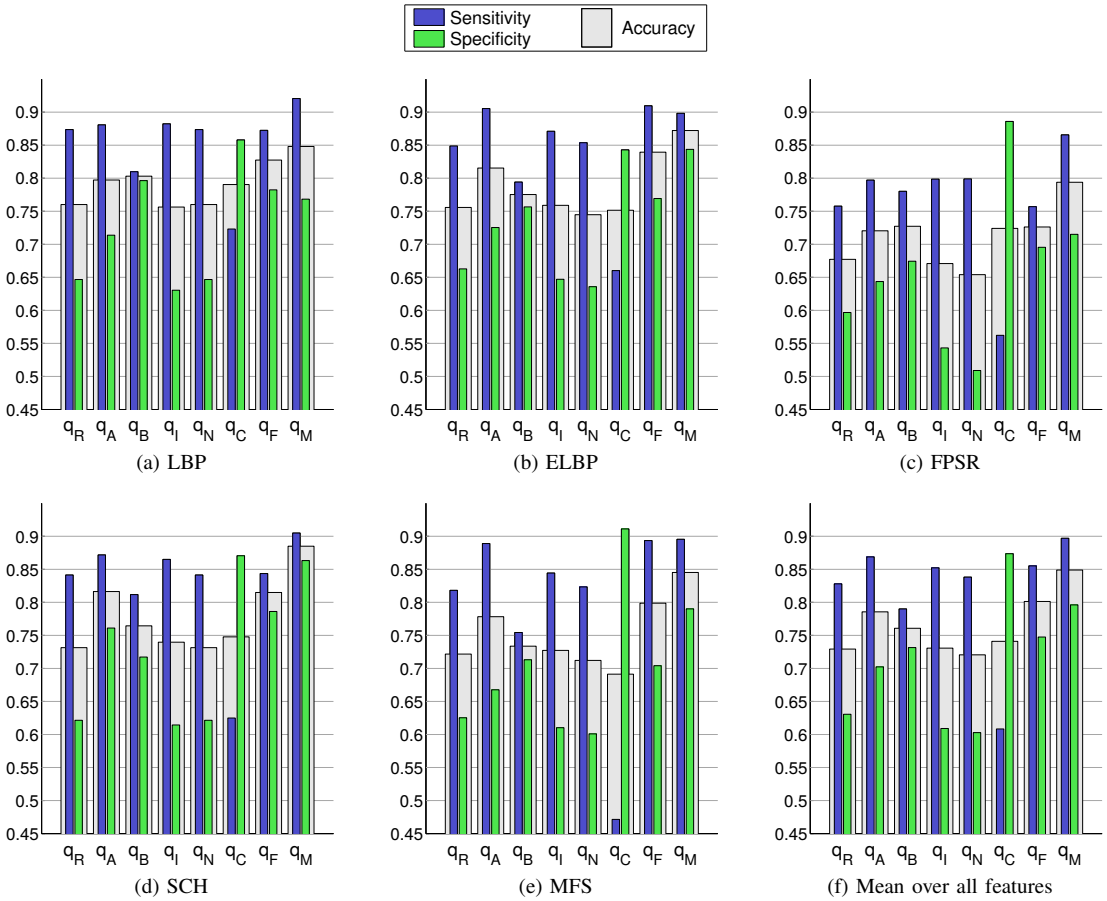
Our experiment is based on one database for parameter optimization consisting of an idealistic patch set and an non-idealistic original image set as well as one database for evaluation consisting of a similarly sized patch and original image set. The patch datasets consist of 300 (151 Marsh-0, 149 Marsh-3) and 312 (155 Marsh-0, 157 Marsh-3) images, respectively. Both original datasets comprise 2752 patch candidates (1376 for each class), automatically extracted from 172 images (i.e. each original image has 16 patch candidates).

For classification, the k-nearest neighbor classifier is used. We utilize this simple classifier in order to focus on the patch selection stage. The rates achieved with k values reaching from one to 30 are averaged, to get more stable and significant results rather than to get the highest possible rates.

B. Feature Extractors

For the experimental analysis, we deploy the following feature extraction techniques, which proved to the adequate for celiac disease classification in previous work [13], [5]:

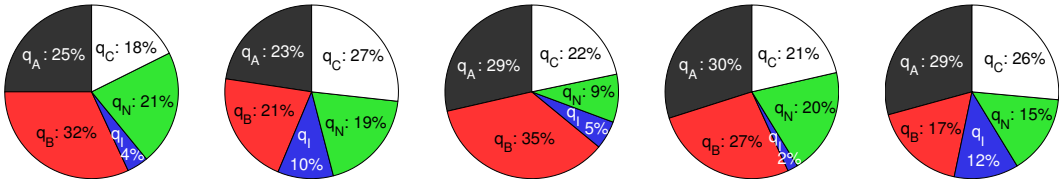
- Local Binary Patterns [14] (LBP):
  The commonly used Local Binary Patterns describe a texture by computing the joint distribution of binarized intensity differences within a certain neighborhood. This widely used feature extraction technique is used with eight neighbors and a radius (i.e. the distance to the neighbors) of two pixels.
- Extended Local Binary Patterns [15] (ELBP):
  ELBP is an edge based derivative of Local Binary Patterns. Instead of capturing intensity differences (as done by LBP), this feature captures differences of the edge magnitude. As LBP it is used with eight neighbors and a radius of two pixels.
- Fourier Power Spectra Rings [16] (FPSR):
  To get this descriptor, first the Fourier power spectra of the image patches are computed, in a way that the low frequencies are in the image center. Afterwards, a ring with a fixed inner and outer radius is extracted and the median of the values in this ring are calculated. For our experiments, we use an inner radius of seven and an outer radius of eight pixels, which turned out to be suitable in previous work [16]. Although this feature has a dimensionality of one, it turned out to be suitable [16] for celiac disease diagnosis.
- Shape Curvature Histogram [17] (SCH):
  SCH is a shape feature, especially developed for celiac disease diagnosis. After detection of significant regions (using the Canny-edge detector [18]), the shape curvatures are computed by measuring the differences of the gradient orientations in a 3×3 neighborhood. Finally a histogram collects the occurrences of the contour curvature values in the significant regions. As in the original work, we consider a histogram bin count of eight that turned out to be optimal for celiac disease diagnosis.
- Multi-Fractal Spectrum [19] (MFS):
  The local fractal dimension is computed for each pixel using three different types of measures for computing the local density. The final feature vector is built by concatenation of these fractal dimensions. Previous work [6] showed that this declared viewpoint invariant feature is suitable for celiac disease diagnosisis.



(a) LBP (b) ELBP (c) FPSR

(d) SCH (e) MFS (f) Mean over all features

Fig. 2: Accuracies, sensitivities and specificities achieved with the specific quality measures separately for each feature.



(a) LBP (b) ELBP (c) FPSR (d) SCH (e) MFS

Fig. 3: Distribution of weights evaluated for the fused quality measure q_F. For each measure q_i, w_i indicates the weighting (relative to the sum of all weights).

C. Results

In Fig. 2, the performances of the patch quality measures are visualized separately for each feature. The wide bars indicate the classification accuracies and the thin bars indicate the sensitivities (left) and the specificities (right).

First, we focus on the accuracies. As anticipated, we notice that the manual patch selection q_M corresponds to the highest accuracies. However, especially the approach q_F, combining all quality measures is able to outperform a random patch selection by far in case of each feature. Moreover, the fusion in each case corresponds to the highest (or equal) rates compared to all of the single quality measures (q_A, q_B, q_I, q_N and q_C). As far as the different features are concerned, the highest accuracies are achieved with the shape based SCH in case of manual patch selection and the edge based ELBP in case of automatized (q_F) patch selection. The highest accuracies considering single quality metrics are achieved with q_A, which measures the average illumination, followed by the blur measure q_B.

If focusing on the balance between sensitivity and specificity, quite interesting aspects can be detected. Considering the random patch selection, in case of all features a high sensitivity faces a very low specificity. This behavior is due to the fact that a low quality patch, especially if suffering from blur, usually has a higher similarity to an image showing diseased mucosa than to a healthy mucosa which shows the villi structure. Even with the manual patch selection, a similar but less distinct behavior can be observed. If focusing on the automatic quality measures, we notice that in case of q_B and especially in case of q_C this behavior is balanced or even reversed. In order to

get at the root of these things, the properties blur and contrast with respect to mucosal images have to be investigated. Images with a high blur level (i.e. the blur measure is low) and a low contrast level show a higher degree of similarity to diseased patches than to healthy patches. Therefore, if forcing a high measure patches with a higher similarity to healthy patches are preferred. In opposite, forcing a large noise measure (i.e. the noise level should be small), patches with a high similarity to diseased patches are preferred. If considering the fused measure q_F, it can be observed that a balance similar to the manual approach is achieved.

In the following we focus on this metric, to estimate the level of participation of the single measures in the fused measure. In Fig. 3, the distribution of the weights of the single quality measures are shown for the specific features. First of all, we see that except from q_I which penalizes reflections and seems to be less important, the measures are quite good balanced and none of the measures is predominant. A high contrast (q_A), a high average illumination (q_A) and a low blur level (q_B) are similar important for all features. Furthermore, we notice that especially a low noise level plays an important role in case of LBP, ELBP and SCH, which operate in a small neighborhood and are based on rather high image frequencies compared to the other two features. In opposite the Fourier based FPSR which extracts quite low frequencies seems to be less dependent on a low noise level.

IV. Conclusion

We have investigated five quality measures and a weighted combination of them. Finally we have found out that especially the combination corresponds to significantly improved accuracies, compared to an inadequate (random) patch selection. The average improvement is more than seven per cent. None of the single measures is able to outperform the combined approach in case of any feature. Consequently, we conclude that a sensible quality measure must incorporate at least some of the single measures. The analysis of weight distributions showed that the importance of the certain measures depends on the respective feature extraction method. Compared to the manual patch selection, the achieved results are on average less than five per cent below. Although the results of the manual selection cannot be achieved, it have to be mentioned that its accuracies would decrease in case of an inappropriate manual selection (e.g. by inexperienced or inattentive medical doctors). By selecting more appropriate patches per image or even some images per patient combined with a decision-level or score-level fusion, we assume that further improvements can be achieved. However, for such an investigation, a larger database consisting of a significant number of images for each patient would be required.

References

[1] 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*, vol. 11, pp. 1185–1194, Nov. 1999.

[2] 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*, vol. 163, pp. 286–92, February 2003.

[3] E. J. Ciaccio, C. A. Tennyson, S. K. Lewis, S. Krishnareddy, G. Bhagat, and P. Green, "Distinguishing patients with celiac disease by quantitative analysis of videocapsule endoscopy images," *Computer Methods and Programs in Biomedicine*, vol. 100, no. 1, pp. 39–48, Oct. 2010.

[4] E. J. Ciaccio, C. A. Tennyson, G. Bhagat, S. K. Lewis, and P. H. R. Green, "Classification of videocapsule endoscopy image patterns: comparative analysis between patients with celiac disease and normal individuals," *BioMedical Engineering Online*, vol. 9, no. 1, pp. 1–12, 2010.

[5] 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*, vol. 41, no. 6, pp. 313–325, Jun. 2011.

[6] 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.

[7] S. Hegenbart, A. Uhl, and A. Vécsei, "Impact of histogram subset selection on classification using multiscale LBP," in *Proceedings of Bildverarbeitung für die Medizin 2011 (BVM'11)*, ser. Informatik aktuell, Lübeck, Germany, March 2011, pp. 359–363.

[8] ——, "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, Sep. 2011, pp. 715–720.

[9] T. Tamaki, J. Yoshimuta, M. Kawakami, B. Raytchev, K. Kaneda, S. Yoshida, YoshitoTakemura, K. Onji, R. Miyaki, and S. Tanaka, "Computer-aided colorectal tumor classification in NBI endoscopy using local features," *Medical Image Analysis*, vol. 17, no. 1, pp. 78 – 100, 2013.

[10] C. Barata, J. S. Marques, and T. Mendonça, "Bag-of-features classification model for the diagnose of melanoma in dermoscopy images using color and texture descriptors," in *Proceedings of the International Conference on Image Analysis and Recognition (ICIAR'13)*, 2013, pp. 547–555.

[11] P. Marziliano, F. Dufaux, S. Winkler, T. Ebrahimi, and G. Sa, "A no-reference perceptual blur metric," in *Proceedings of the IEEE International Conference on Image Processing (ICIP'02)*, 2002, pp. 57–60.

[12] B. Weile, B. F. Hansen, I. Hägerstrand, J. P. H. Hansen, and P. A. Krasilnikoff, "Interobserver variation in diagnosing coeliac disease, a joint study by danish and swedish pathologists," *APMIS*, vol. 108, no. 5, pp. 380–384, 2000.

[13] M. Gadermayr, M. Liedlgruber, A. Uhl, and A. Vécsei, "Evaluation of different distortion correction methods and interpolation techniques for an automated classification of celiac disease," *Computer Methods and Programs in Biomedicine*, vol. 112, no. 3, pp. 694–712, Dec. 2013.

[14] 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.

[15] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Li, "Learning multi-scale block local binary patterns for face recognition," in *Advances in Biometrics*. Springer, 2007, pp. 828–837.

[16] M. Gadermayr, A. Uhl, and A. Vécsei, "Barrel-type distortion compensated fourier feature extraction," in *Proceedings of the 9th International Symposium on Visual Computing (ISVC'13)*, ser. Springer LNCS, vol. 8033, Jul. 2013, pp. 50–59.

[17] M. Gadermayr, M. Liedlgruber, A. Uhl, and A. Vécsei, "Shape curvature histogram: A shape feature for celiac disease diagnosis," in *Medical Computer Vision. Large Data in Medical Imaging (Proceedings of the 3rd International MICCAI - MCV Workshop 2013)*, ser. Springer LNCS, vol. 8331, 2014, pp. 175–184.

[18] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Recognition and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, Sep. 1986.

[19] Y. Xu, H. Ji, and C. Fermüller, "Viewpoint invariant texture description using fractal analysis," *International Journal of Computer Vision*, vol. 83, no. 1, pp. 85–100, 2009.