©Elsevier. This is the authors version of the work. It is posted here by permission of Elsevier for personal use. Not for redistribution or commercial use. The definitive version is available at www.sciencedirect.com.

Disponible en ligne sur
ScienceDirect
www.sciencedirect.com

Elsevier Masson France
EM|consulte
www.em-consulte.com

IRBM

IRBM 37 (2016) 31–39

Medical Image Analysis CAD

Fully automated decision support systems for celiac disease diagnosis

M. Gadermayr [a,*], A. Uhl [b], A. Vécsei [c]

[a] *Institute of Imaging and Computer Vision, RWTH Aachen, Germany*
[b] *Department of Computer Sciences, University of Salzburg, Austria*
[c] *St. Anna Children's Hospital, Department of Pediatrics, Medical University Vienna, Austria*

Received 13 February 2015; received in revised form 22 June 2015; accepted 1 September 2015

Available online 16 October 2015

Abstract

In most recent computer aided celiac disease diagnosis approaches, image regions (patches) showing discriminative features necessarily need to be manually extracted by the medical doctor, prior to the automated classification pipeline. However, although the obtained classification outcomes based on such semi-automated systems are attractive, a human interaction finally is undesired. In this work, fully automated approaches are investigated which are based on the measurement of several image quality properties. Firstly, we investigate a method based on optimization of single quality measures as well as an approach based on weighted combinations of these metrics. Furthermore, a weighted decision-level and a weighted feature-level fusion method are investigated which are not based on the selection of one single best patch, but on a weighted combination. In a large experimental setting, we evaluate these methods with respect to the achieved overall classification rates. Finally, especially the proposed feature-level fusion method supplies the best performances and comes close to manual experts' patch selection as far as the accuracy is concerned.
© 2015 AGBM. Published by Elsevier Masson SAS. All rights reserved.

Keywords: Endoscopy; Celiac disease diagnosis; Fully automated diagnosis; Medical imaging

1. Introduction

1.1. Celiac disease

Celiac disease, also known as gluten intolerance, is a complex autoimmune disorder which affects the small intestine in genetically predisposed individuals of all age groups after the introduction of gluten containing food. Characteristic for this disease is the inflammatory reaction in the mucosa of the small intestine. During the course of the disease the mucosa looses its absorptive villi and hyperplasia of the enteric crypts occurs, leading to a diminished ability to absorb any nutrients.

Endoscopy in combination with biopsy is currently considered as the gold standard for the diagnosis of celiac disease. During standard upper endoscopy at least four biopsies are taken. Microscopic changes within these specimen are then classified in a histological analysis according to the Marsh classification [1]. Subsequently, Oberhuber et al. proposed the modified Marsh classification scheme [2] which distinguishes between classes Marsh-0 to Marsh-3, with subclasses Marsh-3A, Marsh-3B, and Marsh-3C, resulting in a total number of six classes. According to the modified Marsh classification scheme, Marsh-0 denotes a healthy mucosa (without visible changes of the villous structure) and Marsh-3C designates a complete absence of villi (villous atrophy).

In accordance to previous work [3–5], we consider the four classes Marsh-0 and Marsh-3A to Marsh-3C only, since visible changes in the villi structure can be observed only for classes Marsh-3A to Marsh-3C. In this work we focus on the two-class case only (i.e. Marsh-0 and Marsh-3) since if considering this problem definition, the image data set available is well balanced with respect to the images in each class. Furthermore, this two classes case is most relevant for clinical practice.

The overall prevalence [6] of celiac disease in the USA is about one per cent. Fig. 1 shows example images, captured during standard upper endoscopy.

---

[*] Corresponding author at: Institute of Imaging and Computer Vision, RWTH Aachen, Germany.
E-mail address: mgadermayr@cosy.sbg.ac.at (M. Gadermayr).

http://dx.doi.org/10.1016/j.irbm.2015.09.009
1959-0318/© 2015 AGBM. Published by Elsevier Masson SAS. All rights reserved.

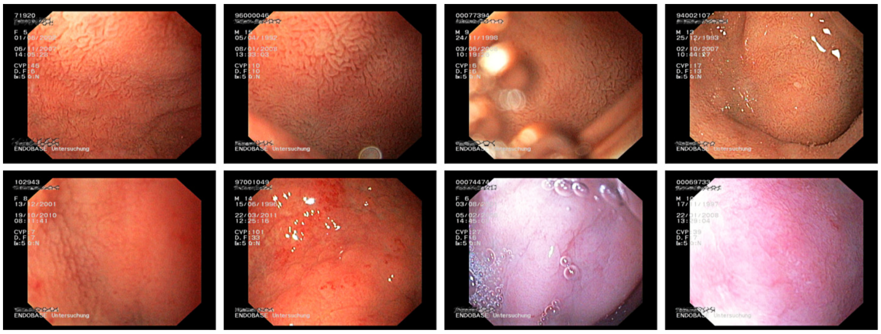


Fig. 1. Endoscopic images of healthy mucosa (top row) clearly showing the villi structure and of diseased mucosa (bottom row). In some regions the markers for celiac disease are visualized well whereas others suffer from strong degradations.

1.2. Computer aided diagnosis

For most computer aided celiac disease diagnosis approaches [3,7–10,4,11], reliable (ideal) image regions (e.g. patches with a size of 128 × 128 pixels) showing discriminative features have to be identified prior to the automated classification. This must be done to get ideal data which is free from strong image degradations as in case of strong degradations the classification accuracy of the decision support system decreases strongly [12,13]. The identification of reliable regions could be done manually [9,10] on the one hand or by means of a computer based method. Existing approach for detection of informative regions [14,15], do not directly focus on a succeeding computer aided diagnosis and are certainly not optimized for celiac disease diagnosis. Although the manual method works effectively if done by experienced medical doctors [13], there are two incentives to use a computer based selection method: First of all, a human interaction during endoscopy is time consuming and annoying which probably leads to a diminished acceptance of the decision support system by physical doctors. Apart from that, especially in case of physicians which are inexperienced, inattentive or just unfamiliar with the (new) decision support system, a weak selection automatically leads to decreased classification accuracies [13]. This can furthermore lead to an even more decreased acceptance of the semi-automated system.

The reason for the decreased classification accuracies in case of randomly or inappropriately 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. Recent work [12] showed that image degradations definitely affect the feature extraction stage and consequently lead to a reduced classification accuracy. Such degradations are blur, noise, a lack of contrast, underexposure and overexposure (reflections). They are potentially prevalent in any real world image data, however, endoscopic images are particularly affected because of the difficult capturing conditions. Blur occurs because the difficult handling does not allow to adjust the distance to the surface (mucosa) precisely. Furthermore motion often cannot be prevented. The small sensors used in the endoscopic devices are prone to noise. This liability is amplified in case of underexposure which is caused by the spotty lightning (as endoscopes are equipped with one or two spotty lights). Unfortunately, these spotty lightning not only leads to underexposed regions, but also to overexposed ones as well as small reflexion (bright spots). Example endoscopic images with various kinds of degradations are shown in Fig. 1.

1.3. Contribution

This article collects the two approaches for fully automated celiac disease diagnosis from our previous publications [13,16]. As previous methods for semi-automated celiac disease diagnosis [3,9,10,4,11] are optimized for manually extracted patches with a size of 128 × 128 pixels, focus of work on fully automated diagnosis [13,16] is on the selection of one or more such reliable patches. These reliable patches are sub-images which clearly show markers for a visual distinction (between class Marsh-0 and class Marsh-3) and which do not strongly suffer from image degradations. Thereby, after the selection of reliable patches, methods for semi-automated celiac disease diagnosis can be used to obtain a final decision.

In both papers on fully automated diagnosis, the availability of theoretically numerous small sub-images (in our case patches with a size of 128 × 128 pixels) in each original endoscopic image (768 × 576 pixels) is exploited in order to extract data for subsequent classification. Consequently, the initially required task is to automatically extract numerous potential sub-images (at fixed, predefined positions) distributed over the original endoscopic image. Then, the availability of large data firstly allows to select one best patch per original image as done in the first work on fully automated celiac disease diagnosis [13]. Additionally, it facilitates a redundant processing [16] (i.e. feature extraction and classification) of these multiple available patches aiming at improving the classification accuracy. In order to generate one final decision for each image, these redundant threads have to be fused. This can be done on different levels [17], such as feature-, score- or decision-level as successfully deployed in biometric systems [18,17,19]. As the simple (unweighted) fusion does not lead to improved accuracies, we utilize patch quality measures to introduce a weighting. Based on this weighting, a weighted decision-level as well as a weighted feature-level fusion method is investigated.

In this work, the techniques for fully automated celiac disease diagnosis are characterized as well as extensively analyzed and compared with each other. Therefore, several novel experimental scenarios are created. Additionally, the required computational runtimes are evaluated and compared. For training of the classification model, we investigate and compare two different scenarios (which has not been done before [13,16]). The first one is based on training with manually extracted ideal patches. The second is based on training with (also) automatically extracted data. It should be mentioned that the manual stage (required in the first scenario) can be done beforehand by experts and does not require any interaction during medical treatment. The ground-truth, which has been determined by histological examination of biopsies, is available for each original image and can be directly taken for all patches extracted from the respective image.

1.4. Outline

The paper is organized as follows: First, in Section 2 the quality measures, which are furthermore required for fully automated celiac disease diagnosis, are introduced. In Section 3 the method based on patch selection as well as the two fusion based approaches are introduced. In Section 4 the experimental results are extensively analyzed and discussed. Finally, Section 5 concludes this paper.

2. Quality metrics

First of all, we define a set of sensible quality measures which are required for our metric based approaches.

- 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 [20]. For computing this metric, first in one direction the edges are identified by extracting all local minima and maxima. Finally the ratio between the lengths and the pixel differences of the edges is computed which indicates the blur level. As all of our images suffer from more or less significant sensor noise, the patches are previously denoised using a Gaussian filter G_2 where σ is 2.0 pixels.

$$q_B(P) = -b(P * G_2)\,. \tag{3}$$

- To detect noisy image patches, we sum up the differences between the original image and a denoised version of the same image

$$q_N(P) = \sum_{z \in Z} |P - G_1 * P|\,. \tag{4}$$

  The denoised image is achieved by filtering the original image with a Gaussian G_1 where σ is 1.0 pixel.
- 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.

$$q_I(P) = \begin{cases} 1, & \text{if } \max(P) < T \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

  T is set to 245 (eight bit gray scale), which turned out to be appropriate for separating extremely bright regions (by manual inspection of a set of training images).

For further processing, these quality measures are min-max-normalized to be within the interval between zero and one. Experiments provide strong evidence that any single quality measure is unable to represent the quality of a patch with respect to the classification performance. Therefore, we do not focus on single measures but instead introduce methods based on a combination of these metrics. How this could be done is explained in the following section.

3. Methods: approaches for fully automated diagnosis

3.1. Selection based on single quality metrics (SEL-SIN)

First we investigate the effectiveness of single quality measures as introduced in Section 2. Based on a set of automatically extracted patches in an image, the patch with the maximum concerning the respective quality measure is selected.

3.2. Selection based on combined quality metric (SEL-COM)

Let Q be a matrix containing the row vectors $(q_A, q_C, q_B, q_N, q_I)$ of each patch of one original image and let W be a properly chosen column vector containing a weight for each quality measure. Then the column vector $Q \cdot W$ is the weighted summed overall quality measure. Our first approach is based on maximizing this weighted measure [13]. Therefore, the row with the maximum value of this product is evaluated and the corresponding image is used for feature extraction and classification. Classification in this context refers to the discrimination between images showing healthy and diseased mucosa (i.e. between class Marsh-0 and Marsh-3).

3.3. Information fusion: general remarks

Whereas the first two methods (SEL-SINGLE and SEL-SUMMED) rely on the selection of one single best patch per image, now we focus on a classification based on the fusion of the available patch data. In the following we use the vector $Q \cdot W$ which maps a real number to each patch. By computing the element-wise exponentiation of $Q \cdot W$ with the properly chosen exponent k, the ratio between the impact of high and low quality patches can be adjusted furthermore ($^\circ$ denotes the element-wise matrix exponentiation which corresponds to the repeated Hadamard matrix product). Using element-wise exponentiation has some positive aspects: First the ratio between the impact of high and low quality patches can be adjusted (theoretically) continuously. Additionally, by selecting specific values for k, a fallback to less sophisticated methods can be performed efficiently. In case of setting k to zero, the quality measures and the weights are ignored and each image finally has the same impact. The thereby achieved fusion methods (unweighted decision-level (DLF) and unweighted feature-level fusion (FLF)) are compared with the weight based methods in the experimental section. If assigning a large value to k ($k \to \infty$), the methods converge to the patch selection strategy as small values are thereby suppressed.

In the following two subsections we show how the quality vector $(Q \cdot W)^{\circ k}$ can be used in patch fusion. For the experiments, W and k are evaluated during exhaustive search based on a separate data set.

3.3.1. Weighted decision-level fusion (W-DLF)

The first method based on the computed quality vector $(Q \cdot W)^{\circ k}$ operates on the decision level. That means, for each patch in an original image, first the classifier's decision is computed by means of traditional feature extraction and classification. All decisions for one original image are stored in the row vector D, where 1 stands for a positive and -1 stands for a negative decision. By computing

$$D_f = \mathrm{sgn}(D \cdot (Q \cdot W)^{\circ k}) , \tag{6}$$

the single decisions are multiplied with the corresponding weights (image qualities), summed up and finally thresholded using the sign function sgn. We have to content with the rather simple sum rule, as more elaborate decision-level fusion approaches like the behavior-knowledge space [21] or decision templates [22] are developed for fusing different classifiers and not different input data.

3.3.2. Weighted feature-level fusion (W-FLF)

In opposite to W-DLF, W-FLF operates on the feature level. This implies that the features are fused prior to the classification step. In this approach the classification step that corresponds to a loss of information is postponed and applied to the fused features, which could be a benefit compared to the simpler decision-level fusion. The fused feature vector F_f which is used for classification is calculated by

$$F_f = F \cdot \frac{(Q \cdot W)^{\circ k}}{\|(Q \cdot W)^{\circ k}\|} , \tag{7}$$

where F is a matrix containing the feature vectors (columns) for each patch. The quality vector $(Q \cdot W)^{\circ k}$ is normalized to ensure that the sum of all contributions is one. The column vector F_f contains the element-wise weighted sum of all feature vectors and can be directly given to the classifier. F_f could be interpreted as a weighted average feature vector. We pursue this strategy, as it intuitively allows a weighting of the individual features, which cannot be achieved easily in case of a feature concatenation. The averaging theoretically requires that the decision boundaries are linear as otherwise the averaging of two features of one class could lead to an averaged descriptor located in the subspace of the other class. However, in the experiments we do not restrict to linear classification. To investigate the impact of the decision boundary on our approach, the utilized features are individually analyzed with respect to this problem in Section 4 with variable classifier adjustments.

In this work we focus on decision-level as well as feature-level fusion approaches but we do not investigate score-level approaches. This is done because decision- and feature-level methods can be generally applied whereas score-level techniques highly depend on the classifier that is utilized.

3.4. Computational runtime analysis

The major steps, as far as computational effort is concerned, consist of

- quality measurement (consisting of five single measures) and
- feature extraction.

The classification step is not considered as it is known to be quite fast (as the model can be computed in advance). Whereas in the fused approach the quality measures as well as the features must be computed for each patch, in case of patch selection [13] the feature must be computed only for the best patch. The overall computation time[1] for all quality measures on 128 × 128 pixel gray value patches is 37 milliseconds (ms) (q_A: 1 ms, q_C: 16 ms, q_B: 1 ms, q_N: 1 ms, q_I: 18 ms). The computation time for the features ranges from 6 to 142 ms (6 ms (LBP), 6 ms (ELBP), 13 ms (SCH), 142 ms (MFS), 2 ms (FPS)). For example in case of fusion based classification with LBP or ELBP and extracting 16 patches per original image, for each original image the computation time would be about 688 ms where 592 of them are consumed for quality measurement and only 96 are used for feature extraction. In case of patch selection based classification, it would take 598 ($592 + 6$) ms which is not significantly faster. Thus, we claim that the small additional computational effort is justified if the fusion leads to increased accuracies.

---

[1] Runtime tests are executed on an Intel i5 architecture with 3.1 MHz. All functions are implemented in MATLAB 2013a.

4. Experiments

4.1. Materials/patients: 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).

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 [2]. 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-class case is more relevant in practice [10]. Another incentive for preferring the 2-class case is that the distinction between the different stages of the disease is considerably subjective even as far as the histological examination is concerned [23]. Thereby, the ground-truth and furthermore the evaluation in a multi-class case would be less reliable.

Our experiments are based on a balanced database containing 612 patches (i.e. patches 306 per class) which are used for classifier training and 172 original images that are used for evaluation. From each original image, 16 non-overlapping 128 × 128 pixel patches are automatically extracted and furthermore used for fused classification. The patch size is chosen in order to be able to compare the results with the manual extraction that is done by a highly experienced endoscopist. The original images are captured during endoscopies from 72 different patients. To allow an efficient parameter estimation, this database (consisting of 612 patches and 172 original images) is divided into two equally sized sets (DB 1 and DB 2). In case of multiple images of one patient, we had to ensure that they are assigned to the same set. The weight vector W as well as the exponent k are evaluated during exhaustive search, based on the opposing data set as follows: In order to evaluate the accuracies based on the original images of DB 1, the patches of DB 2 are utilized for training and the original images of DB 2 are used for parameter estimation. The same procedure is applied (vice versa) to evaluate DB 2. Thereby a strict separation between training set, test set and evaluation set is achieved. The search space for each element of W is between 0.0 and 1.0 with a step-size of 0.33 and k is within $\{2^{-1}, 2^0, 2^1, \ldots, 2^6, 2^7\}$.

We perform two different experiments. Experiment A corresponds to the natural fusion of patches extracted from one distinct original image. Experiment B should show if the accuracy improvements are limited by the correlations within one original image. Such correlations are quite natural, as degradations like blur or noise often do not occur only in a small region, but sometimes even compromise a whole image. Therefore, in this experiment the patches of each patient are randomly interchanged across the images leading to virtual images consisting of patches from the same patient, but from different original endoscopic images. This is done as the patches from the new virtual images are supposed to be less correlated and the used database does not contain enough patients to fuse all patches from one patient.

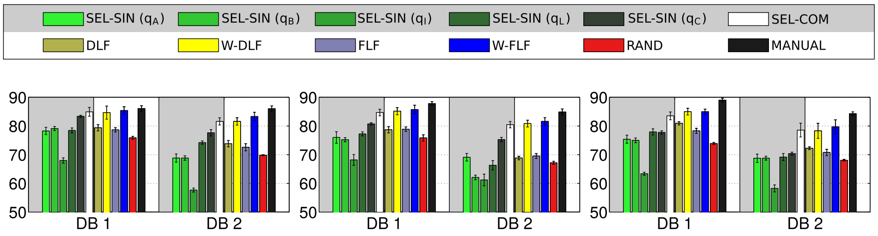
Each of the experiments (A and B) is performed twice. Once (Experiments A.1 and B.1) the classification model is trained based on the patches. This data is manually extracted from the original endoscopic images. Additionally, we make experiments (A.2 and B.2) based on a model trained with automatically extracted data. In this case, the training data is similarly generated as the data for evaluation. This is done because it is not clear if it is advantageous to train with more or less degradation-free (ideal) data (A.1, B.1) or with image data that is similarly generated as the evaluation data (A.2, B.2). These two different experiments are performed, because previous work on adaptive classification [24] suggests that similarity could be more important than a high degree of image quality.

For classification, the k-nearest neighbor classifier is used. We utilize this rather simple classifier in order to focus on the effect of different settings rather than on achieving the highest overall classification rates. For the first experiments (A and B), the rates achieved with odd k values reaching from 1 to 31 are averaged, to get highly stable results (as the classification rates of 16 runs are averaged) rather than to get the highest possible rates. In a further experiment (see Fig. 6) we investigate the impact of the flexibility of the classifier by varying the classifier's k value. This effective adjustment with significant impact on the flexibility (non-linearity) of the decision boundaries is another incentive to deploy this well-known and easily understandable classification model for experimentation.

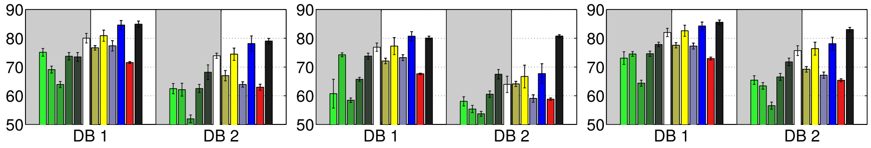
4.2. Image descriptors

For the experimental analysis we deploy the following feature extraction techniques which proved to the adequate for celiac disease classification in previous work [11]:

- Local Binary Patterns [25] (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 [26] (ELBP):
  ELBP is an edge based derivative of Local Binary Patterns. As LBP it is used with eight neighbors and a radius of two pixels.
- Fourier Power Spectra Rings [27] (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 [27].



(a) LBP (b) ELBP (c) SCH



(d) MFS (e) FPS (f) MEAN

Fig. 2. Experiment A.1: These plots show the overall classification rates achieved with patch selection (SEL-SIN and SEL-COM), decision-level fusion (DLF and W-DLF), feature-level fusion (FLF and W-FLF), a random patch selection (RAND) and the manual patch selection (MAN). Training is based on the (ideal) patch image data.

- Shape Curvature Histogram [28] (SCH):
  SCH is a shape feature, especially developed for celiac disease diagnosis. After detection of significant locations, a histogram collects the occurrences of the contour curvature values in these regions. As in the original work, we consider a histogram bin count of eight.
- Multi-Fractal Spectrum [29] (MFS):
  The local fractal dimension is computed for each pixel using three different types of measures for computing the local density. The feature vector is built by concatenation of these fractal dimensions.

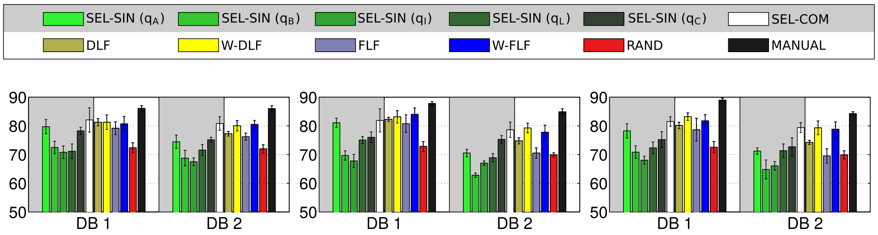
4.3. Results and discussion

Fig. 2 shows the overall classification rates achieved with patch selection based on single measures (SEL-SIN), patch selection based on combined measures (SEL-COM), unweighted and weighted decision-level fusion (DLF and W-DLF), unweighted and weighted feature-level fusion (FLF and W-FLF), random patch selection (RAND) and patch selection based on the human experts (MAN) for experiment A. We notice that based on single quality metrics it is not possible to select appropriate image data for a subsequent classification. Among those methods, selection based on q_A and q_C turned out to be most effective. These two methods are at least able to continuously outperform a random patch selection. Considering the approach based on the combination of quality measures SEL-COM, the obtained accuracies are already relatively good and stable. Obviously the combination definitely is necessary to get a meaningful overall quality metric to optimize subsequent classification. Interestingly, it can be seen that the unweighted feature-level fusion method FLF as well as the unweighted decision-level method DLF are unable to compete with the SEL-SIN approach in case of any feature. The rates obtained with the manual selection are totally out of reach. Considering the weight based methods we recognize that especially the weighted feature-level based method W-FLF is able to outperform the patch selection method SEL-COM in case of all features and all databases with differing extent. Considering MFS, LBP and ELBP, the accuracies of the manual patch selection can be virtually reached. Quite high differences are observed in case of SCH and FPS. A quite interesting aspect is the difference between the two weighted fusion techniques. In almost each case, the feature based W-FLF corresponds to the higher accuracy compared to W-DLF. Obviously the early fusion prior to the (information reduction) classification has a positive impact on the final discriminative power.

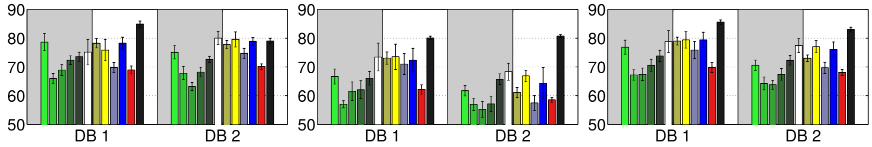
Fig. 3 shows the obtained accuracies based on training with the automatically extracted data. We notice that the performance in this scenario generally is lower. Especially the methods based on information fusion (W-FLF as well as W-DLF) seem to be less appropriate if training is performed with non-ideal data. Due to the generally lower accuracies, it should be noted that the generation of ideal training data in a manual sense definitely can be advantageous for a classification system.

In Fig. 4 the results of experiment B, which is based on randomly interchanged patches across images of the same patient, are shown. As in this experiment not only patches extracted from the same image are fused, but patches from different images, we have expected that in this scenario more significant improvement could be obtained in case of information fusion. Actually, on average (see Fig. 2(f)) the rates with the weight based fusion methods are quite similar.

Considering the results of experiment B with non-ideal training data (see Fig. 5), it can be seen that a similar performance compared to experiment B.1 is obtained. Interestingly, we again observe that especially the feature-level fusion method (W-FLF) generates less competitive results with the non-ideal training data. In opposite, the best outcome on average is obtained

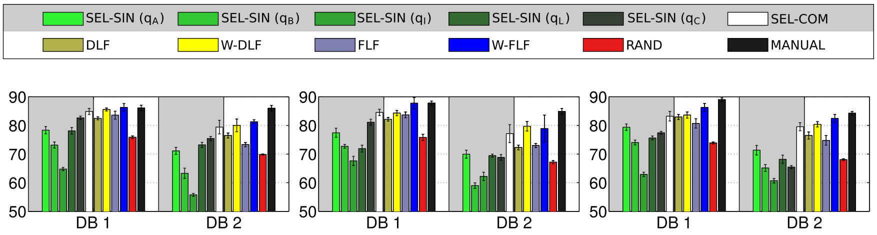


(a) LBP (b) ELBP (c) SCH

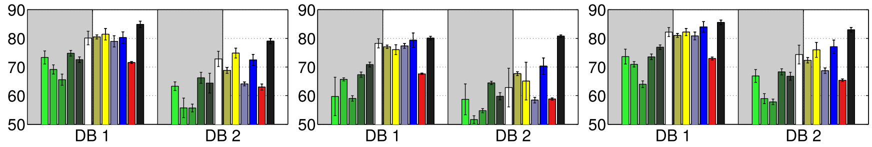


(d) MFS (e) FPS (f) MEAN

Fig. 3. Experiment A.2: These plots show the overall classification rates achieved with patch selection (SEL-SIN and SEL-COM), decision-level fusion (DLF and W-DLF), feature-level fusion (FLF and W-FLF), a random patch selection (RAND) and the manual patch selection (MAN). Training is based on automatically extracted data.



(a) LBP (b) ELBP (c) SCH



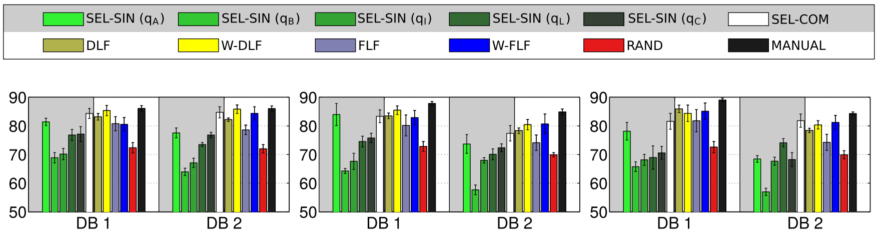
(d) MFS (e) FPS (f) MEAN

Fig. 4. Experiment B.1: These plots show the accuracies with the same strategies as in Fig. 2. In opposite to experiment A, in experiment B the patches of one patient are randomly interchanged.

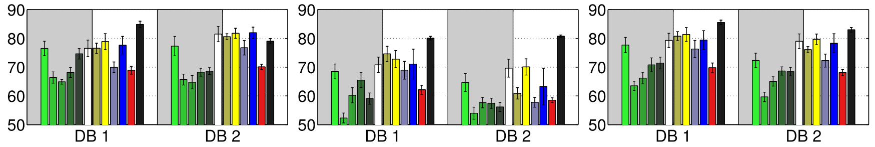
with the decision level based approach (W-DLF). Obviously the specific settings of this medical decision support system has a major impact on the effectiveness of the different selection and fusion approaches.

As especially the weight based feature-level fusion method in most experiments leads to high accuracies, we expect that a fusion on the one hand across all patches in an image (derived from experiment A.1) and on the other hand across all images, captured during endoscopy of one distinct patient (derived from experiment B.1) could improve the rates from our experiments once again. Unfortunately, the data currently available is not large enough for such an experimental evaluation.

So far, we experimentally showed that the W-FLF approach is mostly the best method to make our decision support system totally automated. However, we have not regarding the theoretical issues of this method in case of non-linear decision boundaries. Finally we investigate the impact of the classifier's decision boundary on the effectiveness of W-FLF. As stated in Section 3.3.2, the feature averaging theoretically requires that the decision boundaries are linear as otherwise the averaging of two features of one class could lead to an averaged descriptor located in the subspace of the other class. To investigate how often an averaged feature of two correctly classified images would be incorrectly classified, now we consider all correctly classified images (from the ideal patches data set). For each pair of these images, the average feature is computed and classified with varying settings (different k values). This is done as especially small k values correspond to highly non-linear decision

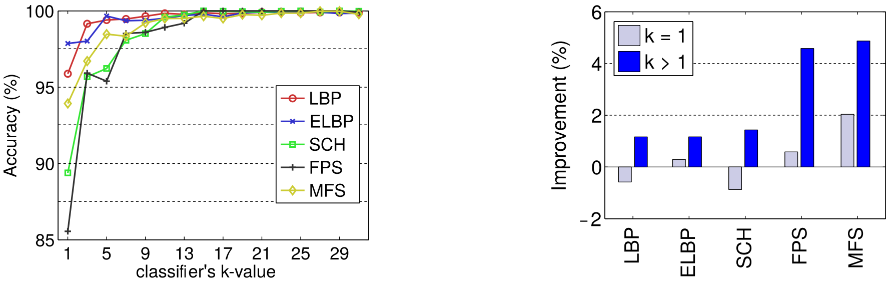


(a) LBP (b) ELBP (c) SCH



(d) MFS (e) FPS (f) MEAN

Fig. 5. Experiment B.2: These plots show the accuracies with the same strategies as in Fig. 3. In opposite to experiment A, in experiment B the patches of one patient are randomly interchanged.



(a) Analysis of the decision boundaries. (b) Impact of varying k-values on the accuracies.

Fig. 6. Analysis of effects in weighted feature-level fusion W-FLF.

boundaries, whereas with higher k values this effect is softened. Fig. 6(a) shows that especially in combination with low dimensional features (FPS, SCH, MFS) and small k values (majorly for $k = 1$), the feature averaging leads to decreased classification accuracies (as 100% accuracy is expected in case of linear classification).

In Fig. 6(b), the impact of a small k value ($k = 1$) compared to the averaging (with k reaching from 3 to 31) on the improvement achieved with W-FWF compared to patch selection SEL-SIN is shown. Apart from the classifier settings, the same setup as in experiment A.1 is used. As expected, if k is set to one, the improvements of W-FLF are (especially in combination with the low dimensional features) considerably smaller or even negative, which is expected to be due to the highly non-linear decision boundaries. Therefore, we finally recommend to take care about the classifier choice using the proposed method for weighted feature-level fusion.

5. Conclusion

We have investigated several techniques to fully automatize decision support systems for celiac disease diagnosis. It has been shown that a patch selection system based on single quality measures does not lead to adequate performances. Nevertheless, the combination of several metrics turned out to be already quite effective. The exploitation of data redundancy by means of information fusion furthermore leads to distinct improvements. The best performances are finally obtained with an unconventional method based on feature-level fusion. It has been shown that the measurement of image quality has a major impact not only in case of a single patch selection, but also in case of information fusion. Experiments based on specific scenarios show that the choice of the best method distinctly depends on the training data. Getting quite close to the classification rates of manual patch selection, this work brings us one step closer to fully automated non-interactive celiac disease diagnosis.

Acknowledgement

This work is partially funded by the Austrian Science Fund (FWF) under projects No. 24366 and No. KLI 429.

References

[1] Marsh M. Gluten, major histocompatibility complex, and the small intestine. A molecular and immunobiologic approach to the spectrum of gluten sensitivity (‘celiac sprue’). Gastroenterology 1992;102(1):330–54.

[2] Oberhuber G, Granditsch G, Vogelsang H. The histopathology of coeliac disease: time for a standardized report scheme for pathologists. Eur J Gastroenterol Hepatol 1999;11:1185–94.

[3] Gschwandtner M, Liedlgruber M, Uhl A, Vécsei A. Experimental study on the impact of endoscope distortion correction on computer-assisted celiac disease diagnosis. In: Proceedings of the 10th international conference on information technology and applications in biomedicine (ITAB'10). Nov. 2010.

[4] Gschwandtner M, Hämmerle-Uhl J, Höller Y, Liedlgruber M, Uhl A, Vécsei A. Improved endoscope distortion correction does not necessarily enhance mucosa-classification based medical decision support systems. In: Proceedings of the IEEE international workshop on multimedia signal processing (MMSP'12). Sept. 2012. p. 158–63.

[5] Liedlgruber M, Uhl A, Vécsei A. 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). July 2011.

[6] Fasano A, Berti I, Gerarduzzi T, Not T, Colletti RB, Drago S, et al. Prevalence of celiac disease in at-risk and not-at-risk groups in the United States: a large multicenter study. Arch Intern Med 2003;163:286–92.

[7] Ciaccio EJ, Tennyson CA, Bhagat G, Lewis SK, Green P. Classification of videocapsule endoscopy image patterns: comparative analysis between patients with celiac disease and normal individuals. Biomed Eng Online 2010;9(1):1–12.

[8] Ciaccio EJ, Tennyson CA, Lewis SK, Krishnareddy S, Bhagat G, Green P. Distinguishing patients with celiac disease by quantitative analysis of videocapsule endoscopy images. Comput Methods Programs Biomed 2010;100(1):39–48.

[9] Hegenbart S, Uhl A, Vécsei A. Impact of histogram subset selection on classification using multiscale LBP. In: Proceedings of Bildverarbeitung für die Medizin 2011 (BVM'11). Informatik aktuell. 2011. p. 359–63.

[10] Vécsei A, Amann G, Hegenbart S, Liedlgruber M, Uhl A. Automated marsh-like classification of celiac disease in children using an optimized local texture operator. Comput Biol Med 2011;41(6):313–25.

[11] Gadermayr M, Liedlgruber M, Uhl A, Vécsei A. Evaluation of different distortion correction methods and interpolation techniques for an automated classification of celiac disease. Comput Methods Programs Biomed 2013;112(3):694–712.

[12] Hegenbart S, Uhl A, 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). Sept. 2011. p. 715–20.

[13] Gadermayr M, Uhl A, Vécsei A. Getting one step closer to fully automatized celiac disease diagnosis. In: Proceedings of the 4th IEEE international conference on image processing theory, tools and applications 2014 (IPTA'14). Oct. 2014. p. 13–7.

[14] Bashar MK, Kitasaka T, Suenaga Y, Mekada Y, Mori K. Automatic detection of informative frames from wireless capsule endoscopy images. Med Image Anal 2010;14(3):449–70.

[15] Atasoy S, Mateus D, Lallemand J, Meining A, Yang GZ, Navab N. Endoscopic video manifolds. In: Proceedings of the international conference on medical image computing and computer assisted intervention (MICCAI'10). Lecture notes in computer science, vol. 6362. 2010. p. 437–45.

[16] Gadermayr M, Uhl A, Vécsei A. Quality based information fusion in fully automatized celiac disease diagnosis. In: Proceedings of the German conference on pattern recognition (GCPR'14). Lecture notes in computer science, vol. 8753. 2014. p. 1–12.

[17] Ross A, Jain A. Information fusion in biometrics. Pattern Recognit Lett 2003;24(13):2115–25.

[18] Prabhakar S, Jain AK. Decision-level fusion in fingerprint verification. Pattern Recognit 2002;35(4):861–74.

[19] Uhl A, Wild P. Single-sensor multi-instance fingerprint and eigenfinger recognition using (weighted) score combination methods. Int J Biom 2009;1(4):442–62 [special issue on multimodal biometric and biometric fusion].

[20] Marziliano P, Dufaux F, Winkler S, Ebrahimi T, Sa G. A no-reference perceptual blur metric. In: Proceedings of the IEEE international conference on image processing (ICIP'02). 2002. p. 57–60.

[21] Raudys Š, Roli F. The behavior knowledge space fusion method: analysis of generalization error and strategies for performance improvement. In: Proceedings of the 4th international conference on multiple classifier systems, MCS'03. Springer-Verlag; 2003. p. 55–64.

[22] Kuncheva LI, Bezdek JC, Duin RPW. Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognit 2001;34(2):299–314.

[23] Weile B, Fischer Hansen B, Hägerstrand I, Hansen JPH, Krasilnikoff PA. Interobserver variation in diagnosing coeliac disease, a joint study by Danish and Swedish pathologists. APMIS, Acta Pathol Microbiol Immunol Scand 2000;108(5):380–4.

[24] Gadermayr M, Uhl A, Vécsei A. Degradation adaptive texture classification: a case study in celiac disease diagnosis brings new insight. In: Proceedings of the international conference on image analysis and recognition (ICIAR'14). Lecture notes in computer science, vol. 8815. 2014. p. 263–73.

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

[26] Liao S, Zhu X, Lei Z, Zhang L, Li S. Learning multi-scale block local binary patterns for face recognition. In: Advances in biometrics. Springer; 2007. p. 828–37.

[27] Gadermayr M, Uhl A, Vécsei A. Barrel-type distortion compensated Fourier feature extraction. In: Proceedings of the 9th international symposium on visual computing (ISVC'13). Lecture notes in computer science, vol. 8033. 2013. p. 50–9.

[28] Gadermayr M, Liedlgruber M, Uhl A, Vécsei A. 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). Lecture notes in computer science, vol. 8331. 2014. p. 175–84.

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