Degradation Adaptive Texture Classification: A Case Study in Celiac Disease Diagnosis Brings new Insight

Michael Gadermayr[1], Andreas Uhl[1], and Andreas Vécsei[2]

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

Abstract. Degradation adaptive texture classification has been claimed to be a powerful instrument for classifying images suffering from degradations of dissimilar extent. The main goal of this framework is to separate the image databases into smaller sets, each showing a high degree of similarity with reference to degradations. Up to now, only scenarios with different types of synthetic degradations have been investigated. In this work we generalize the adaptive classification framework and introduce new degradation measures to extensively analyze the effects of the approach on real world data for the first time. Especially computer aided celiac disease diagnosis based on endoscopic images, which has become a major field of research, is investigated. Due to the weak illuminations and the downsized sensors, the images often suffer from various distortions and the type as well as the strength of these degradations significantly varies over the image data. In a large experimental setup, we show that the average classification accuracies can be improved significantly.

Keywords: Adaptive classification, endoscopic images, celiac disease

1 Introduction

The degradation adaptive classification framework [1] has been investigated with reference to "idealistic" degradations. For this purpose, based on the Kylberg texture database [2], the three types of degradations isotropic Gaussian blur, Gaussian white noise and isotropic scale variations have been simulated separately. The authors showed that the approach is able to improve the classification rates (overall classification accuracies), with most configurations. While it is highly interesting that the method works in case of the respective simulated scenarios, the impact in case of real world image degradations is not clear yet.

In this work, a real world classification scenario is investigated with endoscopic images, suffering from various distortions. The final task is to discriminate between images of a healthy person and a person suffering from celiac disease [3–6]. The database shows the following divergences from the previous scenario [1] based on synthetic distortions:

- The real world image degradations differ from the simulated ones. For example, in the previous work [1] only isotropic scale variations are simulated,

but in endoscopy, due to varying viewing angles, anisotropic scale variations, perspective distortions and even non-linear deformations [7] are omnipresent.
- The distortions on average are less significant and furthermore the distribution of the distortion strengths differs (in the preceding paper [1], a uniform distribution is generated).
- In this database, not just one single distortion, but even combinations of distortions are prevalent. In the previous paper [1], three scenarios (one for each degradation type) are considered separately.
- The available training set is significantly smaller (about 300 instead of 10,000 images) which potentially induces very small training subsets in case of adaptive classification.

In previous work [8], the impact of various degradations such as blur and noise on computer aided celiac disease diagnosis has been investigated. The authors showed that especially noise and blur have a major impact on the classification accuracy. Furthermore, the impact of lens distortions as well as distortion correction has been prospected in a large experimental setting [7]. In another study [9], the authors investigate the impact of varying scales with respect to computer aided celiac disease diagnosis.

A concept related to degradation adaptive classification is given by domain adaption [10]. Although the nomenclature suggests a high similarity, in opposite to the approach investigated in our paper, this method has been developed to allow domain shifts between the training and the evaluation dataset. Consequently, domain adaption would be applicable if the type or the extent of the prevalent degradations significantly differs between the training and the evaluation set, which is not the case in our scenario. Due to collisions in nomenclature, we have to mention that the term adaptive classification in the following always refers to degradation adaptive classification.

In this paper, we investigate the impact of adaptive classification on real world image data, which has not been done before. Therefore, seven different similarity (or degradation) measures are deployed. Three of them have already been investigated [1] and have been deployed to capture the three simulated degradations. Four more measures which do not directly address specific degradations are introduced and investigated with respect to the final classification rates. These new measures are implemented as degradation adaptive classification turned out to be advantageous even if no (simulated) degradations are prevalent [1]. Consequently, we suspect that metrics which do not directly measure degradations, but other image properties, could be powerful as well. Furthermore, we extend the adaptive classification framework, to additionally investigate multi-dimensional similarity measures. This potentially is useful as we have to cope with multiple degradations in one database. Finally, the results are extensively analyzed in order to understand, how the adaptive classification works.

This paper is structured as follows: In Sect. 2, the distortion adaptive classification framework is explained and extended to a multidimensional adaptive classification framework and the degradation measures are described. In Sect. 3, the experimental results are presented and discussed. Section 4 finally concludes this paper.

2 Degradation Adaptive Classification

In recent work [1], we showed that an absolutely robust feature is harder to achieve than a relatively robust one. Relative robustness in this context means that the classification accuracy does not decrease strongly if all images in a database (i.e. all images in the training and the evaluation set) similarly suffer from a degradation. In opposite, absolute robustness (or invariance) means that the accuracy can even be maintained if the images in the training and the evaluation set suffer from the same degradation, but with a dissimilar extent. Degradation adaptive classification [1] exploits this knowledge by dividing a dataset into several smaller datasets with similar properties with respect to degradation type and severity. Thereby absolute robustness becomes less decisive.

Based on a normalized degradation measure $D : \Omega \to [0, 1)$ (Ω is the image domain) the original training set $T \subset \Omega$, is divided into the subsets

$$T_i = \{I \in T : d \leq D(I) \cdot C - i < d + 1\}. \tag{1}$$

where $i \in \{0, 1, ..., C - 1\}$, C denotes the cardinality of the set of generated subsets and d defines the overlap.

In a similar manner, the evaluation set $E \subset \Omega$ is partitioned into the subsets

$$E_i = \{I \in E : 0 \leq D'(I) \cdot C - i < 1\}, \tag{2}$$

where $D'(I) = \max(\min(D(I), 1 - \epsilon), 0)$ and ϵ is a small constant, to ensure that each sample belongs to exactly one evaluation subset. Finally for each i, the evaluation set E_i is classified by the discriminant generated by T_i.

This methodology could also be interpreted in terms of a classifier selection system [11]. Classifier selection is done by means of a degradation measure, based on all images in the training set. The decision of this selection defines one specific classifier (which is based on a specific training set) to compute the final decision.

A subdivision using soft-assignment clustering (e.g. kmeans) instead of the linear intervals which has been also considered, did not lead to improved results.

2.1 Multidimensional Adaptive Classification

The degradation adaptive classification framework allows the use of one-dimensional degradation measures ($D : \Omega \to [0, 1)$). In order allow the usage of measures of an arbitrary dimensionality n ($D : \Omega \to [0, 1)^n$), the definition has to be slightly adapted. The training set has to be divided into the subsets

$$T_{i_1,...,i_n} = \{I \in T : \bigwedge_{j=1}^{n} d_j \leq \pi_j(D(I)) \cdot C_j - i_j < d_j + 1\}. \tag{3}$$

where $i_j \in \{0, 1, ..., C_j - 1\}$, C_j denotes the cardinality of the set of generated subsets for each dimension and d_j defines the overlap for each dimension separately. The projection π_j selects the j^{th} element of an n-tupel. In the experiments, for each j, C_j is set to the same value (C) and the same is done for d_j, in order to limit the search space.

In a similar manner, the evaluation set E is partitioned into the subsets

$$E_{i_1,...,i_n} = \{I \in E : \bigwedge_{j=1}^{n} 0 \leq \pi_j(D'(I)) \cdot C_j - i_j < 1\}. \quad (4)$$

Finally for each n-tupel $(i_1, ..., i_n)$, the evaluation set $E_{i_1,...,i_n}$ is classified by the discriminant generated by $T_{i_1,...,i_n}$.

2.2 Similarity Measures

In order to divide a dataset into several smaller ones with higher similarities (and/or similar degradations), a normalized metric D to capture this similarity or degradation is required. In this work we focus on the following metrics which are furthermore min-max normalized to the interval $[0, 1)$:

- Noise Metric (D_n): Noise being prevalent in images can be measured by computing the total pixelwise sum of the absolute difference between an image and the Gaussian filtered ($\sigma = 1$) version of this image.
- Blur Metric (D_b): In oder to measure blur, the metric which has been introduced in [12] is deployed. After identifying the edges in the horizontal direction by extracting all local minima and maxima for each row, the ratio between the overall lengths and the magnitudes of these edges are computed indicating the blur level.
- Scale Metric (D_s): For estimation the texture scale, the scale-space method introduced in [13] is utilized. The global scale of an image is estimated by first constructing a scale space by convolving an image with Laplacian-of-Gaussians (LoG) in varying scales. As proposed in [13], for the LoGs, the scales $\sigma = \hat{c}\sqrt{2}^k, k \in \{-4, -3.75, ..., 7.75, 8\}$ are chosen with $\hat{c} = 2.1214$. The scales for each pixel are obtained by using the index of the highest filter responses. The final global scale of an image is estimated by computing the histogram of this scale values over all pixels, followed by a Gauss-fitting.
- Mean Metric (D_m): This measure captures the average overall gray value. The intention is to separate properly illuminated images from weakly illuminated ones which usually show more noise in combination with a lack of contrast.
- Contrast Metric (D_{c_*}): Contrast is not only able to measure degradations, but it is already able to discriminate between healthy and diseased patients. Although it seems to work in a different way, we would like to investigate the effect of such a discriminating metric on the adaptive classification framework. For that, the three contrast neighborhood sizes of two (D_{c_S}), four (D_{c_M}) and six pixels (D_{c_L}) are utilized.
- Two-dimensional Metric Combinations: Finally we combine the most appropriate measure D_{c_L} (i.e. the measure corresponding to the highest accuracies on average in Sect. 3), with the second ($D_{c_M:c_L}$), the third ($D_{c_S:c_L}$) and the fourth placed measure ($D_{b:c_L}$) utilizing the multidimensional adaptive classification framework.

3 Experiments

3.1 Experimental Setup

The image testsets utilized for the experiments contain images of the duodenal bulb and the pars descendens taken during duodenoscopies at the St. Anna Children's Hospital using pediatric gastroscopes (Olympus GIF N180 and Q165). In a preprocessing step, texture patches with a fixed size of 128×128 pixels have been manually extracted. The size turned out to be suitable in earlier experiments [5]. Prior to feature extraction, all patches are converted to gray scale images. In Fig. 1 example texture patches are shown.

To get the ground truth for the texture patches, the condition of the mucosal areas covered by the images has been determined by histological examination of biopsies from corresponding regions. The severity of the villous atrophy has been classified according to the modified Marsh classification [14]. Although it is possible to distinguish between different stages of the disease, we aim in distinguishing between images of patients with (Marsh-3) and without the disease (Marsh-0), as this two classes case is most relevant in practice. Our experiments are based on a training and an evaluation dataset containing 300 (151 Marsh-0 and 149 Marsh-3) and 312 (155 Marsh-0 and 157 Marsh-3) images, respectively.

After defining nine sensible subset counts ($C \in \{1, 5, 9, 13, 17, 21, 25, 29, 33\}$) as well as six sensible overlaps ($d \in \{0, 2, 4, 6, 8, 10\} \cdot 10^{-2}$), training and evaluation sets for all combinations (C, d) are generated, the classifiers are trained and the decisions for each evaluation set are processed. Finally, the most appropriate configuration is predicted in a leave-one-patient-out cross validation.The experiment is repeated with reversed training and evaluation set and the mean rates, featuring a higher reliability, are finally utilized.

To extensively study the effect of the proposed approach on the overall classification accuracy, eight different feature extraction techniques which turned out to be appropriate for celiac disease classification are investigated. The chosen parameters turned out to be optimally suited in earlier experiments.

- Graylevel Contrast (CON) [15]: This feature vector consists of the Haralick feature contrast [15] computed from the gray level co-occurrence matrices with different offset vectors $(0, c)^T$, $(c, 0)^T$ and $(c, c)^T$. For the experiments c is fixed to four pixels.
- Edge Co-occurrence Matrix (ECM) [16]: After applying eight differently orientated directional filters, the orientation is determined individually for each

Fig. 1: Example image patches of healthy patients (left) clearly showing the villi structure and patches of diseased patients (right) suffering from villous atrophy. It can be seen that the discrimination in some cases is quite difficult.

pixel, followed by masking out pixels with a low gradient magnitude (75% below the maximum response). Then the methodology of the gray-level co-occurrence matrix is applied to obtain the ECM for one specific displacement (two pixels in our experiments).

- Local Binary Patterns (LBP) [17]: LBP describes a texture by computing the joint distribution of binarized pixel intensity differences represented by binary patterns. We deploy the uniform LBP version, capturing only patterns with at most two bitwise transitions, in combination with eight circular samples and a radius of two pixels.
- Extended Local Binary Patterns (ELBP) [18]: 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 Spectrum Features (FF) [19]: To get this descriptor, first the Fourier power spectra of the patches are computed in a way that the low frequencies are located in the center. After that, rings with a fixed inner and outer radius are extracted and the medians of these values are calculated. For our experiments we use one single ring with an inner radius of seven and an outer radius of eight pixels.
- Shape Curvature Histogram (SCH) [20]: SCH is a shape feature especially developed for computer aided celiac disease diagnosis. A histogram collects the occurrences of the contour curvature magnitudes. The proposed histogram bin count of eight is used in our experiments.
- Histogram of Gradients (HOG) [21]: Similar to SCH, this well known feature computes the distribution of gradient orientations to describe a texture. The standard bin count of nine is used, which corresponds to an angular resolution of 20 degrees.
- Random Feature (RDF): Finally we investigate a random feature, which returns a random value between zero and one, independently from the input signal. Although this feature is not useful in practical scenarios, it helps us to understand the behavior of adaptive classification.

As it could have been shown [1] that the classifier has a significant impact on the effectiveness of the adaptive classification approach, we deploy three very different classifiers consisting of the (highly non-linear) nearest neighbor classifier (NNC), a linear (Bayes normal) classifier (LDC) [22] as well as a support vector classifier (SVC) [22] which is based on a polynomial kernel. To evaluate if an improvement is significant, McNemar's test [23] ($\alpha = 0.01$) is deployed.

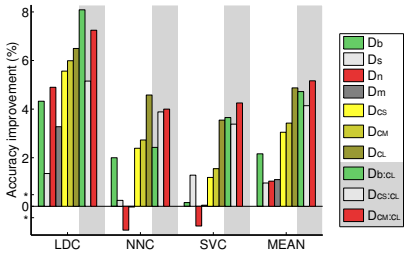
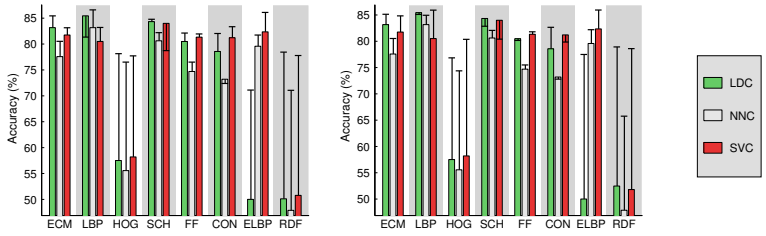


Fig. 2: Average classification accuracies achieved with the three classifiers. Improvements or aggravations exceeding the asterisk symbol (*) are significant.

3.2 Experimental Results

First of all, in order to identify the most appropriate similarity measure, in Fig. 2 the accuracy increases (and decreases) for all features are averaged separately for each mode and each classifier. Additionally the mean over the three classifiers is given which shows that the most appropriate measures on average are given by D_{c_L} (followed by D_{c_M}, D_{c_S} and D_b considering one dimensional measures) and $D_{c_M:c_L}$ (which is the overall best measure). The scale based D_s and the noise based D_n correspond to the lowest average improvements. Obviously scale and noise differences do not deeply affect the traditional classification, compared to the simulated scenario [1] which showed significant accuracy increases using these measures. For this experiment, the random RDF feature is not considered, as it is not relevant in practice. It can be observed that the two-dimensional features which are highlighted by the gray background, do not significantly outperform the one-dimensional D_{c_L}. This behavior is supposed to be due to the limited dataset sizes as especially a large C in case of two dimensions causes extremely small training sets.

Furthermore, it can be seen that the LDC classifier on average profits more significantly than the NNC and the SVC classifiers from the new framework. Such a behavior has already been reported and discussed in recent work [1]. In another study [13] which investigates the impact of classifying images of varying scales the nearest neighbor classifier is analyzed extensively. Considering for each subject in the evaluation set only the corresponding training subject with the smallest distance, the NNC classifier induces a highly non-linear decision boundary. The authors have shown, that this classifier is able to most likely ("implicitly") choose a similarly degraded corresponding subject as nearest neighbor. The achieved accuracy benefit with this classifier on average is supposed to be smaller as adaptive classification also aims in grouping similarly degraded data. In opposite to the NNC classifier, the LDC classifier has to cope with a linear decision boundary and is not able to simply "ignore" subjects with a larger distance. The SVC classifier is investigated as it is widely used in practice and furthermore represents an intermediate method. Finally we notice that,



(a) D_{c_L} measure (b) $D_{c_M:c_L}$ measure

Fig. 3: This figure shows the obtained performances in combination with the two similarity measures. The bars indicate the rates achieved with traditional and the horizontal lines indicate the accuracies obtained with adaptive classification.

although the LDC classifier corresponds to the highest average improvements, consistent improvements are also achieved with the classifers NNC and SVC in our experiments. These two methods seem to profit similarly from adaptive classification.

In the following we only consider the measures corresponding to the lowest error rates D_{c_L} (as best one-dimensional measure) and $D_{c_L:c_M}$ (as best overall measure). In Fig. 3, the accuracies are given separately for each feature and each classifier.

Using the D_{c_L} measure, with all features the highest overall accuracies are achieved utilizing the adaptive classification framework. Using the two-dimensional measure $D_{c_M:c_L}$, with all features but SCH the best accuracies are obtained. Altogether, 39 improvements face eight aggravations. The two-dimensional measure corresponds to the highest average improvement, however, the higher continuity (i.e. the number of improvements) is achieved with the one-dimensional measure.

3.3 Discussion

A quite interesting behavior is seen with HOG and the random feature RDF. Although in case of traditional classification with all classifiers the accuracies are very low with HOG and as expected around 50 % with RDF, with the adaptive classification framework, with both features, rates above 75 % can be achieved. Considering the increased accuracies of HOG with adaptive classification, the reader could think that the feature benefits because it is not absolutely robust but quite relatively robust to the distortions prevalent in the database. However, we observe a similar behavior with RDF which does not provide any discriminative power and therefore absolute and relative robustness of this feature do not play any role.

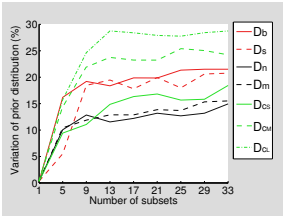


Fig. 4: Variations of prior probabilities

Inspired by these results, we identified another effect of the adaptive classification framework on the database. Whereas in the original dataset the classes are distributed approximately equally, by grouping the training and evaluation set images using adaptive classification this equal prior distribution is not necessarily maintained. To quantify this behavior, in a small experiment we compute an average variance v which measures the distance to a perfectly equal prior distribution over all sets. This is done by summing up the absolute differences between the prior probabilities P_{T_i} of one class (Marsh-0) and 0.5 weighted by the ratio of images in the respective subset, for each training set T_i:

$$v(C) = \sum_{i=0}^{C-1} |P_{T_i}(X = \text{Marsh-0}) - 0.5| \cdot \frac{\#T_i}{\#T} \ . \tag{5}$$

For the introduced one-dimensional similarity measures and varying numbers of subsets C (horizontal axis), these values are presented in Fig 4. We notice that

the prior distribution definitely varies in case of adaptive classification and as expected the variation raises with an increasing number of subsets. Moreover, we observe that the most effective measures (D_{c_L} and D_{c_M}) as far as the accuracy improvement is concerned correspond with the highest variations. This analysis shows that adaptive classification not only divides a database into smaller sets with similar properties, but furthermore changes the prior distributions considerably. This definitely has a positive impact in case of the RDF feature. However, this mechanism is supposed to potentially increase the classification performances in case of all features, especially in combination with the D_{c_*} similarity measures which do not directly capture any degradation but rater act as discriminative features.

4 Conclusion

We have investigated the effect of degradation adaptive classification on a real world classification problem. To additionally analyze two-dimensional degradation measures, the framework has been generalized. Finally the best improvements on average are achieved with the contrast based one-dimensional D_{c_L} and the two-dimensional $D_{c_M:c_L}$ measure. Except for one feature (SCH), the accuracies can be continuously (i.e. in case of both measures) improved. The average increase of the accuracy is 5.2 % ($D_{c_M:c_L}$) and 4.9 % (D_{c_L}), respectively. However, in the studied scenario the introduction of multidimensional measures did not lead to significant improvement compared to one-dimensional ones. In previous work, the authors found out that the framework improves accuracies especially in case of a large range of degradation strengths by separating the original dataset in similarly degraded smaller datasets. We have shown that adaptive classification furthermore improves the classification performance, especially in case of features with low discriminative powers, by changing the prior distributions within the datasets.

References

1. Gadermayr, M., Uhl, A.: Degradation adaptive texture classification. In: IEEE International Conference on Image Processing 2014 (ICIP'14). (October 2014) accepted.
2. Kylberg, G.: The kylberg texture dataset v. 1.0. Technical Report 35, Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, Sweden (2011)
3. Ciaccio, E.J., Tennyson, C.A., Lewis, S.K., Krishnareddy, S., Bhagat, G., Green, P.: Distinguishing patients with celiac disease by quantitative analysis of videocapsule endoscopy images. Comp. Methods and Prog. in Biomedicine **100**(1) (2010) 39–48
4. Ciaccio, E.J., Tennyson, C.A., Bhagat, G., Lewis, S.K., Green, P.H.R.: Classification of videocapsule endoscopy image patterns: comparative analysis between patients with celiac disease and normal individuals. BioMedical Engineering Online **9**(1) (2010) 1–12
5. 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. Computers in Biology and Medicine **41**(6) (2011) 313–325

6. Hegenbart, S., Uhl, A., Vécsei, A., Wimmer, G.: Scale invariant texture descriptors for classifying celiac disease. Medical Image Analysis **17**(4) (2013) 458 – 474
7. Gadermayr, M., Liedlgruber, M., Uhl, A., 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
8. Hegenbart, S., Uhl, A., V'ecsei, 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). (2011) 715–720
9. Hegenbart, S., Uhl, A., V'ecsei, 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) 1–6
10. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: Proceedings of the 11th European Conference on Computer Vision (ECCV'10). (2010) 213–226
11. Giacinto, G., Roli, F.: Methods for dynamic classifier selection. In: Proceedings of the Intern. Conf. on Image Analysis and Processing (ICIAP'99). (1999) 659–664
12. Marziliano, P., Dufaux, F., Winkler, S., Ebrahimi, T., Sa, G.: A no-reference perceptual blur metric. In: IEEE International Conference on Image Processing (ICIP'02). (2002) 57–60
13. Gadermayr, M., Hegenbart, S., Uhl, A.: Scale-adaptive texture classification. In: Proceedings of 22nd International Conference on Pattern Recognition (ICPR'14). (August 2014) accepted.
14. 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** (1999) 1185–1194
15. Haralick, R.M., Dinstein, Shanmugam, K.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics **3** (1973) 610–621
16. Rautkorpi, R., Iivarinen, J.: A novel shape feature for image classification and retrieval. In: Proceedings of the International Conference on Image Analysis and Recognition (ICIAR'04). (2004) 753–760
17. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on feature distributions. Pattern Recognition **29**(1) (1996) 51–59
18. Liao, S., Zhu, X., Lei, Z., Zhang, L., Li, S.: Learning multi-scale block local binary patterns for face recognition. In: Advances in Biometrics. (2007) 828–837
19. 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). Volume 8033. (2013) 50–59
20. Gadermayr, M., Liedlgruber, M., Uhl, A., Vécsei, A.: Shape curvature histogram: A shape feature for celiac disease diagnosis. In: Proceedings of the 3rd International Workshop on Medical Computer Vision (MCV'13). (2014) 175–184
21. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05). (2005) 886–893
22. Duin, R., Juszczak, P., Pacl'ik, P., Pekalska, E., de Ridder, D., Tax, D., Verzakov, S.: PR-Tools4.1, a matlab toolbox for pattern recognition (2007)
23. McNemar, Q.: Note on the sampling error of the difference between correlated proportions of percentages. Psychometrika **12**(2) (1947) 153–157