Degradation Adaptive Texture Classification for Real-World Application Scenarios[1]

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

[a] *Institute of Imaging and Computer Vision, RWTH Aachen University, 52074 Aachen, Germany*
[b] *St. Anna Children's Hospital, Department of Pediatrics, Medical University Vienna, 1090 Vienna, Austria*
[c] *Department of Computer Sciences, University of Salzburg, 5020 Salzburg, Austria*
*e-mail: *Michael.Gadermayr@Ifb.rwth-aachen.de*

Abstract– Images captured under non-laboratory conditions potentially suffer from various degradations. Especially noise, blur and scale-variations are often prevalent in real world images and are known to potentially affect the classification process of textured images. We show that these degradations not necessarily strongly affect the discriminative powers of computer based classifiers in a scenario with similar degradations in training and evaluation set. We propose a degradation-adaptive classification approach, which exploits this knowledge by dividing one large data set into several smaller ones, each containing images with some kind of degradation-similarity. In a large experimental study, it can be shown that our method continuously enhances the classification accuracies in case of simulated as well as real world image degradations. Surprisingly, by means of a pre-classification, the framework turns out to be beneficial even in case of idealistic images which are free from strong degradations.

Keywords: texture classification, invariance, feature extraction, similarity measures, robustness
DOI: 10.1134/S1054661817010035

1. INTRODUCTION

For many decades, texture classification [1–22] has been a fundamental field in image processing and pattern recognition. The main issue in this field of research is to find a lower dimensional representation of textures which captures the intrinsic properties but simultaneously skips extrinsic ones caused by different image acquisition conditions such as illumination and pose. Although it is simple to declare what a good feature extraction method has to do, it is challenging to design such a method. The main question is, how to remove the non-discriminative extrinsic information while maintaining the important discriminative intrinsic information. We focus on three extrinsic properties (or degradations) which are known to affect the classification performances in real world applications if not being considered. These are blur, noise and scale-variations. Generally in medical and especially in endoscopic imaging, these degradations are frequently prevalent and often cannot be circumvented [23–26] due to the downsized sensors, punctual lightnings and difficult handling. In spite of their high relevance in practice, many highly discriminative state-of-the-art texture descriptors are not invariant [27] to these degradations.

[1] The article is published in the original.

A. Related Work

One way to deal with such extrinsic information is to develop features that are invariant to a certain property. There exists significant literature on developing rotation-invariant [9, 15–17, 20], scale-invariant [9, 15–17], affine-invariant [18], view point-invariant [14], deformation-invariant [9] and illumination-invariant [16, 19] feature extraction methods. Furthermore, there exists also literature on making descriptors invariant (or robust) to noise [20, 28] as well as blur [13, 22]. In the latter case, the term robustness is rather used than invariance. The difference in nomenclature should remind us that noise as well as blur are considered as degradations, whereas e.g. a different viewing angle or illumination just gives us another view of the texture. However, in the following we will use the term degradation for all kinds of extrinsic variations, although e.g. a scale change usually is not considered to be a degradation. This is done in order to keep conformity with nomenclature in previous work.

The proposed degradation adaptive classification framework can be interpreted in terms of a multiple classifier system [29]. More to the point it can be interpreted as a special case of a classifier selection system [30] as illustrated in the following section. The final classifier (which is based on a specific training subset) is selected according to the similarity (or degradation) measure. In recent work [31], classifier selection is utilized in a similar way for a different problem definition.

The similarly denoted domain adaptive classification [32, 33] aims at a different classification scenario, which is referred to as domain-change-scenario (see Section 1A1).

1.1. Classification Scenarios

In this sub-section, divergent classification scenarios are defined. In the following we assume to have separate training and evaluation sets for image classification, both containing extrinsic variations. Furthermore, the distributions of the variations in both sets are similar. This scenario is referred to as standard-scenario which is highly relevant for many real world applications.

In analysis on invariant image description often a different scenario is chosen which is based on an idealistic training set and an evaluation set with any kind of extrinsic variation (or degradation). This scenario is furthermore referred to as invariance-scenario.

A third scenario, which is referred to as domain-change-scenario, is investigated in work on domain adaptation [32, 33]. In this case the extrinsic properties in the training set and in the evaluation set significantly differ (e.g. the training set is captured with one camera whereas the evaluation set is captured with another camera). In contrast to the invariance-scenario, this one generally deals with real-world image data.

Although there definitely are applications for all of those three scenarios, in the following we focus on the maybe most relevant standard-scenario as the proposed adaptive-classification framework is only applicable to this scenario.

1.2. Limitations of Invariant Descriptors

It seems to be highly beneficial to have descriptors which are invariant to all occurring extrinsic variations within an image database. Especially considering the invariance-scenario, descriptors necessarily have to be invariant if training is performed on idealistic data whereas evaluation is performed on degraded data. Recent work [34] showed for image data with variations in scale that this scenario is highly difficult and even features, declared to be scale-invariant, actually at best are invariant to a certain degree. However, in the invariance-scenario the necessity of invariance cannot be circumvented easily. The application of non-invariant image description approaches leads to distinct drops in classification accuracy [34].

If focusing on the standard-scenario (with varying but similarly distributed degradations in training and evaluation set), intuitively invariance should also be advantageous. Nevertheless, previous work [23, 34] showed that the utilization of state-of-the-art invariant feature extraction techniques in this scenario often leads to lower performances, compared to other highly discriminative (but non-invariant) descriptors. Obviously, many invariant image descriptors are developed for the invariance-scenario (in which the benefit of invariant features definitely is much higher) rather than the standard-scenario. To put it into a nutshell it can be stated that distinctiveness (discriminative power) often has to be sacrificed for achieving a high degree of invariance.

In the following, we outline why distinctiveness often is lost if making image representation invariant to certain degradations. Generally a feature extraction method can be interpreted as a function f: $\mathbb{R}^{N \times M} \rightarrow \mathbb{R}^{L}$, where N and M are the image dimensions and L is the feature dimensionality. Theoretically, if this function is invariant to a certain property, for two images I_1 and I_2 which are similar apart from the respective extrinsic property, $f(I_1)$ and $f(I_2)$ must be equal. Although such a function seems to be highly desirable, one significant problem is given by the discrete sampling of the processed signals. In the following we assume that f is a scale invariant feature extraction method and an image is captured in two different scales s_1 and s_2. Consequently, $f(I_1)$ must be equal to $f(I_2)$. However, this condition restricts the information content prevalent in $f(I_1)$ $(=f(I_2))$, as the image with the larger scale contains low frequency information which is not prevalent in the image with the smaller scale. On the other hand, the image with the smaller scale contains high frequency content that is not prevalent in the image with the larger scale. This shows that the scale invariant feature extraction method necessarily has to ignore that information. Moreover, we notice that a reasonable scale-invariant feature can never be completely scale invariant, but only invariant within a specific scale range, as otherwise the information content of the descriptor would totally deflate. Similar effects also apply to other degradations.

1.3. Impact of the Classifier

Especially in the standard-scenario, besides the feature extraction technique, the classification model has a major impact on the achieved classification accuracies. We consider two simple but highly intuitive classifiers, namely the k-nearest neighbor classifier and the Parzen-window classifier [35]. In case of the k-nearest neighbor classifier, the choice of the k-value adjusts the degree of non-linearity (i.e. the flexibility) of the decision boundaries. A similar behavior is exhibited by the Parzen window classifier when varying the kernel size. To optimize a classifier to a specific problem definition these values can be adjusted. A small k (or a small kernel width) in general more likely leads to overfitting whereas a larger value more likely leads to underfitting.

Again considering the standard-scenario with variable degraded images in the data set (leading to increased intra-class variations), we notice that a more flexible (non-linear) model can be advantageous to fit

Table 1. Overview of the content of this work compared to previous publications [37–39].

Publication	Scale-Adaptive Class. [37]	Adaptive Classification I [38]	Adaptive Classification II [39]	current work
Image data sets	1 synthetic set 1 real-world set	3 synthetic sets	1 real-world set	4 synthetic sets 5 real-world sets
Classifier	k-nearest neighbor	no restriction	no restriction	no restriction
Image descriptors	basic	basic	basic	basic and state-of-the-art
Dimensionality	1-dimensional	1-dimensional	multi-dimensional	multi-dimensional
Measure	scale	scale, blur, noise	scale, blur, noise, contrast	scale, blur, noise, contrast
Analysis	classification accuracies impact: similarity threshold classifier's scale selection	classification accuracies impact of classifier impact: reduced degradations	classification accuracies impact of classifier prior class probabilities	classification accuracies not-related degrad. measures impact of training set ratio obtained class probabilities impact of training set size runtime analysis

the data. However, a flexible model potentially leads to overfitting and generally requires a larger training set. State-of-the-art classifiers can be adjusted similar to the k-nearest neighbor and the Parzen window classifier. For example the support vector machine [36] can be made more flexible by using a non-linear kernel. Neural networks can be adjusted by varying the number of neurons.

B. Contribution

We have been inspired by our previous work on scale adaptive texture classification [37] which is based on another concept but a similar idea. The main restriction arises from the utilized classification model (k-nearest neighbor classifier) which cannot be changed easily. In a consecutive work [38], this limitation has been removed, by introducing a more general method which can be combined with arbitrary models. However, validation is only provided by few synthetic experiments without real world data. Motivated by promising outcomes, further work [39] based on the same idea investigating endoscopic image data has been performed later on. Additionally, in this work the adaptive classification approach has been extended to multiple dimensions in order to allow multiple degradations.

In a significantly larger experimental setup, in the current work we employ degradation measures which are not directly related to the prevalent degradations. Whereas in past work mainly fast and lean feature extraction methods are investigated, in this work more complex well known state-of-the-art texture descriptors [1, 2, 21] are additionally utilized to place more general statements. As the image database has a significant impact on the effect of degradation adaptive classification [38, 39], the method is applied to new image data sets [40–42] with and without strong degradations. Furthermore, the simulated data sets are adjusted in order to meet a more realistic scenario with fewer training data. Finally, several details are outlined and the effects of adaptive classification are explored extensively. We especially focus on the effects occurring during training set division such as the resulting training set sizes, the resulting prior distributions within the training sets as well as the impact of small training sets, which is highly relevant in practice. In Table 1, a brief comparison (based on keywords) to previous work is provided.

C. Outline

This paper is organized as follows: In Section 2 the degradation adaptive texture classification approach including the utilized degradation measures is described. In Section 3, the main experiments to evaluate the performance of our approach as well as experiments to allow greater insight into the method are presented and extensively discussed. Finally, Section 4 concludes this paper.

2. DEGRADATION ADAPTIVE CLASSIFICATION

First, we define two robustness types. If the classification accuracy in a certain problem definition does not strongly decrease when all images in a database

(training and evaluation set) are similarly degraded, a feature is denoted to be relatively robust with reference to a certain degradation. The notation absolute robustness is used, if the accuracy can be preserved even if the training and the evaluation set contain degradations with different extent (but the same type).

The basic idea of the degradation adaptive classification is based on the assumption that absolute robustness generally is harder to achieve than relative robustness (which is shown in Section 3 A). Therefore, we divide our data sets into smaller data sets with similar properties.

To put it into a nutshell, the evaluation set is partitioned into (non-overlapping) subsets, to ensure that the class correspondence of each sample in the evaluation set is computed exactly once. The training set in general is not partitioned, but separated by overlapping intervals to prevent from too small training sets.

A. One-Dimensional Approach

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

$$T_i = \{I \in T : d \le D(I) \cdot C - i < d + 1\}, \quad (1)$$

where $i \in \{0, 1, ..., C - 1\}$, C denotes the cardinality of the set of generated subsets and I is an image in the training set. The parameter d determines how the degradation measures for each training subset would overlap with the degradation measures of adjoining subsets. A large C leads to smaller subsets and consequently a higher similarity within one set. This is quite obvious if considering Eq. 1 as the interval sizes (which is e.g. $\left[\frac{d}{C}, \frac{d+1}{C}\right)$ for $i = 0$) decrease with an increasing C. As the interval gets smaller, the images in that subset are in general fewer in number and more similar in appearance.

If d, which defines the overlap, equals zero, the original data set is partitioned. Especially in case of a large C, it is potentially sensible to create overlapping subsets ($d > 0$), to ensure that the subsets for training do not get too small. Especially in case of a small original training set, an overlap larger than zero should be chosen to prevent the new training sets from getting too small. Actually, the choice of a suitable overlap value d highly depends on the database, the chosen number of partitions C, the classifier and the feature extraction technique. A discussion will follow in the experimental section (Section 3). The normalized degradation measure D must be in the half open interval [0, 1), in order to allow a real partitioning of the trainings set T, if d is set to zero. This kind of normalization is obtained by means of min-max-normalization of the non-normalized degradation measure $\hat{D}\left(D(I) = \frac{\hat{D}(I) - \hat{D}_{\min}}{\hat{D}_{\max} - \hat{D}_{\min} + \epsilon}\right)$, where $\hat{D}_{\min}$ is the lowest and $\hat{D}_{\max}$ is the highest degradation measure in the data set and ϵ is a small constant. The small constant ϵ is fixed to 10^{-6} and is only required to obtain the half open interval (and not for parametrization).

The evaluation set $E \subset \Omega$ is partitioned into the subsets in a similar manner

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

where outliers must be set to 0 or $1 - \epsilon$, respectively

$$D'(I) = \max(\min(D(I), 1 - \epsilon), 0). \quad (3)$$

This must be done, to ensure that each sample belongs to one evaluation subset, which can only be guaranteed when all degradation measures are within the interval (0,1]. Finally for each i, the evaluation set E_i is classified by the discriminative model based on T_i.

We also considered subdivision strategies based on clustering (e.g. k-means) instead of the equidistant degradation metric intervals. However, these more elaborated methods did not lead to improved accuracies.

This methodology could also be interpreted in terms of a classifier selection system [30]. Classifier selection is done by means of a degradation measure D, based on the images in the training set. In case of our approach, the selection step is done quite simply by generating equidistant linear intervals. The decision of this selection defines one specific classifier (which is based on a specific training set) to compute the final decision. In opposite to a multi-classifier system such as ensembles, only the decision of the selected classifier must be evaluated.

In recent work [37] a scale-adaptive classification method has been introduced. In this work, for each element in the evaluation set, a separate training subset is constructed. As a consequence, only classifiers with highly lean learning stages (like the k-nearest neighbor classifier) can be efficiently utilized. The current approach allows the utilization of arbitrary classifiers. The computational costs can potentially even be improved compared to the straight-forward classification, as the training of a set of classifiers based on smaller data sets often is less costly than the training based on one large data set. However, with large overlaps (d), this positive effect vanishes.

B. Multi-Dimensional Approach

The proposed 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

$$T_{i_1,\ldots,i_n} = \{I \in T : \bigwedge_{j=1}^{n} d_j \le \pi_j(D(I)) \cdot C_j - i_j < d_j + 1\}, \quad (4)$$

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 j selects the jth 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

$$E_{i_1,\ldots,i_n} = \{I \in E : \bigwedge_{j=1}^{n} 0 \le \pi_j(D'(I)) \cdot C_j - i_j < 1\}. \quad (5)$$

Finally for each n-tupel (i_1, ..., i_n), the evaluation set $E_{i_1,\ldots,i_n}$ in is classified by the discriminative model based on $T_{i_1,\ldots,i_n}$.

The appropriate choice of C as well as d, is highly decisive in order to raise the classification accuracies. For this work, the subset cardinality C is fixed to a sufficiently large number (32). This restriction on the one hand limits the search space, as only d has to be evaluated further more. On the other hand a too large C value does not affect the classification accuracy if d is adjusted appropriately. We decided to choose d individually for each training set T_i in a way that the training set size equals a fixed number, which allows a more intuitive analysis of the results. This fixed number is referred to as training set ratio (TR). The chosen potential training set sizes are outlined in Section 3.

C. Degradation Measurement

In order to divide a data set into several smaller ones with higher similarities (with reference to a degradation), a metric D to capture this similarity is required. In this work, we especially focus on the degradations noise, blur and scale, which are explicitly captured by three of the following degradation measures. Furthermore, we investigate a contrast measure which does not extract one of the degradations mentioned above, however, previous work provides evidence that this method is appropriate for adaptive classification [39].

A quantitative assessment of the proposed methods with respect to the obtained classification performance is provided in the experimental section.

• **Noise Metric** (D_n): Noise can be quantified by computing the absolute difference between an image and the Gaussian filtered (with $\sigma = 1$ and a kernel size of 3 pixels) version of the same image [38]. Based on the thereby obtained noise image

$$I_n = |I - I * G_1|, \quad (6)$$

the scalar noise measure D_n is obtained by calculating the pixel-wise arithmetic mean over all pixels.

• **Blur Metric** (D_b): To measure blur, the metric introduced in [43] is deployed. For computing this measure, first in the horizontal direction, the edges Z are identified by extracting all local minima and maxima for each row. Each edge $z \in Z$ is a tupel consisting of a length z_l (geometric distance in the image between the minimum and the maximum) and a magnitude z_m (pixel difference between the maximum and the minimum). For blur estimation, we utilize the average ratio between the overall lengths and magnitudes of the edges D_b:

$$D_b = \frac{1}{\#Z} \sum_{z \in Z} \frac{z_l}{z_m}. \quad (7)$$

This ratio is finally deployed for blur estimation. In case of a sharp edge, the length value is small and the magnitude is high which leads to a low blur metric. In opposite a blurry edge has a large length value and a small magnitude which leads to a high blur metric.

• **Scale Metric** (D_s): For scale estimation, we utilize a scale-space based method [37]. To estimate the global scale of an image, first a scale space is constructed by convolving an image with Laplacian-of-Gaussian filters in varying scales σ. As proposed in the previous paper [37], for the Lapacian-of-Gaussians, the scales $\sigma = \hat{c}\sqrt{2}^k$, $k \in \{-4, -3.75, ..., 7.75, 8\}$ are chosen (with $c = 2.1214$). The pixel-wise scales are achieved by using the index of the maximum responses. Finally, the global scale D_s for an image is estimated by computing a histogram of this scale value over all pixels, followed by a Gauss-fitting. The final scale measure is given by the mean of the fitted Gaussian kernel.

• **Contrast Metric** (D_c): Contrast is computed from the gray-level co-occurrence matrix [44] specifying a distinct pixel neighborhood.

The measure D_c is obtained by summing up the absolute differences of all combinations of neighboring pixel values multiplied with the probability p of the occurrence.

$$D_c = \sum_{i,j} |i - j|^2 \cdot p(i, j). \quad (8)$$

As successfully applied in previous work [39], an offset (geometric distance between i and j) of six pixels is utilized. Interestingly, the contrast property is not (only) able to measure degradations, but it has also been used as discriminate feature e.g. in celiac disease diagnosis [45]. Although this method obviously does not explicitly measure any kind of degradation, we would like to investigate the effect of such a "discriminative" metric on the adaptive classification framework.

• **Other Measures:** In previous work [39], further measures such as variance and mean have been investigated. However, as these measures were less powerful than the ones declared in this section, we do not consider them in this work.

Table 2. Execution runtimes for degradation measurement and feature extraction

Degradation measure	Runtime, ms	Feature extraction	Method runtime
Dn (Noise measure)	1	MRLBP	20 ms
Db (Blur measure)	28	ECM	23 ms
Ds (Scale measure)	234	MFS	198 ms
Dc (Contrast measure)	17	DTCWT	71 ms

• **Combinations of Metrics** (e.g. D_{bc}): Furthermore, we investigate the multi-dimensional adaptive classification framework and thereby consider combinations of two and three of the declared degradation metrics. Combined metrics are easily obtained by means of vector concatenation of single measures. Further details on dealing with combined metrics are outlined in Sect. 2 B In the following, the combination of the two metrics e.g. D_b and D_s is referred to as metric D_{bs}.

D. Computational Complexity

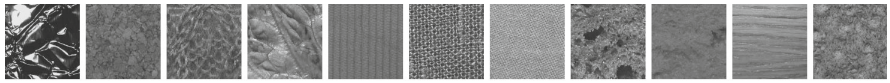
In this section, focus is on the computational complexity of the proposed framework. Here we have to separately investigate the training and the evaluation step. For both steps, the degradation measures for each image have to be computed. During classifier training, first a degradation measure must be computed per image and the training of a large set (depending on the chosen C) of classifiers must be performed. However, the training sets are significantly smaller compared to the traditional scenario (depending on the chosen overlap d). During evaluation, the degradation measures must be computed for each image and consequently one classifier is chosen for computing the final decision.

As different classifiers (with different complexities) can be used and the training set size as well as the optimal parameters C and d vary, it is hard to give a proper estimation about the impact of the degradation adaptive classification framework on the training runtime. Nevertheless, the more important aspect is given by the computational efforts required for evaluation. For this, the only additional time-consuming step in case of the new framework is the computation of the degradation measure, as the subsequent classifier selection is not worth mentioning and the final classification is similarly computationally expensive. In Table 2, the execution times for the degradation measures compared to some feature extraction methods (which are specified in Section 3) are given. Depending on the chosen feature extraction method and degradation measure, the additional effort varies significantly. For example if using two intermediate methods MRLBP and the blur measure the overall evaluation runtime is approximately doubled (from 20 ms which are required for feature extraction to 20 ms + 28 ms for feature extraction plus degradation measuring).

3. EXPERIMENTS

Experiments are performed based on nine different databases. Each database consists of 3 similarly sized data sets. One of them is used for training, one for parameter estimation and one for evaluation. In a second run, the evaluation and the estimation data set are switched and the overall classification accuracies (for analysis) are averaged to increase robustness. Five of the databases are based on the original Kylberg database [46], which consists of 28 classes and 40 images per class. The KB-STD database consists of the original Kylberg set, cropped to a size of 128 × 128. KB-SCALE is based on the same images, which are randomly downscaled and also cropped to 128 × 128 pixels. For each image, one the of downscaling factors $\{2^{0.00}, 2^{0.25}, 2^{0.50}, 2^{0.75}, 2^{1.00}, 2^{1.25}, 2^{1.50}, 2^{1.75}, 2^{2.00}\}$ is randomly chosen. By using the original 576 × 576 Kylberg patches for downscaling, the size of 128 × 128 can be preserved (even with the largest downscaling-factor 2 × 2). KB-BLUR is constructed similarly, by randomly adding blur to the images. The blurred images are simulated by applying a Gaussian filter with randomly chosen a values within {0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0}. The same is done in case of KB-NOISE. For that, Gaussian white noise is applied with a being within {0, 4, 8, 12, 16, 20, 24, 28, 32}. The effect of the different kinds of degradations is shown in Fig. 2. The KB-ALL database consists of all images of KB-SCALE, KB-BLUR and KB-NOISE and thereby contains all three kinds of simulated degradations. KTH2 is the abbreviation for the popular KTH-TIPS2 database [41] which consists of different textures and real (non-simulated) scale, pose and illumination variations (see Fig. 1a). The CELIAC database [47] (see Fig. 1b) consists of endoscopic images captured during esopha-gogastroduodenoscopies at the St. Anna Children's hospital. The goal of this problem definition is to discriminate between healthy patients and patients suffering from celiac disease, based on visual markers. Furthermore our approach is tested with the well known CURET database [40]. The samples are downscaled by factor two for more efficient computation. To generate three distinct data sets, the samples (of each class) beginning with "01", "02" and "03" are assigned to the respective data sets. Finally tests are executed with the UIUC database [42]. Similarly, these images are downscaled by factor two for boosting efficiency. Three data sets are generated by randomly partitioning the original set into three similarly sized image data sets. For a concise summary of the image data used, we refer to Table 3.

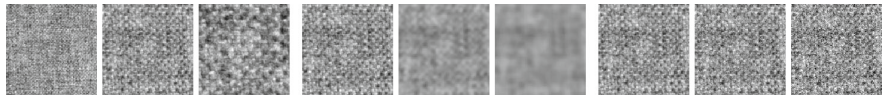
For degradation adaptive texture classification, the training set size must be evaluated for each configuration. As mentioned in Section 2, we do not fix the overlap d, but instead fix the ratio between the number



(a) The KTH2 database: This figure shows one example patch per class

(b) The CELIAC database: This figure shows three images of healthy (left) and three images of diseased mucosa (right).

Fig. 1. Example texture patches of the different databases.



(a) Simulated scale-variations (KB-SCALE) (b) Simulated blur (KB-BLUR) (c) Simulated noise (KB-NOISE)

Fig. 2. This figure shows the three strengths of simulated degradations (original, moderate and maximal degradations) in case of KB-SCALE, KB-BLUR and KB-NOISE and one specific texture patch.

of images in the original training set and in the new subsets. For this purpose, we defined sensible training set ratios (*TR*) {1, 1/2, 1/4, 1/8, 1/16, 1/32, 1/64, 1/128}. For example, a *TR* value of 1/4 means, that for each training subset in adaptive classification, *d* is individually adjusted that the number of images in the new training set is one forth of the overall training set. *TR* is evaluated based on the separate data set for optimization. The number of subsets *C* has been fixed to 32. Experiments showed that a *C* larger than 32 does not lead to further accuracy increases.

For final classification, we deploy two different established classification models consisting of a nearest neighbor classifier (NN) and a linear support vector machine (SVM) [36].

The linear SVM is a maximum margin classifier computing a linear decision boundary separating the samples of two classes and equally maximizing the margin between the classes. The nearest neighbor classifier's more flexible decision boundaries are implicitly given by the Voronoi diagram.

By investigating these two opposing classifiers, we aim at getting more insight into the impact of the classifier (and the decision boundaries) on the classification error rates.

Classifiers are implemented in MATLAB. We utilize an in-house NN classifier implementation and liblinear's [36] C-SVM. The cost value of the C-SVM is obtained by means of inner cross validation.

Table 3. Information regarding the databases used in the experiments.

Dataset	Image-size	Degradations	DB-size	Classes
KB-STD	128 × 128	high image quality	1.120	28
KB-SCALE	128 × 128	simulated scale variations	1.120	28
KB-BLUR	128 × 128	simulated Gaussian blur	1.120	28
KB-NOISE	128 × 128	simulated Gaussian white noise	1.120	28
KB-ALL	128 × 128	simulated scale variations, blur, noise	3.360	28
KTH2	100 × 100	real scale variations, pose, illuminations	1.173	11
CELIAC	128 × 128	real scale variations, blur, noise	310	2
CURET	100 × 100	different viewing and illumination directions, noise	2.500	61
UIUC	320 × 240	high image quality	325	25

For feature extraction, the following well known techniques are deployed:

• Multi-Resolution Local Binary Patterns [10] (MRLBP): Local Binary Patterns describe a texture by means of the joint distribution of pixel intensity differences represented by binary patterns. We deploy the uniform version, capturing only patterns with at most two bit-wise transitions with eight neighboring samples. To achieve a higher degree of distinctiveness, the LBP feature vectors with a radius of one and a radius of two are concatenated resulting in this multi-resolution descriptor.

• Edge Co-occurrence Matrix [12] (ECM):

After applying eight differently orientated directional filters, the orientation is determined for each pixel, followed by masking out pixels with a gradient magnitude below some threshold t. Finally, the ECM is achieved by computing the gray-level co-occurrence matrix of these data and a specified displacement υ. For the experiments, t is set to 25% of the maximum response and the displacement vector $\upsilon = (1, 1)$ is used.

• Multi-Fractal Spectrum [14] (MFS):

The local fractal dimension is computed pixel-wise utilizing three different types of measures for obtaining the local density. These three measures consist of (1) the average value, (2) the first derivation and (3) the second derivation. The fractal dimension is obtained by computing the global (averaged) change between 25 different scales. The final multi fractal feature vector is built by concatenation of the three fractal dimensions.

• Dual-Tree Complex Wavelet Transform [21] (DTCWT):

This image descriptor is based on fitting a two-parameter Weibull distribution to the wavelet coefficient magnitudes of sub-bands obtained from the dual-tree variant of the complex wavelet transform. Decomposition is performed on five levels leading to a 60 dimensional image representation.

• Improved Fisher Vectors [2] (IFV):

Fisher Vectors [6], as well as the next descriptor (VLAD), is a global mid-level image representation which is obtained by pooling local image descriptors. These state-of-the-art de facto standard methods build up and improve the idea of the Bag-of-visual-words approach [48] which has become highly popular in past years. In case of Fisher Vectors, the Gaussian mixture model is used to construct a dictionary, based on a local descriptor. For this local descriptor, we use the well known SIFT (Scale-invariant Feature Transform) [49] feature. The final Fisher vector contains information how the parameters of Gaussian mixture model have to be modified to better fit the data. This is done by concatenating the means and the covariance deviation vectors. We use the improved Fisher vectors [2] which are derivatives based on two ideas. Instead of the linear kernel IFV uses the non-linear Hellinger's kernel which is based on the Bhattacharyya distance. Furthermore, the final feature vector is L^2 normalized.

• Vector of Locally Aggregated Descriptors [3] (VLAD): VLAD is technique which is similar to Fisher Vectors. In opposite to Fisher Vectors is does not store any second-order information. Furthermore is uses k-means clustering instead of a Gaussian mixture model to generate the feature vocabulary. The feature vectors finally store information of the difference between the cluster centers and the pooled local descriptors.

• Random Feature (RAND): Finally we investigate a random feature (random scalar) not depending on the input image. Although this feature is not useful in practical scenarios as it does not provide any discriminative power, it helps us to understand the effects of degradation adaptive texture classification.

The methods MRLBP, MFS, DTCWT and ECM are in-house MATLAB implementations. In case of IFV and VLAD, we utilize the VLFeat library [50].

A. Experiment 1. Robustness-Analysis

First of all, we investigate the robustness of three features (MRLBP, ECM and MFS) with respect to specific (simulated) degradations using the databases KB-SCALE, KB-NOISE and KB-BLUR.

We distinguish between two different robustness types. If the classification accuracy does not strongly decrease in case all images in a database (training and evaluation set) are similarly degraded, a feature is denoted to be "relatively robust" with reference to a certain degradation. The notation "absolute robustness" is used, if the accuracy can be preserved even if the training and the evaluation set contain degradations with different extent.

For each degradation type, we construct nine training and nine evaluation data sets, reaching from non simulated degradations (0) to strong degradations (8). Each of the data sets contains the same original images with a dissimilar degree of applied degradation. Details on the degradation simulation are provided in Sect. 3.

In Fig. 3, the robustness of the investigated feature extraction techniques with respect to the three degradations scale, noise and blur is presented. If the training and the evaluation data set continuously suffer from similar degradations, the accuracy only moderately decreases in most combinations of features and modes. These outcomes are shown on the diagonal axis in each subplot. A high value in the bottom-right part of the diagonal indicates that the feature has a high relative robustness. If the level of degradation in

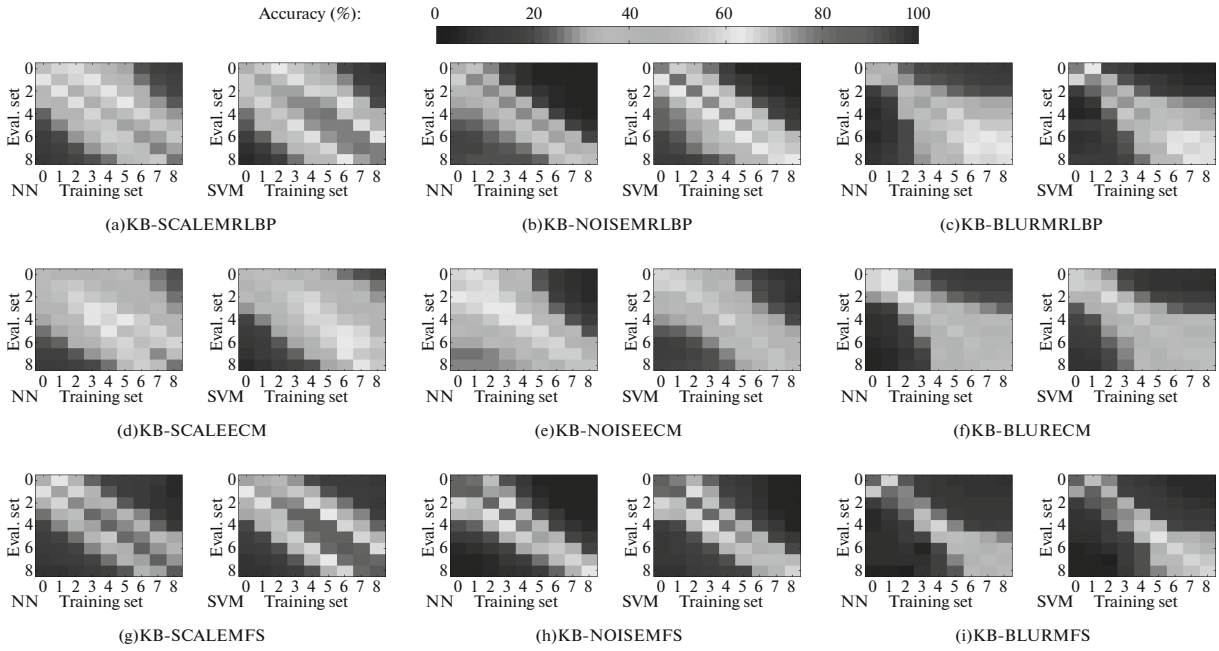


Fig. 3. Classification accuracies in a scenario with degradations of different extent in the training set (horizontal axis) and the evaluation set (vertical axis) for both classifiers (NN, SVM). The value 0 (on x- and y-axis) corresponds to non-degraded data whereas 8 corresponds to the strongest degradations.

training and evaluation set differs, measuring the absolute robustness, the decrease in accuracy is by far more significant in case of all features. This behavior has been expected in case of the scale-degradation, as a different scale in general is not considered to be a "degradation" if all images in a database have the same scale. However, the behavior is similarly significant in case of noise and blur, which is a very interesting outcome. A highly distinct behavior is observed in case of the degradation noise (and especially the features MRLBP and MFS). Apparently it is hard to obtain absolute robustness to noise, whereas relative robustness seems to be achievable quite easily. This behavior is even less distinct in case of scale variations, which is another surprise, as we expected that most features would be highly relatively robust to scale-variations as it does not represent a real "degradation" as explained in the introduction. Considering the features MRLBP and MFS, it can be seen that the differences between relative and absolute robustness are quite similar. In case of ECM, the difference between the relative and the absolute robustness is smaller, which might be due to its generally smaller discriminative power. Considering the two investigated classification models (SVM and NN), we do not detect significant differences related to relative and absolute robustness.

Summing up, it does not matter if textured images are slightly blurred or if they slightly suffer from noise, if all images in a database similarly suffer from the respective inadequacy. But on the other hand if the training and the evaluation set suffer from variable degradation strengths, we observe a strong decrease in accuracy.

The large differences between the relative and the absolute robustness indicate that the accuracies could be increased using degradation adaptive classification, as this method divides the data into several smaller sets with a higher degree of similarity. In the following experiments, the classifier is supposed to have a larger impact as highly non-linear classifiers (e.g. the nearest neighbor classifier) in general are able to preselect similarly degraded features [37] for classification which cannot be performed in case of linear SVM's classification.

B. Experiment 2. Adaptive Classification with simulated data

In Figs. 4 and 5, the classification accuracies obtained with simulated degradations are shown. In these plots, the accuracies achieved with the adaptive classification framework in combination with simulated image degradations (dashed lines) are compared with the classification rates of traditional SVM and NN classification (solid lines). One subplot provides the accuracies (on the vertical axis) for all degradation measures and combinations of degradation measures (on the horizontal axis), for both classifiers, one distinct feature and one distinct database. The short hor-

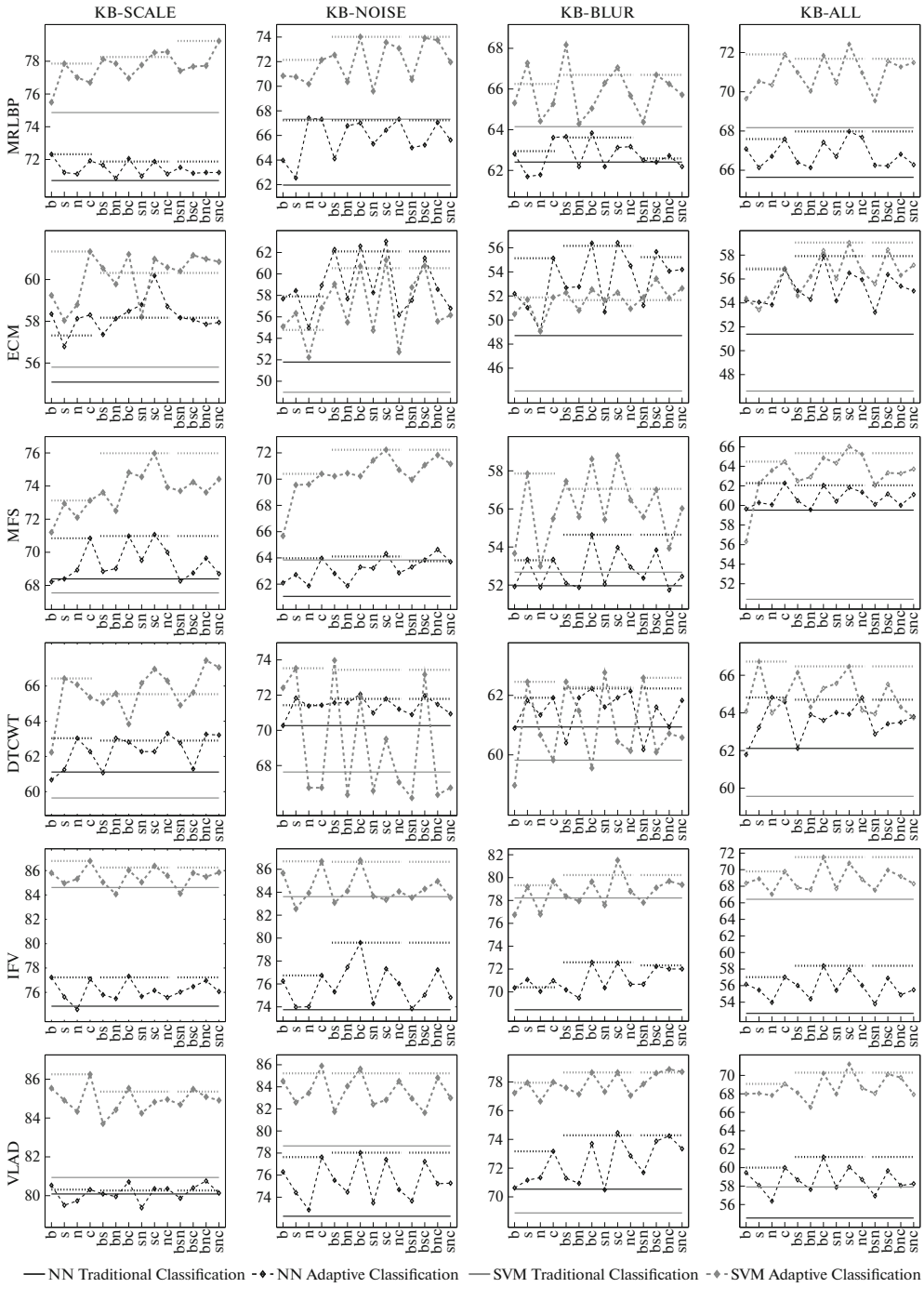


Fig. 4. Accuracies (vertical axis) of adaptive classification in combination with simulated degradations and varying degradation metrics (e.g. bs on the horizontal axis indicates the blur measure D_{bs}). The short horizontal dotted lines indicate the best accuracy based on one-dimensional (left line), at most two-dimensional (center line) and three dimensional metrics (right line).

izontal dotted lines indicate the best accuracy obtained with cross validation based on one-dimensional metrics (left line), at most two-dimensional metrics (center line) and based on all available metrics (right line). Notice that these rates not necessarily correspond to the best achieved accuracies, since the metric in this case is chosen based on the estimation data set to avoid bias due to overfitting.

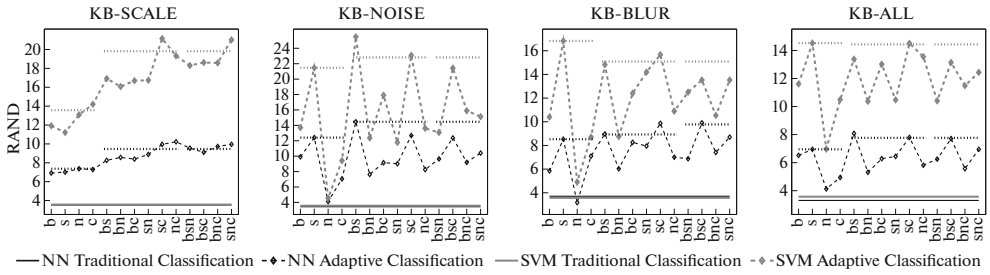


Fig. 5. Accuracies (vertical axis) of adaptive classification in combination with simulated degradations and varying degradation metrics using the random (RAND) feature extraction method (as in Fig. 4).

First, we only consider the one-dimensional (single) degradation measures D_b, D_s, D_n and D_c. We notice improvements in virtually all combinations of a feature, a degradation measure, a database and a classifier. Furthermore is can be seen that especially in case of KB-NOISE, but also in case of KB-BLUR and KB-SCALE, the highest accuracies are not achieved with the corresponding assembled measure. Especially the scale and the contrast measure correspond to good performances in general. One of these measures produces the best rates in almost each case. Interestingly, it is hard to detect any connections between the ideal measure and the simulated degradation (which corresponds to the respective data set). On the other hand, especially the measure D_c seems to be most appropriate on average although it does not directly measure any of the simulated degradations. This is actually a quite interesting outcome. As such a behavior has not been expected, in previous work [38], experiments were only performed with corresponding degradation measures and simulated degradations.

Considering the two different classifiers, we notice that in case of traditional classification (indicated by the horizontal solid line), in some cases the nearest neighbor classifier delivers better or at least competitive results. This is supposed to be due to the fact that the (highly non-linear) nearest neighbor classifier is able to (implicitly) choose a nearest neighbor with a similar degradation level. This effect has been analyzed in recent work [37]. The linear SVM, in opposite, is not able to factor out images with dissimilar degradations, which is disadvantageous in a scenario with different degradation strengths. Especially in the scenario with combined degradations (right row in Fig. 4), the nearest neighbor classifier is quite competitive compared to the linear SVM. However, in case of adaptive classification this effect is mostly reversed. Considering higher degrees of similarity in the smaller data sets, the NN classifier obviously profits less distinctly from the highly nonlinear underlying decision boundaries.

Next, we focus on multi-dimensional degradation measures. Especially we ask, if the best combination of two dimensional features delivers any improvements. This question can be answered if regarding the left dotted lines and the center dotted lines. We observe that in the majority of our settings (38 out of 56), the utilization of two-dimensional metrics leads to improved outcomes. Especially a combination of the best one-dimensional metrics (which are often D_c and D_s or D_c and D_b) seems to be advantageous. On the other hand, utilizing three dimensional degradation measures (right dotted lines), the accuracies hardly ever can be improved further.

Considering the different feature extraction methods, we observe that generally higher improvements are obtained in case of methods with a weaker performance based on traditional classification. Especially with the (artificial) RAND feature (see Fig. 5), quite significant improvements are observed, although this descriptor does not store any distinctive information. Obviously the improvement is only due to the change of the prior distribution in the generated sub data sets. However, even with the high performing methods (MFS, VLAD, IFV) distinct improvements are obtained. With each database, the highest overall accuracies are obtained using adaptive classification, which is maybe most relevant in practice.

C. Experiment 3. Adaptive Classification with real-world data

Now we investigate the impact of adaptive classification on real-world image data without any simulated degradations. Whereas two databases are widely free from any strong image degradations (KB-STD, UIUC), the others suffer from more or less distinct degradations. This is especially the case considering the CELIAC database. An overview of the image databases is given in Table 3. Figure 6 shows the achieved classification performances with the real-world databases, similarly presented as in Fig. 4. Considering the results, we observe that in case of most configurations, again the measures D_s and D_c seem to be most appropriate. Adaptive classification consistently improves the performances in case of almost each combination of a database and a feature extraction method. The best outcome for each individual image database is

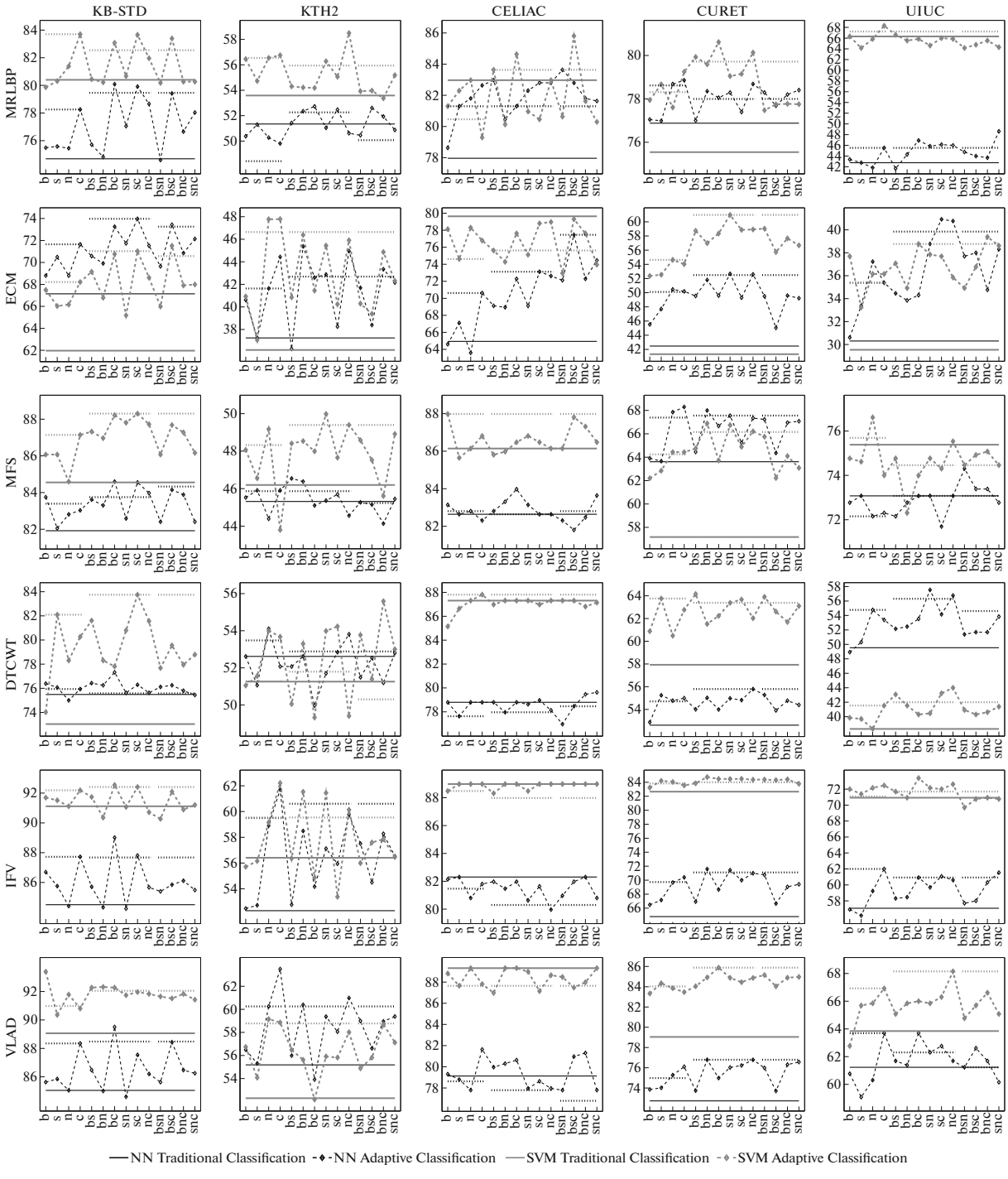
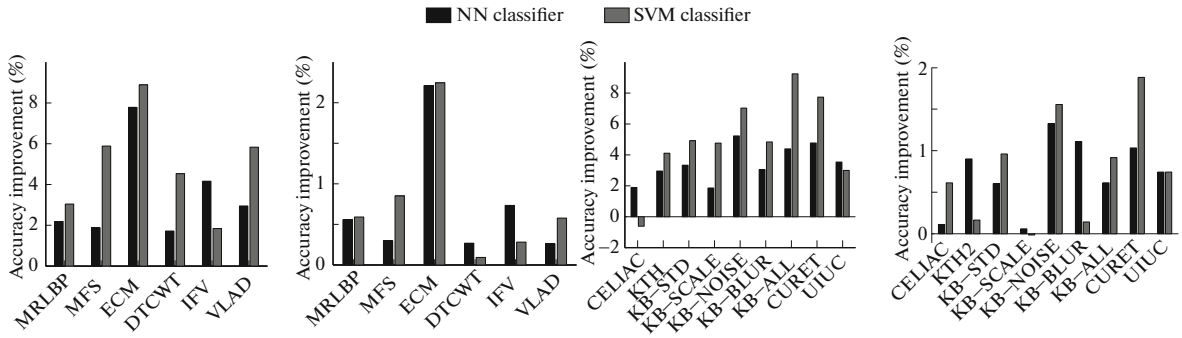


Fig. 6. Accuracies (vertical axis) of degradation adaptive classification in combination with real-world image data, separately for each degradation measure, each database and each feature.

obtained using adaptive classification. As in the synthetic scenario, two-dimensional degradation metrics again seem to be even more effective than one-dimensional ones, whereas three-dimensional measures do not improve the performances in general. Even with the quite idealistic, distortion-free databases KB-STD and UIUC significant improvements are observed. An interesting behavior is shown by the CELIAC data-



(a) Average accuracy improvement of adaptive classification (compared to traditional classification) per feature and classifier

(b) Average accuracy improvement of two-dimensional adaptive classification (compared to one-dimensional) per feature and classifier

(c) Average accuracy improvement (compared to traditional classification) per dataset and classifier

(d) Average accuracy improvement of two-dimensional adaptive classification (compared to one-dimensional) per dataset and classifier

Fig. 7. Overview of the classification accuracy improvements with adaptive classification.

base. Although these images suffer from many different kinds of strong degradations, this database benefits only in some cases. This is supposed to be due to the relatively easy two-classes classification problem. As the classifier has to be distinguished only between two different classes, strong intra-class variations (caused by varying degrees of degradations) could be compensated more easily even without the adaptive classification framework. This hypothesis is supported by the fact that the more flexible NN classifier is highly inferior (even in case of traditional classification) in case of this image data set. Again, we notice that the most appropriate degradation measure does not strongly correlate with the degradations prevalent in the images. For example, the KTH2 database which mostly suffers from scale variations, cannot be most accurately classified using the scale measure. Noise and contrast are appropriate for this data set.

In Fig. 7 the large amount of data is variably summarized to provide and overview from different points of view. Figure 7a shows the average improvements in case of adaptive classification averaged over all databases. We notice that ECM and VLAD on average profit most significantly. A high improvement of ECM has been expected as it is one of the features with the lowest accuracies on average and previous work [38] showed that with low-performing features more distinct improvements are expected. Much more remarkable is the strong improvement of VLAD which is a state-of-the-art method and corresponds to a high distinctiveness in the investigated problem definitions. As already mentioned the SVM classifier profits more (right bar) than the NN classifier (left bar) on average. If considering the improvements between the best one- and the best two-dimensional degradation measure 7b, a similar outcome can be observed. Features which profit more distinctly from adaptive classification in general also profit more distinctly from the second dimension. In Fig. 7b,d, a similar overview is given with respect to the different databases. We notice that with all databases, except for CELIAC, on average improvements are observed. Interestingly, we do not observe a strong connection between the degree of degradation and the classification performance increase. However, we notice a connection between the number of classes and the degree of improvement. Especially with the CURET database (61 classes) strong improvements are observed. On the other hand, with KTH2 (11 classes), UIUC (25 classes) and CELIAC (2 classes) the improvements are significantly smaller. We suppose that this is due to the more difficult classification problem in case of more classes.

D. Experiment 4. Impact of the Training Dataset Size

In recent work [39], we assumed that a small original training set size might affect the adaptive classification framework. The even more decreased number of training set images during adaptive classification could lead to problems during classifier training. To investigate this effect, the KB-STD data set is used with different training set sizes. Therefore, we randomly select a specific number of samples for training, in order to evaluate the impact on the classification accuracy in case of traditional and degradation adaptive classification. In Fig. 8 the accuracies with decreased training set sizes are shown for both classifiers with traditional and adaptive classification. As the outcome is similar for all features, we only show the results for three feature extraction methods and two degradation measures. We observe that the accuracies consistently drop with decreasing training set sizes in case of all features and all degradations measures, which is not surprising. However, this decrease similarly concerns traditional and adaptive classification. As the benefits, generally do not vanish in case of a reduced training set, the adaptive classification framework can be effectively utilized even in case of

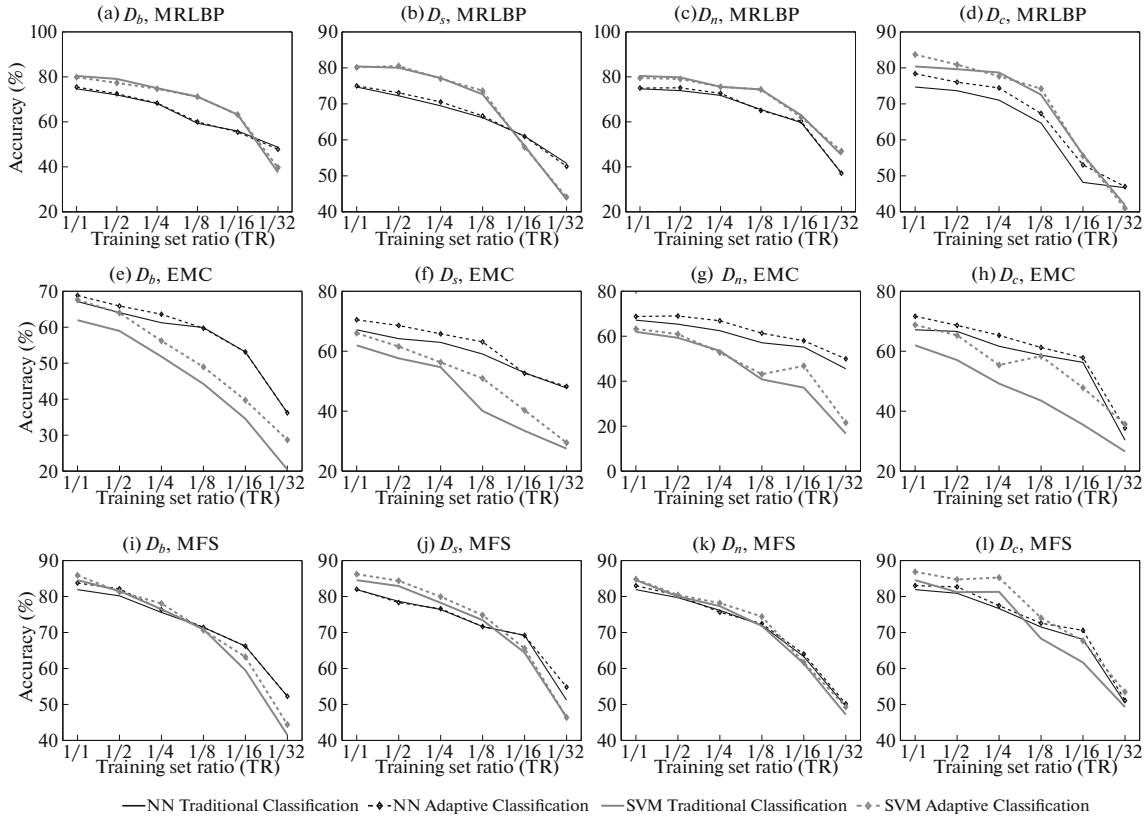


Fig. 8. The impact of randomly reducing the training set size (horizontal axis) on the classification accuracies (vertical axis) based on one image database (KB-STD). *TR* in this case refers to the factor of training data reduction (reaching from 1 to 1/32).

quite small training data sets. Strong accuracy decreases compared to traditional classification are unlikely, as the adaptive classification method evaluates an overlap, which can be set in a way that all elements in the traditional training set are in the new training sets. Thereby it can be stated that degradation adaptive classification is a generalization of traditional classification and a fallback (if adaptive classification is disadvantageous) is automatically (implicitly) provoked if adaptive classification would be adversely.

CONCLUSION

We have shown that relative robustness to degradations is easier to obtain than absolute robustness. This knowledge is exploited by the proposed degradation adaptive classification framework. Experimentation has shown that the classification accuracies can be improved by our method in combination with all evaluated image representations and all classification models in combination with simulated image degradations. Even with real world databases as well as databases showing no strong degradations, distinct improvements are observed. Experiments showed that this is very likely due to the change of the prior probabilities caused by the degradation measures. The most appropriate degradation measure cannot be predicted easily, as the effects of the respective measures highly depend on the utilized image data, the classifier as well as the feature extraction technique. Even if significant degradations of any kind are prevalent it is not clear if the measure capturing this property actually leads to the best outcomes. As expected, the linear classification model benefits more distinct from the new technique compared to the highly non-linear nearest neighbor classifier. Finally it has been shown that improvements are obtained even in case of small training data sets.

ACKNOWLEDGMENT

This work has been funded by the Austria Science Fund (FWF 24366).

REFERENCES

1. M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'14)* (Columbus, OH, 2014), pp. 3606–3613.

2. J. Sánchez, F. Perronnin, T. Mensink, and J. J. Verbeek, Int. J. Comput. Vision **105**, 222 (2013).
3. H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, IEEE Trans. Pattern Anal. Mach. Intellig. **34**, 1704 (2012).
4. R. Timofte and L. J. V. Gool, in *Proc. British Machine Vision Conf. (BMVC'12)* (Univ. of Surrey, Guildford, 2012), pp. 1–12.
5. G. Sharma, S. ul Hussain, and F. Jurie, in *Proc. Europ. Conf. Computer Vision (ECCV'12)* (Firenze, 2012), pp. 1–12.
6. F. Perronnin and C. Dance, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'07)* (Minneapolis, 2007), pp. 1–8.
7. H. Jegou, M. Douze, C. Schmid, and P. Perez, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'10)* (San Francisco, 2010), pp. 3304–3311.
8. L. Liu, P. Fieguth, G. Kuang, and H. Zha, in *Proc. IEEE Int. Conf. on Computer Vision (ICCV'11)* (Barcelona, 2011), pp. 391–398.
9. L. Sifre and S. Mallat, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'13)* (Portland, 2013), pp. 1233–1240.
10. T. Ojala, M. Pietikäinen, and D. Harwood, Pattern Recogn. **29**, 51 (1996).
11. N. Dalal and B. Triggs, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (CVPR'05)* (San Diego, 2005), Vol. 1, pp. 886–893.
12. R. Rautkorpi and J. Iivarinen, in *Proc. Int. Conf. on Image Analysis and Recognition (ICIAR'04)* (Porto, 2004), pp. 753–760.
13. J. Van De Weijer and C. Schmid, in *Proc. IEEE Int. Conf. on Image Processing (ICIP'06)* (Atlanta, 2006), pp. 993–996.
14. Y. Xu, H. Ji, and C. Fermüller, Int. J. Comput. Vision **83**, 85 (2009).
15. R. Manthalkar, P. K. Biswas, and B. N. Chatterji, Pattern Recogn. Lett. **24**, 2455 (2003).
16. Q. Xu and Y. Q. Chen, in *Proc. 18th IEEE Int. Conf. on Pattern Recognition (ICPR'06)* (Hong Kong, 2006), Vol. 4, pp. 29–32.
17. E. H. S. Lo, M. R. Pickering, M. R. Frater, and J. F. Arnold, in *Proc. IEEE Int. Conf. on Image Processing, ICIP'04* (Singapore, 2004), vol. 1, pp. 227–230.
18. J. Zhang and T. Tan, Pattern Recogn. Lett. **36**, 657 (2003).
19. X. Tan and B. Triggs, in *Analysis and Modelling of Faces and Gestures* (2007), Vol. 4778 of LNCS, pp. 168–182.
20. R. Maani, S. Kalra, and Y.-H. Yang, Pattern Recogn. **46**, 2103 (2013).
21. R. Kwitt and A. Uhl, in *Proc. 11th IEEE Int. Conf. on Computer Vision (ICCV'07)* (Rio de Janeiro, 2007), pp. 1–8.
22. K. M. Saipullah and D.-H. Kim, Multimedia Tools Appl. **59**, 717 (2012).
23. A. Vécsei, G. Amann, S. Hegenbart, M. Liedlgruber, and A. Uhl, Comput. Biol. Med. **41**, 313 (2011).
24. M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba, in *Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications,* Ed. by D. F. T. P. Exarchos and A. Papadopoulos (IGI Global, Hershey, PA, 2009), pp. 335–350.
25. A. Sousa, M. Dinis-Ribeiro, M. Areia, and M. Coimbra, in *Proc. 16th IEEE Int. Conf. on Image Processing (ICIP'09)* (Cairo, 2009), pp. 681–684.
26. D. J. C. Barbosa, J. Ramos, and C. S. Lima, in *Proc. 30th Annu. Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBS'08)* (Vancouver, 2008), pp. 3012–3015.
27. S. Hegenbart, A. Uhl, and A. Vécsei, in *Proc. 7th Int. Symp. on Image and Signal Processing and Analysis (ISPA'11)* (Dubrovnik, 2011), pp. 715–720.
28. S. R. Fountain and T. Tan, in *Proc. IEEE Int. Conf. on Image Processing (ICIP'97)* (Santa Barbara, 1997), Vol. 3, pp. 197–200.
29. J. Ghosh, in Multiple Classifier Systems, edited by F. Roli and J. Kittler (Springer Berlin Heidelberg, 2002), vol. 2364 of LNCS, pp. 1–15. doi 10.1007/3-540-45428-4_1.10.1007/3-540-45428-4_1
30. G. Giacinto and F. Roli, in Proceedings of the IEEE International Conference on Image Analysis and Processing (ICIAP'99) (1999), pp. 659–664.
31. P. Aljabar, R. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert, in Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'07) (2007), vol. 4791 of LNCS, pp. 523–531,
32. K. Saenko, B. Kulis, M. Fritz, and T. Darrell, in Proceedings of the 11th European Conference on Computer Vision (ECCV'10) (2010), vol. 6314 of LNCS, pp. 213–226.
33. F. Mirrashed and M. Rastegari, in Proceedings of the IEEE International Conference on Computer Vision (ICCV'13) (2013), pp. 2608–2615.
34. S. Hegenbart, A. Uhl, A. V.ecsei, and G. Wimmer, Medical Image Analysis **17**, 458 (2013).
35. G. Babich and O. Camps, IEEE Transactions on Pattern Analysis and Machine Intelligence **18**, 567 (1996).
36. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, Journal of Machine Learning Research **9**, 1871 (2008).
37. M. Gadermayr, S. Hegenbart, and A. Uhl, in Proceedings of 22nd IEEE International Conference on Pattern Recognition (ICPR'14) (2014).
38. M. Gadermayr and A. Uhl, in Proceedings of the IEEE International Conference on Image Processing 2014 (ICIP'14) (2014).
39. M. Gadermayr, A. Uhl, and A. V.ecsei, in Proceedings of the International Conference on Image Analysis and Recognition (ICIAR'14) (2014), vol. 8815 of Springer LNCS, pp. 263–273.
40. K. J. Dana, B. V. Ginneken, S. K. Nayar, and J. J. Koenderink, in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'97) (San Juan, Puerto Rico, 1997), pp. 151–157.
41. B. Caputo, E. Hayman, and P. Mallikarjuna, in Proceedings of the IEEE International Conference on Computer Vision (ICCV'05) (2005), vol. 2, pp. 1597–1604 Vol. 2.

42. S. Lazebnik, C. Schmid, and J. Ponce, Transactions on Pattern Analysis and Machine Intelligence **27**, 1265 (2005).

43. P. Marziliano, F. Dufaux, S. Winkler, T. Ebrahimi, and G. Sa, in Proceedings of the IEEE International Conference on Image Processing (ICIP'02) (2002), pp. 57–60.

44. R. M. Haralick, Dinstein, and K. Shanmugam, IEEE Transactions on Systems, Man, and Cybernetics **3**, 610 (1973).

45. M. Gadermayr, M. Liedlgruber, A. Uhl, and A. V.ecsei, in Proceedings of the 9th International Conference on Image Analysis and Processing (ICIAP'13) (2013), vol. 8156 of Springer LNCS, pp. 513–522.

46. G. Kylberg, External report (Blue series) 35, Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, Uppsala, Sweden (2011), URL http://www.cb.uu.se/~gustaf/texture/.

47. S. Hegenbart, R. Kwitt, M. Liedlgruber, A. Uhl, and A. V.ecsei, in Proceedings of the 6th International Symposium on Image and Signal Processing and Analysis (ISPA'09) (Salzburg, Austria, 2009), pp. 718–723.

48. M. Varma and A. Zisserman, in Proceedings of the 7th European Conference on Computer Vision (ECCV'02) (Springer-Verlag, 2002), pp. 255–271.

49. D. G. Lowe, in Proceedings of the Seventh IEEE International Conference on Computer Vision (CVPR'99) (IEEE, 1999), vol. 2, pp. 1150 – 1157.

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

Michael Gadermayr received his Dipl.Ing. degree (corresponds to M.Sc. degree) from the University of Salzburg, Department of Computer Sciences, Salzburg, Austria in 2012. Between 2012 and 2015, he was a PhD student working as research assistant at the University of Salzburg. In 2015, he received his Dr. techn. degree. Since 2015, he is working as a postdoctoral researcher at the Institute of Imaging and Computer Vision at RWTH Aachen University. Hi is in charge of the Aachen Center for Biomedical Image Analysis, Visualization and Exploration (ACTIVE). His main research interests are in the field of medical image analysis. Specific focus is currently on the analysis of histopathological as well as magnetic resonance image data.

Dorit Merhof received her Diploma in computer science (2003) and her PhD degree in biomedical image analysis (2007) from the University of Erlangen-Nuremberg, Germany. After two years with Siemens Molecular Imaging, Oxford, UK, she joined the University of Konstanz, Germany, as an assistant professor for Visual Computing. Since 2013, she is full professor and head of the Institute of Imaging and Computer Vision at RWTH Aachen University, Germany. Her research interests include acquisition, processing and visualization of image data originating from various biomedical and industrial applications.

Andreas Vécsei received his MD from the University of Vienna, School of Medicine, Austria in 1993 and did postdoctoral work in Trauma Center Salzburg, Austria, the Medical College Hospital Trivandrum, India, the Regional Hospital Hallein, Salzburg and the Regional Hospital Salzburg. He did his residency in pediatrics at the St. Anna Children's Hospital, Vienna, Austria and became a consulting physician in pediatrics in 2002. In 2003 he started his sub-specialization training in pediatric gastroenterology at the Medical University Vienna (Department of Pediatrics and Adolescent Medicine, Department of Gastroenterology) and continued his training at the Dr. v. Hauner Children's Hospital, Division of pediatric Gastroenterology Ludwig Maximillian University Munich, Germany. He was accredited by the German-speaking society for Pediatric Gastroenterology and Nutrition (GPGE) as pediatric gastroenterologist. In 2011 he became Associate Professor of Pediatrics at the Medical University Vienna.

Andreas Uhl is full professor at the Department of Computer Sciences (University of Salzburg, Austria) where he leads the Multimedia Signal Processing and Security Lab. His research interests include image and video processing and compression, wavelets, media security, medical imaging, biometrics, and numbertheoretical numerics.