Degradation Adaptive Texture Classification: An Extended Analysis Leads to a Different Perspective

Michael Gadermayr Andreas Uhl Andreas Vécsei

Technical Report 2015-07 April 2015

Department of Computer Sciences

Jakob-Haringer-Straße 2
5020 Salzburg
Austria
www.cosy.sbg.ac.at

Technical Report Series

Degradation Adaptive Texture Classification: An Extended Analysis Leads to a Different Perspective

Michael Gadermayr, Andreas Uhl and Andreas Vécsei

***Abstract*—Images captured under non-laboratory conditions often significantly suffer from various degradations such as sensor noise and defocus aberrations as well as variations in view point and illumination. 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 the training and the evaluation data set. In this paper, we propose a degradation-adaptive classification approach, which exploits this knowledge and divides one large database into several smaller ones, each containing images with some kind of similarity. In order to get sensible database divisions, we introduce several similarity criteria. In a large set of experiments with several degradations, classifiers and feature extraction methods, we show that our method continuously enhances the classification accuracies in case of simulated as well as real world image degradations. Surprisingly, the framework turns out to be beneficial even in case of idealistic images which are free from strong degradations. Analyzes show that this is due to the fact that the similarity measure performs a kind of pre-classification by changing the prior class probabilities within the generated smaller sub-training sets.**

***Index Terms*—Texture Classification, Invariant Features, Adaptive Classifications, Feature Extraction, Similarity Measures, Fisher Vectors, Local Binary Patterns;**

I. Introduction

For many decades, texture classification [1] – [26] has been a fundamental field in image processing. 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 issue is, how to remove the non-discriminative extrinsic information while maintaining the important discriminative intrinsic information.

We have identified the following common extrinsic properties which are known to affect the classification performances if not being considered:

- Geometric distortions:
  - Affine transformations (rotation, translation, scaling)
  - Perspective transformations
  - Deformations
- Sensor noise
- Blur
- Illumination

One way to deal with such extrinsic information is to develop features that are invariant to a certain property. There is a lot of literature on developing rotation-invariant [10], [16], [17], [18], [19], [22], scale-invariant [10], [16], [17], [18], [19], affine-invariant [20], deformation-invariant [10], view point-invariant [15] and illumination-invariant [21], [17] feature extraction methods. Furthermore, there is also literature on making descriptors invariant (or robust) to noise [22], [23] as well as blur [24], [26]. 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.

A. Classification Scenarios

In this sub-section, the divergent classification scenarios are declared. 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. In the following, this is referred to as standard-scenario which is mostly relevant for real world applications. However, this has to be clarified, because in analysis on invariant feature extraction often a different scenario is chosen which is based on an idealistic training set and an evaluation set with a kind of 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 [27], [28]. In that case the extrinsic properties in the training set and in the evaluation set significantly differ. This could be the case, e.g. if the training set is captured with one camera whereas the evaluation set is captured with another camera. 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

M. Gadermayr is with the Department of Computer Sciences, University of Salzburg, Multimedia Signal Processing and Security Lab (WAVELAB), Salzburg, Austria.

A. Uhl is with the Department of Computer Sciences, University of Salzburg, Multimedia Signal Processing and Security Lab (WAVELAB), Salzburg, Austria.

A. Vécsei is with the St. Anna Children’s Hospital, Department of Pediatrics, Medical University Vienna, Vienna, Austria.

adaptive-classification framework is only applicable sensibly to this scenario.

B. Invariant Feature Extraction Techniques

It seems to be highly beneficial to have descriptors which are invariant to all occurring extrinsic variations within an image database. Especially in case of 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 [29] showed that in case of scale variations, 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. If applying non-invariant features, the classification accuracies significantly drop. 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 [29], [30] showed that the use of state-of-the-art invariant feature extraction techniques in this scenario often leads to lower classification accuracies, compared to other highly discriminative (but non-invariant) features. Obviously, many invariant features seem to be 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 will outline why distinctiveness often is lost if making features invariant to certain degradations. Generally a feature extractor can be interpreted as a function

$$f : \mathbb{R}^{N \times M} \rightarrow \mathbb{R}^{L} \ , \tag{1}$$

where N and M are the image dimensions and L is the feature dimensionality. Theoretically, if a feature is invariant to a certain property, for two images I_1 and I_2 which are similar apart from the respective property, $f(I_1)$ and $f(I_2)$ must be equal! It seems to be highly desirable to have such features, to be able to extract the intrinsic texture properties. However, 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 and an image is captured in two different scales s_1 and s_2. Then $f(I_1)$ must be equal to $f(I_2)$, as already mentioned. But 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 us 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 absolutely scale invariant, but only invariant within specific scale limits, as otherwise the information content of the feature would totally deflate. Similar effects are prevalent in case of other degradations.

Degradation adaptive texture classification focuses on improving the performances of non-invariant feature extraction techniques. However, as it requires similarly distorted images in the training set and the evaluation set, it is only applicable in the standard-scenario.

C. Impact of the Classifier

Especially in the standard-scenario, besides the feature extraction technique, the method used to classify the feature vectors has a major impact on the achieved classification accuracies. In the following we consider two simple but highly intuitive classifiers, namely the k-nearest neighbor classifier and the Parzen-window classifier [31]. In case of the k-nearest neighbor classifier, the choice of the k-value adjusts the degree of "non-linearity" (or flexibility) of the decision boundaries. A similar behavior is exhibited by the Parzen window classifier when varying the kernel variance. To optimize a classifier to a specific problem definition these values can be adjusted. A small k (or a small kernel variance) in general more likely leads to overfitting whereas a larger value more likely leads to underfitting. Again considering the standard-scenario with degradations, we notice that a more "non-linear" decision boundary can help the classifier to fit on the data. Potentially a smaller k-value (or smaller variance) might be necessary in case of degradations. This can be the case, because an image with any kind of degradation has a different corresponding (variant) feature than an image without this degradation. If the classifier is variable (non-linear) enough, to distinguish between the classes even in case of partly degraded images, the overall classification accuracies are supposed to be better compared to an inflexible (linear) classifier. However, this also introduces classifier 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 classifier [32] can be made more flexible by using a non-linear kernel. Neural networks [33] can be adjusted by varying the number of neurons.

D. Focus and Related Work

In this paper, we first focus (in a simulated scenario) on the extrinsic properties scale, blur and noise that are often prevalent in real world images. In spite of their high relevance in practice, many highly discriminative texture features are not invariant (or robust) [34] to these degradations. One area of application which has to cope with these degradations is endoscopy. There has been high effort on computer aided celiac disease diagnosis [30], colonic lesion classification [35], small bowel tumor detection [36] and gastric cancer detection [37]. Due to the downsized sensors and punctual lightnings, noise and low contrast often cannot be prevented. Moreover as the distance to the surface cannot be precisely adjusted, differences in scale are predominant. Furthermore, images can be partly blurry, mostly caused by a wrong distance between the surface and the lens.

In this work, first we investigate a set of texture features with reference to their robustness to the image degradation types, blur, noise and scale variations. We focus on two robustness types. If the classification accuracy 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).

Based on the knowledge that absolute robustness generally is harder to achieve than relative robustness, we consequently propose an adaptive classification framework. By dividing the data set into small, but similarly degraded ones, the necessity of absolutely robust features can be circumvented. In opposite to invariant feature extraction [10], [15], [16], [17], [18], [19], [21], [22] which removes the extrinsic properties, using the proposed framework certain extrinsic properties (hereinafter referred to as degradation measures) are explicitly computed and furthermore exploited within the classification pipeline. The framework can be interpreted in terms of a multiple classifier system [38]. More to the point it is a special case of a classifier selection system [39] as illustrated in the following (see Sect. II-A). The final classifier (which is based on a specific training subset) is selected according to the similarity (or degradation) measure. In recent work [40], classifier selection is utilized in a similar way for a different problem definition. The authors select atlases (as classifiers) with a high degree of similarity for label fusion of brain MRI images. The utilized similarity measures are based on segmentation, image as well as demographic data.

The similarly denoted domain adaptive classification [27], [28] aims at a different classification scenario, which is referred to as domain-change-scenario (see Sect. I-A). As already mentioned, domain adaptation is utilized if the extrinsic properties vary between the training and the evaluation set. On the other hand the proposed degradation adaptive classification can be utilized if the training as well as the evaluation set contain variably degraded images.

After introducing the one-dimensional adaptive texture classification approach, we furthermore generalize it to multiple dimensions. This is done in order to allow the usage of multiple degradation measures for data set division. This is supposed to be beneficial, as image databases often suffer not just from one but even from a couple of different degradations whereas a degradation measure usually captures only one single degradation. In a large experimental setup, for seven image databases, five feature extraction techniques, four degradation measures and two classifiers, the accuracies in case of traditional classification as well as adaptive classification are computed. Furthermore, we investigate the impact of combined (multi-dimensional) degradation measures. Finally, some effects occurring in degradation adaptive texture classification are visualized and extensively analyzed.

E. Contribution

We have been inspired by our previous work on scale adaptive texture classification [41] which is based on another concept but a similar idea. However, the main problem of this previous work is that it is restricted to a classification based on the k-nearest neighbor classifier. In a consecutive work [42], this limitation has been removed, proposing degradation adaptive texture classification. For validation, in this work only few, highly synthetic experiments have been done. Motivated by promising results, another work [43] based on the same idea investigating endoscopic image data has been published later on.

In a significantly larger experimental setup, in the current work we employ degradation measures which are not directly related to the prevalent degradations. Doing that we succeeded in obtaining very promising results. Moreover, 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], [7], [25] are additionally utilized. This has been done to be able to make a more general statement on a larger variety of methods. As previous work [42], [43] showed that the image database has a significant impact on the effect of the method, additionally the degradation adaptive texture classification is applied to new image databases: First of all, three new image databases are investigated [44], [45], [46]. Especially one of them [46] is of high interest, as it contains images of a very good quality and with only insignificant degradations.

Furthermore, the simulated data sets are changed in order to meet a more realistic scenario. Whereas in previous work [42] each of nine degrees of a degradation is applied to each image in a set, in this work we focus on applying one certain (randomly chosen) degree of degradation to an image. This more realistic scenario is supposed to be harder for adaptive classification as it contains significantly fewer training set images. 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. These experiments give us a better insight into the internals of degradation adaptive texture classification.

In Table I a brief comparison (in keywords) to our previous work is provided.

This paper is organized as follows: In Sect. II the degradation adaptive texture classification approach including the utilized degradation measures is described. In Sect. III, 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, Sect. IV concludes this paper.

II. Degradation Adaptive Classification

The basic idea of the degradation adaptive classification is based on the knowledge that absolute robustness generally is harder to achieve than relative robustness (which is shown in Sect. III-B). 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 each object is classified exactly once. The training set in general is not partitioned, but separated by overlapping intervals to prevent from too small training sets.

TABLE I: Overview of the content of this work compared to previous publications [41], [42], [43].

Publication	Scale-Adaptive Class. [41]	Adaptive Classification I [42]	Adaptive Classification II [43]	**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 restriction	k-nearest neighbor	no restriction	no restriction	no restriction
Image descriptors	basic	basic	basic	basic and state-of-the-art
Degradation dimension	1-dimensional	1-dimensional	multi-dimensional	multi-dimensional
Degradation measures	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 evaluated training set ratios obtained class probabilities impact of training set size runtime analysis

A. One-Dimensional Approach

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

$$T_i = \{I \in T : d \leq D(I) \cdot C - i < d + 1\}, \qquad (2)$$

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. 2 as the interval sizes (which is e.g. $[\frac{d}{C}, \frac{d+1}{C})$ 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 (Sect. III). The normalized degradation measure D should be in the interval [0,1) (excluding 1), in order to be real partitioned in case of an overlap of zero. This is done by means of min-max-normalization of the non-normalized degradation measure $\hat{D}$

$$D(I) = \frac{\hat{D}(I) - \hat{D}_{min}}{\hat{D}_{max} - \hat{D}_{min} + \epsilon}, \qquad (3)$$

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 (10^{-6}). This is especially crucial in case of the evaluation set subdivision where one image should be exactly in one evaluation subset.

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

$$E_i = \{I \in E : 0 \leq D'(I) \cdot C - i < 1\}, \qquad (4)$$

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

$$D'(I) = \max(\min(D(I), 1 - \epsilon), 0) \qquad (5)$$

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 classifier generated for T_i. In Fig. 2 pseudo-code of our method is provided.

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 [39] as schematized in Fig. 1. Classifier selection (S_T) is done by means of a degradation measure D, based on the images in the training set. In our case, 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.

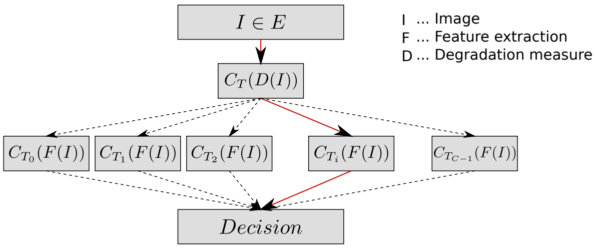


Fig. 1: Schematic visualization of a classifier-selection system.

In recent work [41] 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 usage 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.

```
program degradationAdaptiveClassification
  for i = 0 to C - 1
    Ts(i) = $\{ I \in T : d \leq D(I) \cdot C - i < d + 1 \}$
    Es(i) = $\{ I \in E : 0 \leq D'(I) \cdot C - i < 1 \}$
    model(i) = trainClassifier( Ts(i).imageData, Ts(i).classLabels )
    Es(i).evaluatedClassLabels = evaluate( Es(i).imageData, model(i) )
  end
end program degradationAdaptiveClassification
```

Fig. 2: Pseudo-code of the one-dimensional degradation adaptive classification approach.

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

$$T_{i_1,...,i_n} = \{I \in T : \bigwedge_{j=1}^{n} d_j \leq \pi_j(D(I)) \cdot C_j - i_j < d_j + 1\}. \tag{6}$$

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

$$E_{i_1,...,i_n} = \{I \in E : \bigwedge_{j=1}^{n} 0 \leq \pi_j(D'(I)) \cdot C_j - i_j < 1\}. \tag{7}$$

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

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 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 Sect. III.

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 captured by three of the following degradation measures. Furthermore, we propose a contrast measure which does not extract one of the degradations mentioned above. However, it has proven to be appropriate for adaptive classification in previous work [43].

- **Noise Metric** (D_n):
  The noisy images can be effectively separated from non-noisy ones [42] by computing the total pixelwise sum of 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.
- **Blur Metric** (D_b):
  To measure blur, the metric introduced in [47] is deployed. For computing this rather simple measure, first in the horizontal direction (which has been defined in the publication [47]), the edges are identified by extracting all local minima and maxima for each row. Then we calculate the average ratio between the overall lengths and the magnitudes of the edges. This ratio is directly used 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, the scale-space based method introduced in [41], is utilized. 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. As proposed in the previous paper [41], for the Lapacian-of-Gaussians, the scales $\sigma = \hat{c}\sqrt{2}^k, k \in \{-4, -3.75, ..., 7.75, 8\}$ are chosen (with $\hat{c} = 2.1214$). The pixelwise scales are achieved by using the index of the maximum responses. Finally, the global scale 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 [11] specifying a distinct pixel offset. As successfully used in previous work [43], we utilize an offset of six pixels. Interestingly, contrast is not (only) able to measure degradations, but it has also been used as discriminate feature e.g. in celiac disease diagnosis [48]. 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.
- **Other Measures**:

In previous work [43], further measures such as variance and mean have been investigated. However, as these measures turned out to be less useful than the ones declared in this section, we will not consider them in this paper.

- **Combinations of Metrics** (e.g. D_{bc}):
  Furthermore, we exploit the multi-dimensional adaptive classification framework and investigate combinations of two and three of the declared degradation metrics. In the following, the combination of the 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 necessary 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 II, the execution times for the degradation measures compared to some feature extraction methods (which are specified in Sect. III) 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, in case of adaptive-classification 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).

TABLE II: Execution runtimes for degradation measurement and feature extraction

Method	Runtime	Degradation Measure	Feature Extraction
D_n (Noise measure)	1 ms	✓	
D_b (Blur measure)	28 ms	✓	
D_s (Scale measure)	234 ms	✓	
D_c (Contrast measure)	17 ms	✓	
MRLBP	20 ms		✓
ECM	23 ms		✓
MFS	198 ms		✓
DTCWT	71 ms		✓

III. Experiments

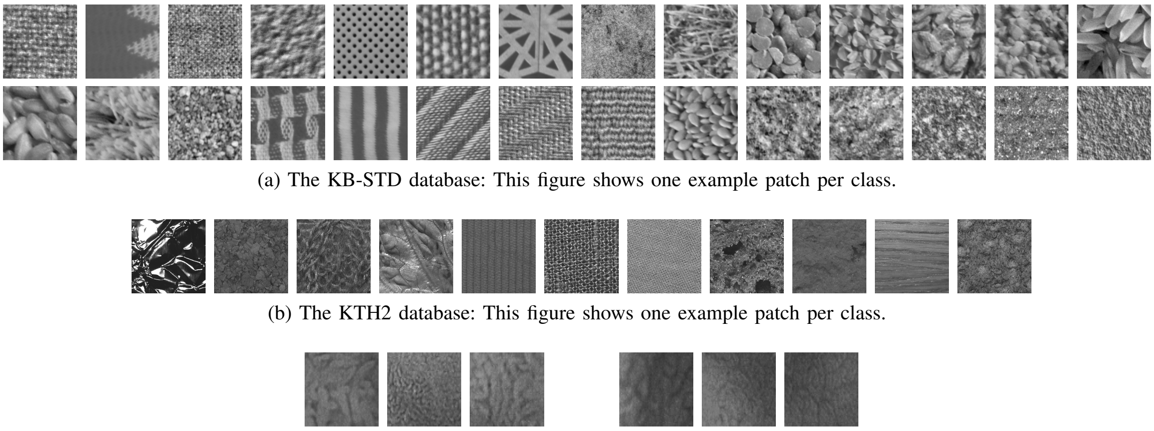
A. Setup

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 [49], which consists of 28 classes and 40 images per class (see Fig. 3a). 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 σ 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 σ being within $\{0, 4, 8, 12, 16, 20, 24, 28, 32\}$. The effect of the different kinds of degradations is shown in Fig. 4. 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 [45] which consists of different textures and real (non-simulated) scale, pose and illumination variations (see Fig. 3b). The CELIAC database [50] (see Fig. 3c) consists of endoscopic images captured during esophagogastroduodenoscopies 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 [51]. The images contain a variety of real degradations. Furthermore our approach is tested with the well known CURET database [44]. 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 [46]. Again the 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 III.

For degradation adaptive texture classification, the training set size must be evaluated for each configuration. As mentioned in Sect. II, we do not fix the overlap d, but instead fix the ratio between the number of images in the original training set and in the new subsets. For this purpose, we defined sensible training set ratios (TR) $\{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}\}$. For example, a TR value of $\frac{1}{4}$ 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

TABLE III: 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



(a) The KB-STD database: This figure shows one example patch per class.

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

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

Fig. 3: Example texture patches of the different databases.

overall training set. The most appropriate value 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 classifiers consisting of a nearest neighbor classifier (NN) and a linear support vector classifier (SVM) [32]. The linear support vector classifier is used, because of its currently high relevance in pattern recognition. The nearest neighbor classifier has been chosen to feature a highly different behavior compared to the SVM. Whereas the SVM corresponds to linear decision boundaries, the nearest neighbor classifier corresponds to highly non-linear decision boundaries. By investigating these two opposing classifiers, we aim in getting more insight into the impact of the classifier (and the decision boundaries) on the error rates, in case of the standard-scenario.

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

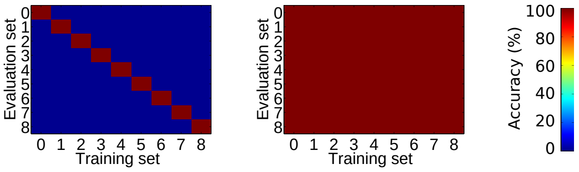
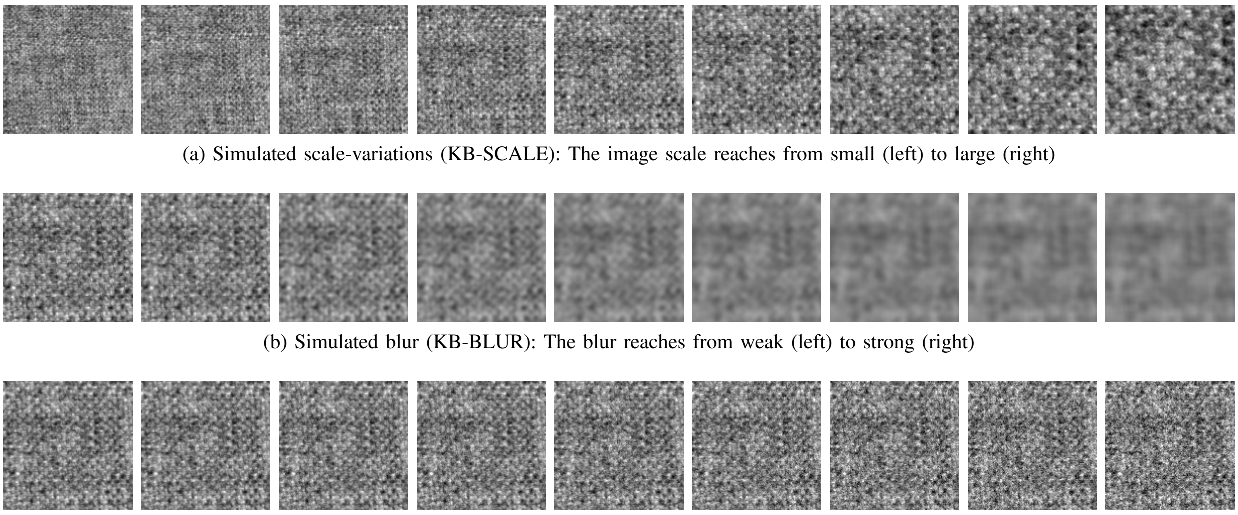


Fig. 5: Visualization of theoretical perfect relative robustness only (left) and perfect absolute robustness (right). These plots are based on fake data, for means of tangibility.

- Multi-Resolution Local Binary Patterns [12] (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 [14] (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 v. For the experiments, t is set to 25% of the maximum response and the displacement vector $v = (1, 1)$ is used.
- Multi-Fractal Spectrum [15] (MFS): The local fractal dimension is computed for each pixel using three different types of measures for computing the local density. The feature vector is built by concatenation of these fractal dimensions.
- Dual-Tree Complex Wavelet Transform [25] (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

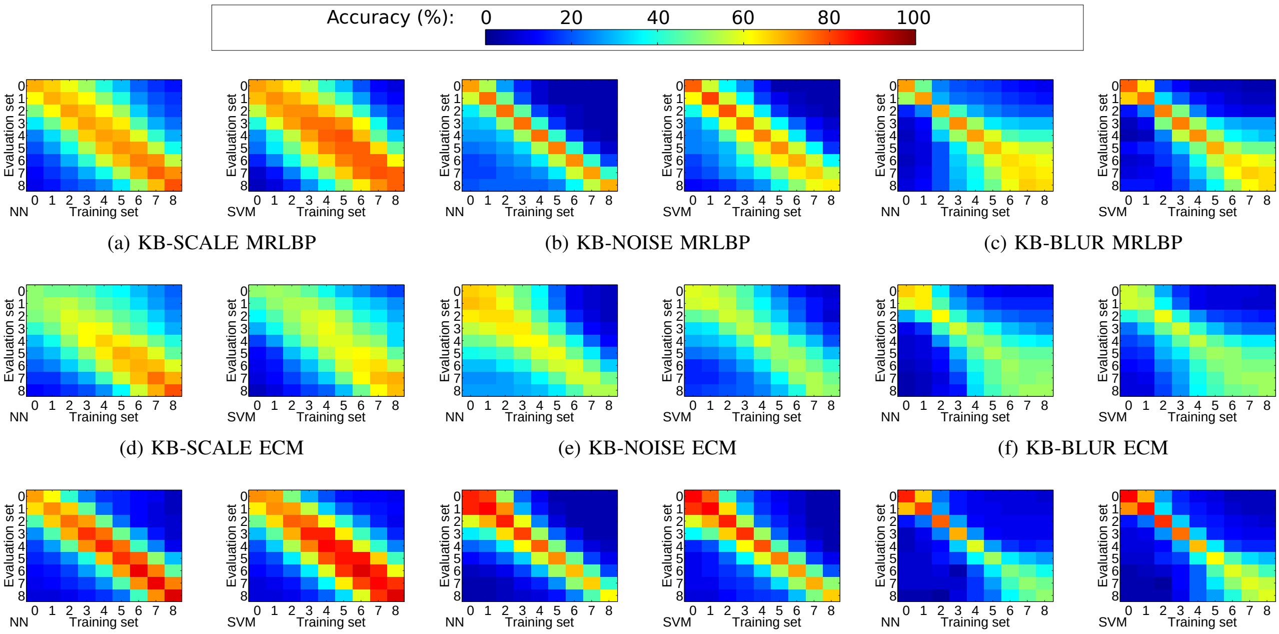


(a) Simulated scale-variations (KB-SCALE): The image scale reaches from small (left) to large (right)

(b) Simulated blur (KB-BLUR): The blur reaches from weak (left) to strong (right)

(c) Simulated noise (KB-NOISE): The noise reaches from weak (left) to strong (right)

Fig. 4: This figure shows the nine strengths of simulated degradations in case of KB-SCALE, KB-BLUR and KB-NOISE and one specific texture patch.



(a) KB-SCALE MRLBP (b) KB-NOISE MRLBP (c) KB-BLUR MRLBP

(d) KB-SCALE ECM (e) KB-NOISE ECM (f) KB-BLUR ECM

(g) KB-SCALE MFS (h) KB-NOISE MFS (i) KB-BLUR MFS

Fig. 6: 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.

complex wavelet transform. Decomposition is performed on five levels.

- 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 [52] 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) [53] 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 [7] 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, which returns a random scalar value between zero and one, independently of the input signal. 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 implementations. In case of IFV and VLAD, we utilize the VLFeat library [54].

B. 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.

As already mentioned in Sect. I-D, 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.

Firstly, Fig. 5 visualizes the theoretic outcomes (classification accuracies indicated by the color) with divergent training and evaluation set degradations with an ideal relatively (but not absolutely) robust image descriptor and an ideal absolutely robust image descriptor. This figure is not based on experimental results but on fake data only, for means of tangibility. In case of the relatively robust descriptor, the classification rates along the diagonal do not drop, whereas if considering the absolutely robust feature, the rates do not drop at all.

In Fig. 6, the robustness of the investigated features with respect to the three degradations scale, noise and blur are visualized. Obviously, 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 achieved accuracies are shown in the diagonal axis in the subplots. 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 training and evaluation set differs, measuring the absolute robustness, the loss 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 shown by noise (and especially the features MRLBP and MFS). Apparently it is hard to achieve absolute robustness to noise, whereas relative robustness seems to be easier achievable. 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 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 classifiers, the ratio between relative and absolute robustness are similar.

Obviously, 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, there is a strong decrease in accuracy. This problem can be compensated using domain adaptation [27], [28]. However, we will focus again on the standard-scenario, with degradations of different strengths in the training and the evaluation set. The large differences between the relative and the absolute robustness shows 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 focus on similarly degraded features [41]. Such a kind of selection definitely cannot be performed by a linear classifier such as the linear support vector classifier.

C. Experiment 2: Adaptive Classification with simulated data

In Fig. 7, the achieved overall classification accuracies of the adaptive classification framework in combination with simulated image degradations are shown (dashed 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 horizontal solid lines indicate the accuracies achieved with traditional instead of degradation adaptive classification. The shorter horizontal dotted lines indicate the best accuracy achieved 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, because the metric in this case is chosen based on the estimation data set to avoid any bias (overfitting).

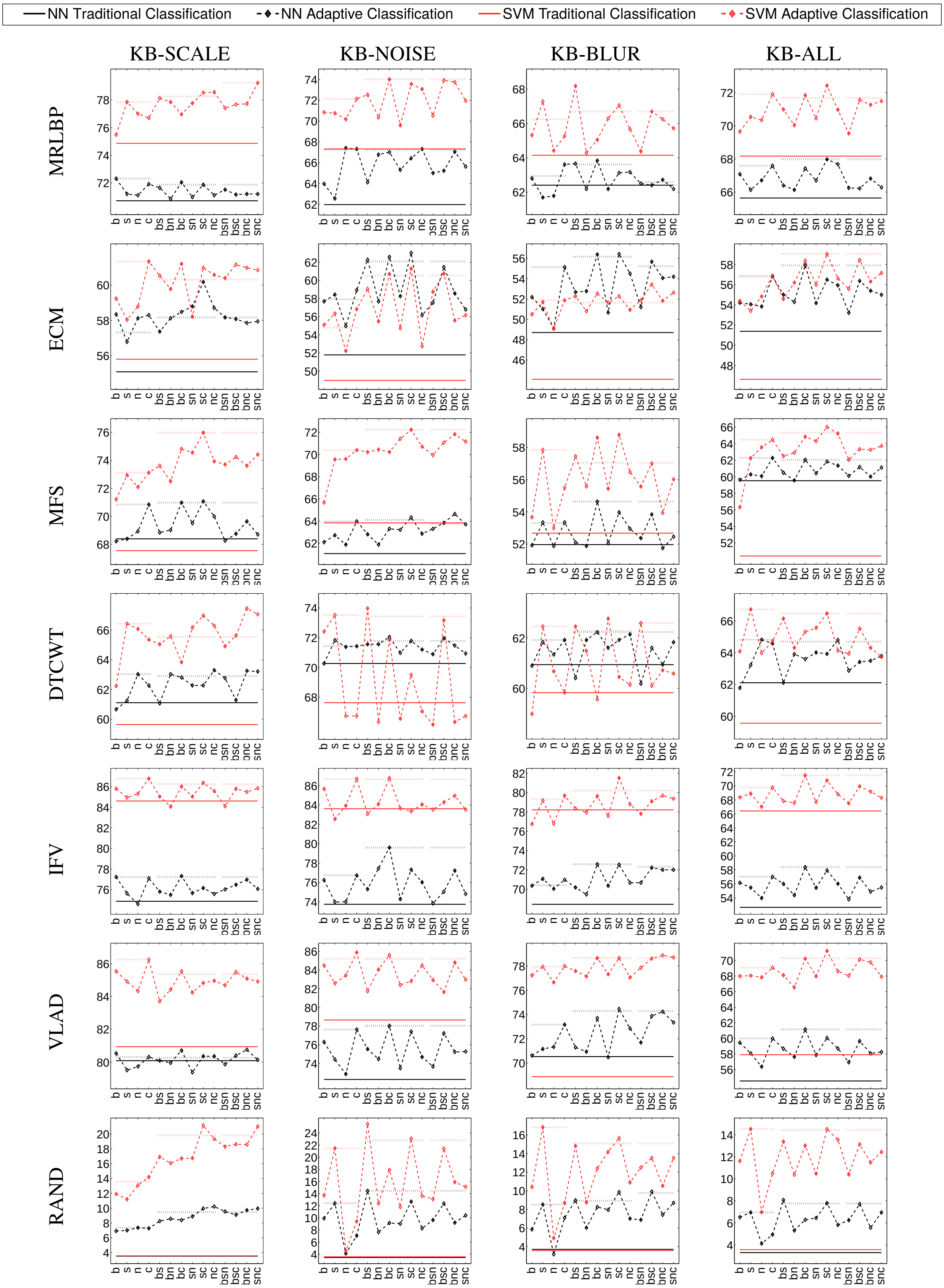


Fig. 7: Accuracies (vertical axis) of adaptive classification in combination with simulated degradations and varying degradation metrics (e.g. bs on the horizontal axis indicates the blur measure D_{bs}). The short horizontal dotted lines indicate the best accuracy based on one-dimensional (left line), at most two-dimensional (center line) and three dimensional metrics (right line).

First, we only consider the one-dimensional (single) degradation measures D_b, D_s, D_n and D_c. We notice that in case of most combinations of a feature, a degradation measure, a database and a classifier, improvements can be obtained. Interestingly, we furthermore notice 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. 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 actually is a quite interesting outcome. As such a behavior has not been supposed, in previous work [42] experiments have been 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 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 [41]. On the other hand, the linear SVM 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. 7), the nearest neighbor classifier is quite competitive compared to the linear support vector classifier. However, in case of adaptive classification this effect is mostly reversed. In case of the higher degree of similarity in the smaller data sets, the NN classifier obviously profits less distinctly from the highly non-linear 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 can be found out if considering the left dotted lines and the center dotted lines. We observe that in the majority of cases (38 out of 56), the utilization of two-dimensional metrics delivers improved classification accuracies. 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, using three dimensional degradation measures (right dotted lines), the accuracies hardly ever can be improved furthermore.

Considering the different feature extraction methods, we observe that generally higher improvements can be achieved with methods with a weaker performance in case of traditional classification. Especially with the (artificial) RAND feature, 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. This effect is particularly investigated in Sect. III-E. However, even with the high performing methods (MFS, VLAD, IFV) distinct improvements are obtained in general. With each database, the highest overall accuracies are obtained using adaptive classification, which is maybe most relevant in practice.

D. 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. Two databases are widely free from any strong image degradations (KB-STD, UIUC). On the other hand, the others suffer from more or less distinct degradations. This is especially the case considering the CELIAC database. An overview above the image databases is given in Table III. Figure 8 shows the achieved classification performances with the real-world databases, similarly presented as in Fig. 7. In case of most configurations, again the measures D_s and D_c seem to be most appropriate in combination with all features. Adaptive classification consistently improves the performances in case of almost each combination of a database and a feature extraction method. The highest overall accuracies for each individual image database is 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 further more. Even with the quite idealistic databases KB-STD and UIUC significant improvements are observed in general. An interesting behavior is shown by the CELIAC database. 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 simple two-classes classification problem. As it has to be distinguished only between two different classes, strong intra-class variations (caused by varying degrees of degradations) could be compensated more easily by the classifier 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. Once 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 seem to be more appropriate in this case.

In Fig. 9 the large amount of data is differently summarized to consider the results from different points of view. Fig. 9a shows the average improvements in case of adaptive classification averaged over all databases. We notice that ECM and VLAD profit most signficantly. A high improvement of ECM has been expected as it is one of the features with the lower accuracies on average and previous work [42] 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. 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 9b, a similar outcome can be observed. Features which

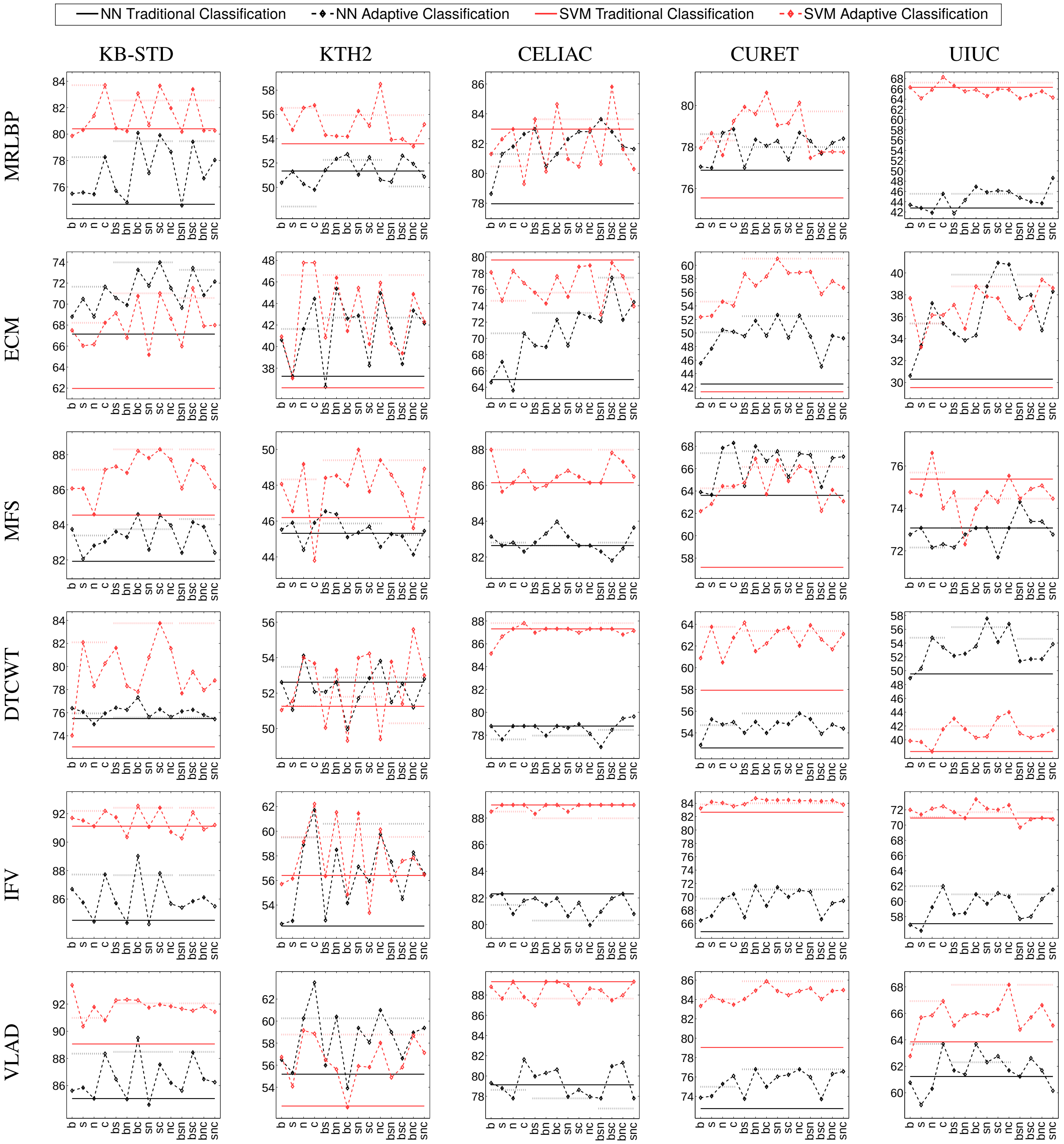


Fig. 8: Accuracies (vertical axis) of degradation adaptive classification in combination with real-world image data, separately for each degradation measure, each database and each feature.

profit more distinctly from adaptive classification in general profit more distinctly from the second dimension. In Fig. 9b and Fig. 9d, 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 accuracy improvement. However, we do 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 because of the more difficult classification problem in case of more classes. Finally in Fig. 9e and Fig. 9f, focus is on the (one-dimensional) degradation measures. If considering the different features (Fig. 9e), it can be seen that there is almost no visible trend. The contrast based metric D_c outperforms all the others in case of each feature extractor. A similar outcome is shown if focusing on the different databases (Fig. 9f). However, in case of one database (CURET), the noise measure D_n outperforms D_c, which is highly inferior considering the other databases. Whereas the feature seemingly does not have a strong impact, the image database has an impact on the choice of the best degradation measure.

E. Experiment 4: Adaptive-Dataset Analysis

In this subsection, effects occurring during adaptive classification and leading to increased classification accuracies are investigated. First, we focus on the chosen overlap threshold d. As already mentioned in Sect. II-B, we do not fix the overlap d, but instead fix the ratio (TR) between the images in the new adaptive training sets and the original training set. For this, the overlap d must be chosen individually, for each individual adaptive training set. In the following, we consider the TR value which is evaluated separately for each configuration based on a separate data set.

In Fig. 10, for each feature (on the horizontal axis), each classifier, each real world data set and each degradation measure, the evaluated ratio TR is shown on the vertical axis. The wide bars in background indicate the average TR value above all databases, separately for each feature, each classifier and each degradation measure. Considering the different features, it can be seen that ECM, which corresponds to the most significant improvements in case of adaptive classification, also attends the smallest training set ratios. However, the effect of the feature extraction technique is much smaller than the effect of the image database. We observe that using the CELIAC or the UIUC database mostly above-average TR values are chosen. In case of the CELIAC database, we suppose a connection between the relatively slight improvements and the large TR value. Based on these data, small TR values do not lead to further improvements and thereby the accuracy benefit is lower compared to other databases. Considering the UIUC database this is supposed to be due to the relatively few training samples (13) which are available per class (see Table III). This leads to very small training data sets especially if small TR values are chosen.

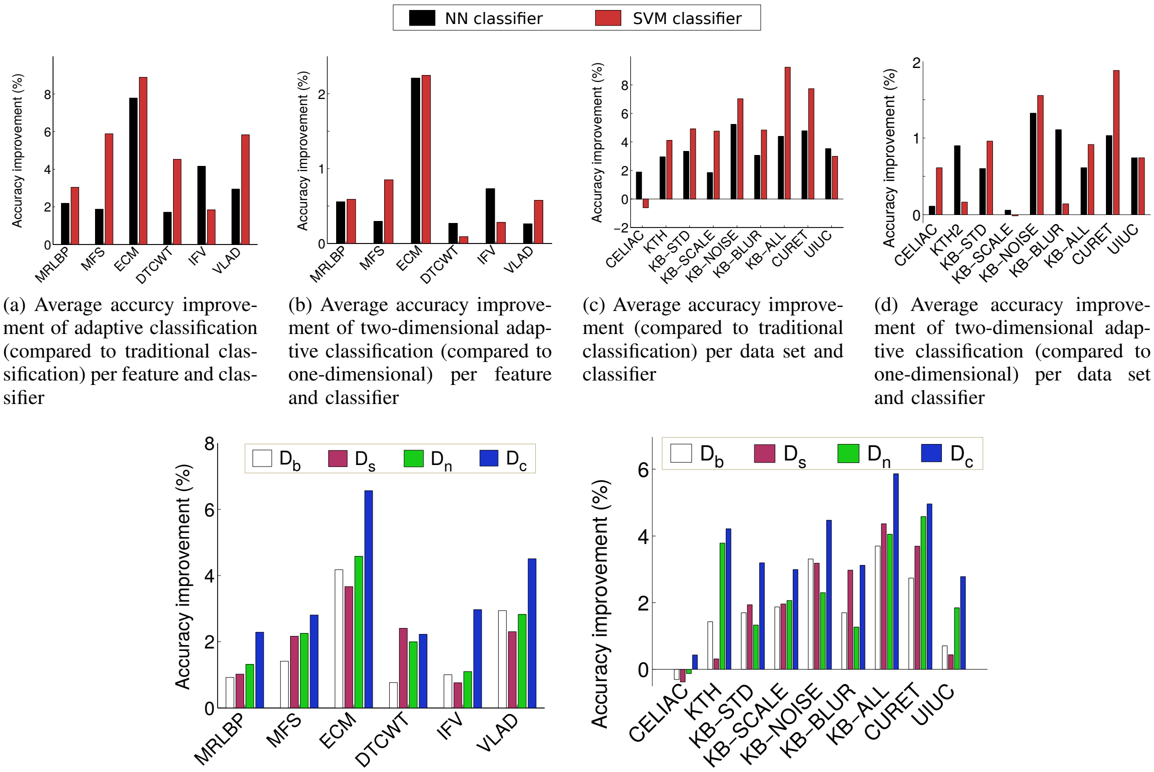
Looking at the two classifiers, the chosen average training set ratio on average is smaller if the nearest neighbor classifier is utilized. Obviously this classifier is able to cope with smaller data sets during adaptive classification than the linear SVM. This is quite interesting as the SVM classifier mostly benefits more distinctly from adaptive classification considering the accuracy improvements. However, this shows that the TR chosen during experimentation does not indicate, how strong the benefit of the degradation adaptive classification is, compared to traditional classification.

From this analysis, we have learned that the most appropriate training set ratio depends on many factors and cannot be determined easily (without a separate optimization data set or cross-validation). Now we exemplarily investigate the impact of varying training set ratios on the overall classification rate. In Fig. 11, for one specific database (KB-STD), the four one-dimensional degradation measures and four different features, the impact of TR is visualized. It can be seen that the curves are quite continuous and do not show a high degree of deceptiveness. Although the best TR values are different, most shown methods reach a accuracy-peak between a TR values of $\frac{1}{4}$ and $\frac{1}{2}$. If considering very small TR values (e.g. $\frac{1}{8}$) the accuracies mostly significantly decrease.

As shown in recent work [43], degradation adaptive classification not only collects similarly degraded images in one database, but also changes the prior class distributions. In the following we investigate the change of the prior distributions in case of one-dimensional degradation measures and the data set without any strong degradations (KB-STD). We utilize the (more or less) idealistic KB-STD database, because we did not expect (but obtained) improvements with these image data using adaptive classification. In Fig. 12, a stacked area chart shows how the originally uniform distribution (right column) is changed by choosing different TR values. Each single area represents the prior probability of a certain class for the variably degraded training sets, indicated by the degradation measure on the horizontal axis. Adjusting d in a way that each training set contains $\frac{1}{2}$ of the overall training set (third column), the uniform distribution is already changed slightly. Especially in regions with very high or very low degradation measures, the prior probabilities are clearly changed. Decreasing the training set size further more to $\frac{1}{4}$ and $\frac{1}{8}$ of the original size, respectively (see left sub figures), an even more significant behavior is obtained. Again the most significant behavior is given by training sets with the highest and the lowest degradations.

F. Experiment 5: Impact of the Training Dataset Size

In recent work [43], it has been assumed that a small original training set size might affect the adaptive classification framework. The even more decreased number of training set images could lead to problems during classification. To investigate this assumption, 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. 13 the accuracies



(a) Average accurcy 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 data set and classifier

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

(e) Average accuracy improvement of adaptive classification (compared to traditional classification) per feature and degradation metric

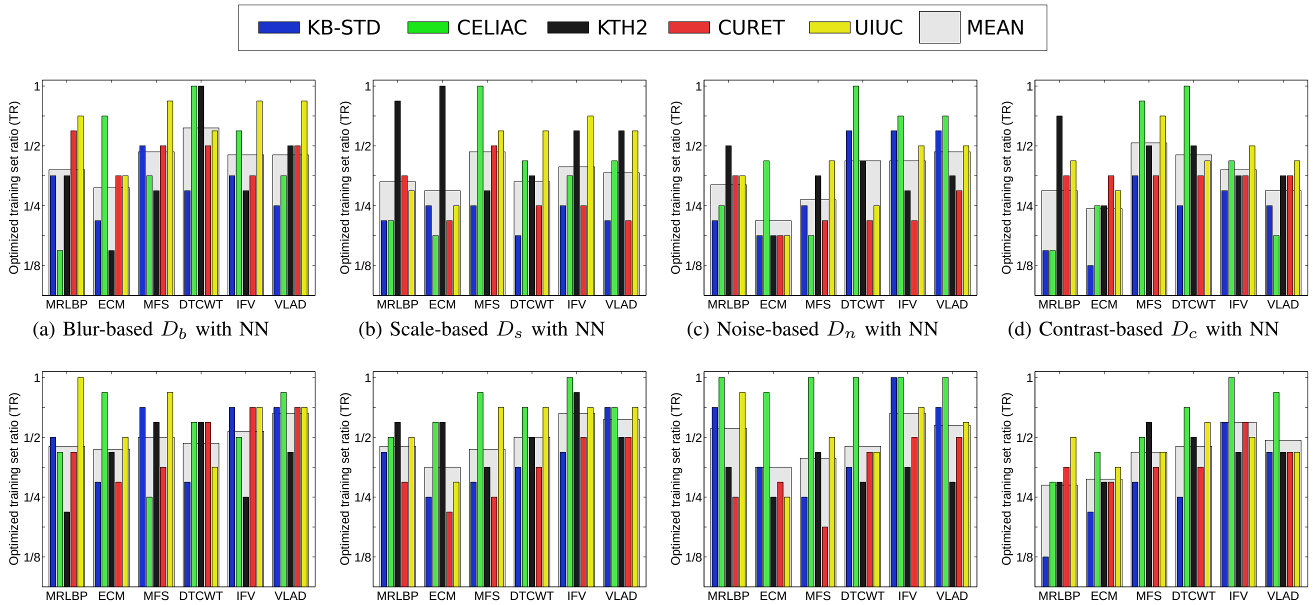
(f) Average accuracy improvement of adaptive classification (compared to traditional classification) per data set and degradation metric

Fig. 9: Overview of the classification accuracy improvements with adaptive classification.

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 features 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 no surprise. However, interestingly this decrease similarly concerns traditional and adaptive classification. As the benefits, as far as accuracy in concerned, mostly do not vanish in case of a reduced training set, the adaptive classification framework can be considered even in case of quite small training sets. In general, 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 the traditional classification and a fallback (if adaptive classification is disadvantageous) is automatically (implicitly) provoked.

IV. Conclusion

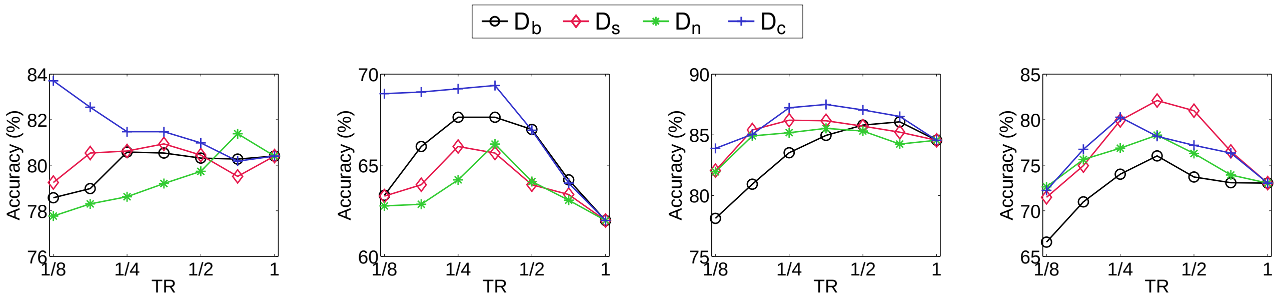
We have shown, that relative robustness to degradations is rather achieved than absolute robustness. Based on this knowledge, we have proposed the degradation adaptive classification framework which exploits this fact in scenarios with variably degraded images in the data sets. Experimentation has shown that the classification accuracy can be improved by our method in combination with all evaluated features and all classifiers in case of simulated image degradations. Surprisingly, enhanced accuracies are not only obtained if the image data is strongly affected by degradations. Even with databases showing no strong degradations, very reasonable improvements are observed. Experiments showed that this is very likely due to the change of the prior probabilities caused by the degradation measures. In such cases, the degradation measure rather acts as a pre-classifying similarity measure, by instantly removing dissimilar images from the respective training data set. Therefore the most appropriate degradation (or similarity) measure cannot be predicted easily, as the effects of the respective measures highly depend on the utilized database, 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 outcome. On average, the contrast based degradation measure generated the most accurate classification results considering one-dimensional measures. We furthermore notice, that the investigated linear classifier benefits more from this new technique compared to the highly non-linear nearest neighbor classifier. Finally it has been proven that even in case of small training data sets, adaptive classification can be effectively used to increase the classification accuracies.



(a) Blur-based D_b with NN (b) Scale-based D_s with NN (c) Noise-based D_n with NN (d) Contrast-based D_c with NN

(e) Blur-based D_b with SVM (f) Scale-based D_s with SVM (g) Noise-based D_n with SVM (h) Contrast-based D_c with SVM

Fig. 10: The evaluated training set ratios TR for each feature, each classifier, each database and each degradation measure. The wide bars in background indicate the average values over all databases.



(a) KB-STD, MRLBP (b) KB-STD, ECM (c) KB-STD, FRACTAL (d) KB-STD, DTCWT

Fig. 11: Each subplot shows the overall classification accuracies with varying TR for a specific feature and one database (KB-STD).

Acknowledgment

This work is partially funded by the Austrian Science Fund (FWF) under Project No. 24366.

Biographies

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. He currently is a Ph.D. student at the University of Salzburg and works as a research assistant. His main research interests are in the field of texture analysis and medical imaging.

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

Andreas Vécsei is medical doctor at the Department of Pediatrics (St. Anna Children's Hospital, Vienna, Austria). Furthermore he is involved in research in gastroenterology at the Medical University Vienna (Austria).

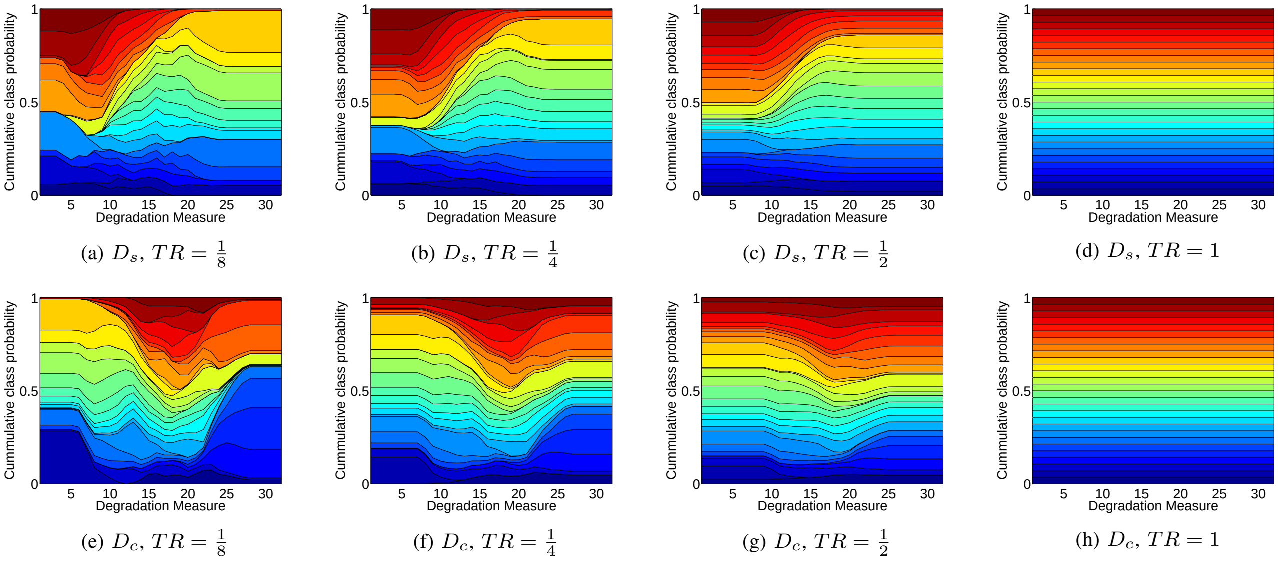
References

[1] M. Cimpoi, S. Maji, I Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*, 2014, pp. 3606–3613.

[2] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision (IJCV)*, vol. 105, no. 3, pp. 222–245, 2013.

[3] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 9, pp. 1704–1716, 2012.

[4] Radu Timofte and Luc J. Van Gool, "A training-free classification framework for textures, writers, and materials," in *Proceedings of the British Machine Vision Conference (BMVC'12)*, 2012, pp. 1–12.



(a) D_s, $TR = \frac{1}{8}$ (b) D_s, $TR = \frac{1}{4}$ (c) D_s, $TR = \frac{1}{2}$ (d) D_s, $TR = 1$

(e) D_c, $TR = \frac{1}{8}$ (f) D_c, $TR = \frac{1}{4}$ (g) D_c, $TR = \frac{1}{2}$ (h) D_c, $TR = 1$

Fig. 12: Stacked area charts showing the prior probabilities for the different degradation measures and image databases. One specific color corresponds to one class of the image data set.

[5] Gaurav Sharma, Sibt ul Hussain, and Frédéric Jurie, “Local higher-order statistics (LHS) for texture categorization and facial analysis,” in *Proceedings of the European Conference on Computer Vision (ECCV’12)*, 2012, vol. 7578 of *Springer LNCS*, pp. 1–12.

[6] F. Perronnin and C. Dance, “Fisher kernels on visual vocabularies for image categorization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’07)*, June 2007, pp. 1–8.

[7] F. Perronnin, Yan Liu, J. Sanchez, and H. Poirier, “Large-scale image retrieval with compressed fisher vectors,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’10)*, June 2010, pp. 3384–3391.

[8] H. Jegou, M. Douze, C. Schmid, and P. Perez, “Aggregating local descriptors into a compact image representation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’10)*, June 2010, pp. 3304–3311.

[9] Li Liu, P. Fieguth, Gangyao Kuang, and Hongbin Zha, “Sorted random projections for robust texture classification,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV’11)*, Nov. 2011, pp. 391–398.

[10] L. Sifre and S. Mallat, “Rotation, scaling and deformation invariant scattering for texture discrimination,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’13)*, June 2013, pp. 1233–1240.

[11] R. M. Haralick, Dinstein, and K. Shanmugam, “Textural features for image classification,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, pp. 610–621, Nov. 1973.

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

[13] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (CVPR’05)*, 2005, vol. 1, pp. 886–893.

[14] R. Rautkorpi and J. Iivarinen, “A novel shape feature for image classification and retrieval,” in *Proceedings of the International Conference on Image Analysis and Recognition (ICIAR’04)*, 2004, vol. 3211 of *LNCS*, pp. 753–760.

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

[16] R. Manthalkar, P. K. Biswas, and B. N. Chatterji, “Rotation and scale invariant texture features using discrete wavelet packet transform,” *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2455–2462, 2003.

[17] Q. Xu and Y. Q. Chen, “Multiscale blob features for gray scale, rotation and spatial scale invariant texture classification,” in *Proceedings of 18th IEEE International Conference on Pattern Recognition (ICPR’06)*, Sept. 2006, vol. 4, pp. 29–32.

[18] S. Cui and Y. Wang, “Redundant wavelet transform in video signal processing,” in *Proceedings of the International Conference on Image Processing, Computer Vision, & Pattern Recognition*, Las Vegas, NV, USA, June 2006.

[19] E. H. S. Lo, M. R. Pickering, M. R. Frater, and J. F. Arnold, “Scale and rotation invariant texture features from the dual-tree complex wavelet transform,” in *Proceedings of the IEEE International Conference on Image Processing, ICIP ’04*, Singapore, Oct. 2004, vol. 1, pp. 227–230.

[20] J. Zhang and T. Tan, “Affine invariant classifiction and retrieval of texture images,” *Pattern Recognition Letters*, vol. 36, no. 3, pp. 657–664, Mar. 2003.

[21] Xiaoyang Tan and Bill Triggs, “Enhanced local texture feature sets for face recognition under difficult lighting conditions,” in *Analysis and Modelling of Faces and Gestures*, Oct. 2007, vol. 4778 of *LNCS*, pp. 168–182.

[22] Rouzbeh Maani, Sanjay Kalra, and Yee-Hong Yang, “Noise robust rotation invariant features for texture classification,” *Pattern Recognition*, vol. 46, no. 8, pp. 2103–2116, 2013.

[23] S. R. Fountain and T.N. Tan, “Extraction of noise robust rotation invariant texture features via multichannel filtering,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP’97)*, Oct. 1997, vol. 3, pp. 197–200.

[24] Joost Van De Weijer and Cordelia Schmid, “Blur robust and color constant image description,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP ’06)*, Atlanta, United States, 2006, pp. 993–996.

[25] R. Kwitt and A. Uhl, “Modeling the marginal distributions of complex wavelet coefficient magnitudes for the classification of zoom-endoscopy images,” in *Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV’07)*, Oct. 2007, pp. 1–8.

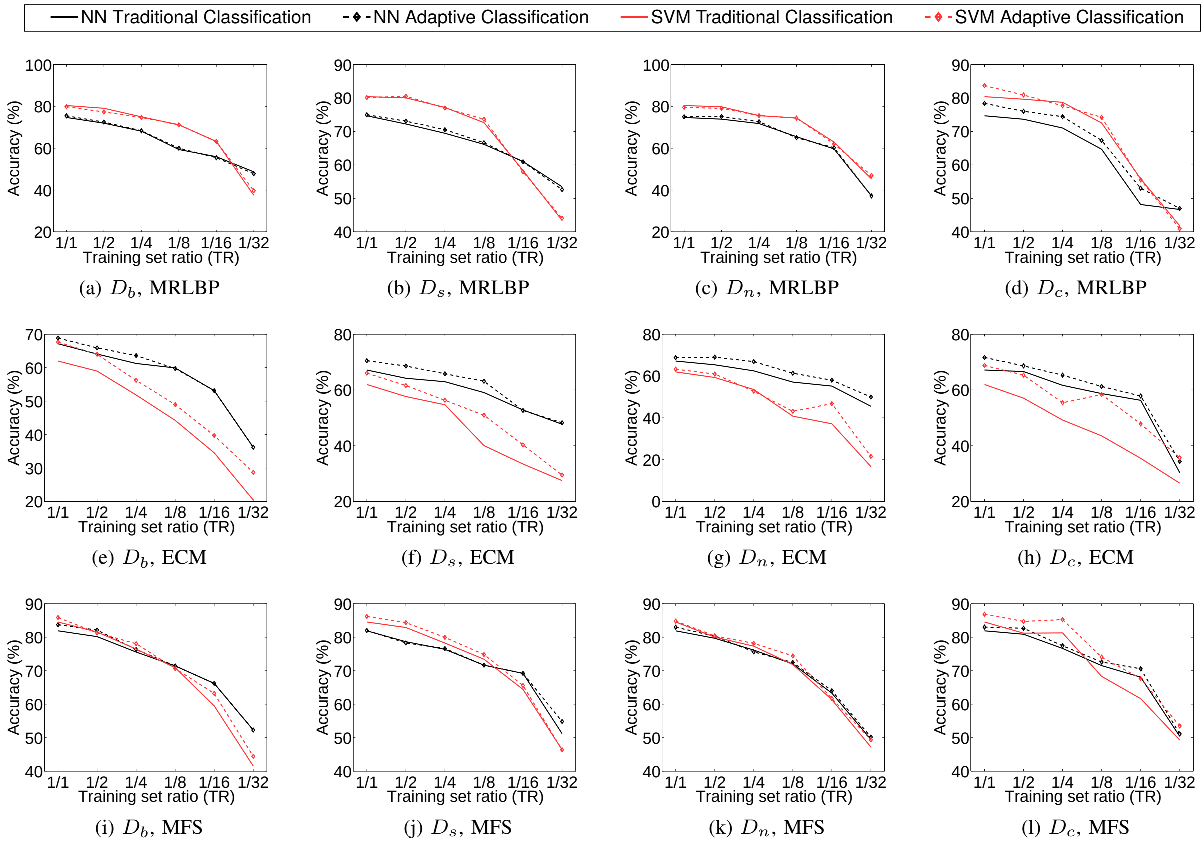
[26] Khairul Muzzammil Saipullah and Deok-Hwan Kim, “A robust texture feature extraction using the localized angular phase,” *Multimedia Tools and Applications*, vol. 59, no. 3, pp. 717–747, 2012.

[27] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell, “Adapting visual category models to new domains,” in *Proceedings of the 11th European Conference on Computer Vision (ECCV’10)*, 2010, vol. 6314 of *LNCS*, pp. 213–226.

[28] Fatemeh Mirrashed and Mohammad Rastegari, “Domain adaptive classification,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV’13)*, 2013, pp. 2608–2615.

[29] Sebastian Hegenbart, Andreas Uhl, Andreas Vécsei, and Georg Wimmer, “Scale invariant texture descriptors for classifying celiac disease,” *Medical Image Analysis*, vol. 17, no. 4, pp. 458 – 474, 2013.

[30] A. Vécsei, G. Amann, S. Hegenbart, M. Liedlgruber, and A. Uhl, “Au-



(a) D_b, MRLBP (b) D_s, MRLBP (c) D_n, MRLBP (d) D_c, MRLBP

(e) D_b, ECM (f) D_s, ECM (g) D_n, ECM (h) D_c, ECM

(i) D_b, MFS (j) D_s, MFS (k) D_n, MFS (l) D_c, MFS

Fig. 13: 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 $\frac{1}{32}$).

tomated marsh-like classification of celiac disease in children using an optimized local texture operator," *Computers in Biology and Medicine*, vol. 41, no. 6, pp. 313–325, June 2011.

[31] G.A. Babich and O.I. Camps, "Weighted parzen windows for pattern classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 5, pp. 567–570, May 1996.

[32] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[33] S.A. Ahmed, S. Dey, and K.K. Sarma, "Image texture classification using artificial neural network (ann)," in *National Conference on Emerging Trends and Applications in Computer Science (NCETACS)*, Mar. 2011, pp. 1–4.

[34] S. Hegenbart, A. Uhl, and A. Vécsei, "Impact of endoscopic image degradations on lbp based features using one-class svm for classification of celiac disease," in *Proceedings of the 7th International Symposium on Image and Signal Processing and Analysis (ISPA'11)*, Dubrovnik, Croatia, Sept. 2011, pp. 715–720.

[35] M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba, "Pit pattern classification using multichannel features and multiclassification," in *Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications*, D.I. Fotiadis T.P. Exarchos, A. Papadopoulos, Ed., pp. 335–350. IGI Global, Hershey, PA, USA, 2009.

[36] Daniel J. C. Barbosa, Jaime Ramos, and Carlos S. Lima, "Detection of small bowel tumors in capsule endoscopy frames using texture analysis based on the discrete wavelet transform," in *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'08)*, 2008, pp. 3012–3015.

[37] A. Sousa, M. Dinis-Ribeiro, M. Areia, and M. Coimbra, "Identifying cancer regions in vital-stained magnification endoscopy images using adapted color histograms," in *Proceedings of the 16th IEEE International Conference on Image Processing (ICIP'09)*, 2009, pp. 681–684.

[38] Joydeep Ghosh, "Multiclassifier systems: Back to the future," in *Multiple Classifier Systems*, Fabio Roli and Josef Kittler, Eds., vol. 2364 of *LNCS*, pp. 1–15. Springer Berlin Heidelberg, 2002.

[39] G. Giacinto and F. Roli, "Methods for dynamic classifier selection," in *Proceedings of the IEEE International Conference on Image Analysis and Processing (ICIAP'99)*, 1999, pp. 659–664.

[40] Paul Aljabar, R. Heckemann, Alexander Hammers, Joseph V. Hajnal, and Daniel Rueckert, "Classifier selection strategies for label fusion using large atlas databases," in *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'07)*, 2007, vol. 4791 of *LNCS*, pp. 523–531.

[41] Michael Gadermayr, Sebastian Hegenbart, and Andreas Uhl, "Scale-adaptive texture classification," in *Proceedings of 22nd IEEE International Conference on Pattern Recognition (ICPR'14)*, Aug. 2014.

[42] Michael Gadermayr and Andreas Uhl, "Degradation adaptive texture classification," in *Proceedings of the IEEE International Conference on Image Processing 2014 (ICIP'14)*, Oct. 2014.

[43] Michael Gadermayr, Andreas Uhl, and Andreas Vécsei, "Degradation adaptive texture classification: A case study in celiac disease diagnosis brings new insight," in *Proceedings of the International Conference on Image Analysis and Recognition (ICIAR'14)*, 2014, vol. 8815 of *Springer LNCS*, pp. 263–273.

[44] K. J. Dana, B. V. Ginneken, S. K. Nayar, and J. J. Koenderink, "Reflectance and texture of real world surfaces," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'97)*, San Juan, Puerto Rico, 1997, pp. 151–157.

[45] B. Caputo, E. Hayman, and P. Mallikarjuna, "Class-specific material categorisation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV'05)*, Oct. 2005, vol. 2, pp. 1597–1604 Vol. 2.

[46] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine region," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1265–1278, 2005.

[47] Pina Marziliano, Frederic Dufaux, Stefan Winkler, Touradj Ebrahimi, and Genimedia Sa, "A no-reference perceptual blur metric," in

Proceedings of the IEEE International Conference on Image Processing (ICIP'02), 2002, pp. 57–60.

[48] Michael Gadermayr, Michael Liedlgruber, Andreas Uhl, and Andreas Vécsei, "Problems in distortion corrected texture classification and the impact of scale and interpolation," in *Proceedings of the 9th International Conference on Image Analysis and Processing (ICIAP'13)*, Sept. 2013, vol. 8156 of *Springer LNCS*, pp. 513–522.

[49] Gustaf Kylberg, "The kylberg texture dataset v. 1.0," External report (Blue series) 35, Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, Uppsala, Sweden, Sept. 2011.

[50] S. Hegenbart, R. Kwitt, M. Liedlgruber, A. Uhl, and A. Vécsei, "Impact of duodenal image capturing techniques and duodenal regions on the performance of automated diagnosis of celiac disease," in *Proceedings of the 6th International Symposium on Image and Signal Processing and Analysis (ISPA'09)*, Salzburg, Austria, Sept. 2009, pp. 718–723.

[51] W. Dickey and D. Hughes, "Prevalence of celiac disease and its endoscopic markers among patients having routine upper gastrointestinal endoscopy," *American Journal of Gastroenterology*, vol. 94, pp. 2182–2186, Aug. 1999.

[52] Manik Varma and Andrew Zisserman, "Classifying images of materials: Achieving viewpoint and illumination independence," in *Proceedings of the 7th European Conference on Computer Vision (ECCV'02)*. 2002, pp. 255–271, Springer-Verlag.

[53] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision (CVPR'99)*. 1999, vol. 2, pp. 1150 – 1157, IEEE.

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