

Degradation Adaptive Texture Classification

Michael Gadermayr

Andreas Uhl

Technical Report 2014-04

February 2014

Department of Computer Sciences

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

Technical Report Series

DEGRADATION ADAPTIVE TEXTURE CLASSIFICATION

Michael Gadermayr, Andreas Uhl

Department of Computer Sciences, University of Salzburg, Austria

ABSTRACT

Image degradations such as noise, blur and scale-variations are known to significantly affect the classification process of textured images. However, due to difficult visual according conditions, such degradation are often prevalent in digital real-world images. We show that these degradations not necessarily strongly affect the discriminative powers of features, in a scenario where similarly degraded images are classified. Contrarily, if the training and the evaluation set contain differently degraded images, the accuracies are decreasing extremely. In this paper, we exploit this knowledge and propose an approach which divides one large database into several smaller ones, each containing similarly degraded images. In order to get sensible database divisions, we use criteria adapted to the respective degradation. In experiments with several degradations, classifiers and feature extraction methods, we show that our method continuously and significantly enhances the classification accuracies.

Index Terms— texture, classification, endoscopy

1. INTRODUCTION

For decades, texture classification has been a fundamental challenge in image processing. On the one hand, feature extraction methods are required to capture image properties, which are significant for texture discrimination. On the other hand, for practical usage it is highly necessary to deal with image degradations which are often prevalent in real world images. In feature extraction, often a balance act between a high discriminative power (to accurately classify in case of idealistic images) and a high invariance to certain degradations has to be performed.

One area of application, where degraded images are prevalent is endoscopy (e.g. celiac disease diagnosis [1], colonic lesion classification [2], small bowel tumor detection [3] and gastric cancer detection [4]). Due to the small sensors and punctual lightnings, noise and low contrast often cannot be circumvented. Moreover as the distance to the surface cannot be precisely adjusted, differences in scale as well as blur (if the surface is out of focus) are potentially predominant.

In this work, first we investigate a set of features with reference to their robustness to three types of image degradations (blur, noise and scale variations). We focus on two robustness types. If the classification accuracy does not strongly

decrease if 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 extend (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 (partitioning) the dataset into smaller, but similarly degraded ones, the necessity of absolute robust features can be circumvented.

In [5], a scale adaptive classification approach has been introduced. However, in this work only degradations due to scale-variations are investigated. Furthermore, the implementation is computationally expensive, as the classifier must be trained separately for each image in the evaluation set. This is why only classifiers with highly lean learning stages (as the k-nearest neighbor classifier) can be efficiently utilized. The framework introduced in this paper overcomes that difficulty. Using other classifiers, we anticipate an even higher advantage of the adaptive classification, as the k-nearest neighbor classifier has proved to be suitable to classify datasets containing variable degradations. In [5] this effect which is referred to as “implicit scale estimation” is investigated.

This paper is structured as follows. In Sect. 2, the proposed degradation adaptive classification approach is introduced. In Sect. 3, the relative and absolute robustness of various features are investigated and the classification improvements achieved with the new approach are presented and discussed. Section 4 finally concludes this paper.

2. 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. Therefore, we divide our datasets into smaller datasets with similar properties.

First for each image I in the training set T , a degradation measure (see Sect. 2.1) D , which measures the level of the respective degradation, is computed and normalized by

$$D_N(I) = \frac{D(I) - D_{min}}{D_{max} - D_{min} + \epsilon}, \quad (1)$$

to be within the interval $[0, 1)$, where D_{min} and D_{max} are the lowest and the highest measure in the training set and ϵ is a small positive constant (we use $\epsilon = 10^{-5}$). After that, the original training set T is divided into the subsets T_i

$$T_i = \{I \in T : d \leq D_N(I) \cdot C - i < d + 1\}. \quad (2)$$

where $i \in \{0, 1, \dots, C - 1\}$ and C defines the cardinality of the set of generated subsets. A large C leads to smaller subsets and consequently higher similarity within one set. If d , which defines the overlap, equals zero, the original dataset 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.

Next, the evaluation set E is partitioned into several subsets E_i as

$$E_i = \{I \in E : 0 \leq D'_N(I) \cdot C - i < 1\}, \quad (3)$$

where

$$D'_N(I) = \max(\min(D_N(I), 1 - \epsilon), 0), \quad (4)$$

in order to ensure that each image in the evaluation set belongs to exactly one training set. Finally for each i , the evaluation set E_i is classified, by the discriminant generated by T_i .

In [5], 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 datasets often is less costly than the training based on one large dataset.

2.1. Degradation Measurement

In order to divide a dataset into several smaller ones with higher similarities, a metric D to capture this similarity is required. In this work, we focus on the following three types of image degradations (and corresponding metrics):

2.1.1. Noise Metric (D_n)

The noisy images can be effectively separated from unnoisy ones 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.

2.1.2. Blur Metric (D_b)

To measure blur, the metric introduced in [6] is deployed. For computing this measure, first in the horizontal direction, the edges are identified by extracting all local minima and maxima for each row. Finally the ratio between the overall lengths and the magnitudes of the edges indicates the blur level.

2.1.3. Scale Metric (D_s)

For scale estimation, the scale-space based method introduced in [5], 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 [5], for the Laplacian-of-Gaussians, the scales $\sigma = \hat{c}\sqrt{2^k}$, $k \in \{-4, -3.75, \dots, 7.75, 8\}$ 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.

3. EXPERIMENTS

3.1. Setup

The experiments are based on the Kylberg texture database [7], consisting of 28 materials with 160 unique texture patches per class, captured at a single scale. The unique textures are divided in 4 separate sets (labeled with A,B,C and D) each containing 40 patches per class (i.e. the total number of images per set is 1120). The image degradations are achieved by simulating blur, noise and scale-variations. In the experiments, these three degradation modes are referred to as mode B (blur), mode N (noise) and mode S (scale-variations). For each degradation, 9 levels are introduced, leading from level 1 (original quality) to level 9 (highest degradations).

Scale-variations are achieved by downscaling the original patches (576×576 pixels) with factors within the set $\{2^{0.00}, 2^{0.25}, \dots, 2^{1.75}, 2^{2.00}\}$. In order to have equal sized patches for all experiments, finally all patches are cropped to a size of 128×128 pixels. Noisy images are computed by adding Gaussian white noise with σ being within $\{0, 4, 8, 12, 16, 20, 24, 28, 32\}$. The blurred images are simulated by applying a Gaussian filter with σ being within $\{0.0, 0.5, \dots, 3.5, 4.0\}$.

For final classification, we deploy three different classifiers consisting of the nearest neighbor classifier (NN), the k-nearest neighbor classifier (KNN) and a linear (Bayes normal) classifier (LIN) [8]. The k value of the KNN classifier is chosen to be one forth of the dataset size, divided by the class count. We chose these 3 classifiers because of their traceability as well as their dissimilarities. On the one hand, in case of the linear classifier, the decision boundary is restricted to a hyperplane. On the other hand, the nearest neighbor classifier has highly complex decision boundaries given by the Voronoi diagram. The KNN classifier lies inbetween the others.

In order to decide whether a classification method is significantly more accurate than another one, we deploy McNemar's test [9] (with $\alpha = 0.01$).

For feature extraction, 4 significantly different well known techniques are deployed:

- **Local Binary Patterns [10] (LBP):**
LBP describes a texture by utilizing the joint distribution of pixel intensity differences represented by binary patterns. We deploy the uniform version (capturing only patterns with at most 2 bitwise transitions) using the standard 8-neighborhood with a radius of 1 pixel.
- **Multi-Fractal Spectrum [11] (FRA):**
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.
- **Histogram of Gradients [12] (HOG):**
The distribution of gradient orientations is used to describe a texture. This feature is used with the standard bin count of 9, which corresponds to an angular resolution of 20° .
- **Edge Co-occurrence Matrix [13] (ECM):**
After applying eight differently orientated directional filters, the orientation (maximal response) 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.

In a first experiment we investigate the relative as well as the absolute robustness of the investigated features (in case of the nearest neighbor classifier) with reference to one specific degradation. Therefore one of the four Kylberg datasets (A) is used for training and another one (set B) is used for evaluation. All combinations with degradations 1 to 9 in the training set and in the evaluation set are examined.

The main experiment is based on training and evaluation sets, each containing a distinct degradation but all levels of this degradation. Dataset A (which contains now 1120×9 images), is used for training and dataset B (with the same size) is used for evaluation. The same is done for datasets C and D. In order to optimize the parameters of the adaptive classification, the rates for all combinations of C and d must be processed. The parameter space for our experiments is fixed to P :

$$P = \{(C, d) \mid C \in \{1 + 4 \cdot i \mid i \in \mathbb{N}, i \leq 8\} \wedge d \in \{0.02 \cdot i \mid i \in \mathbb{N}_0, i \leq 5\}\}. \quad (5)$$

Finally, the classification accuracies for both experiments are achieved by used the indexes of the maxima over all accuracies (varying C and d) of the opposite experiment. Using the accuracies with the gathered indexes, an overfitting is avoided. For the analyses, the accuracies of these two experiments are averaged and considered as one measure.

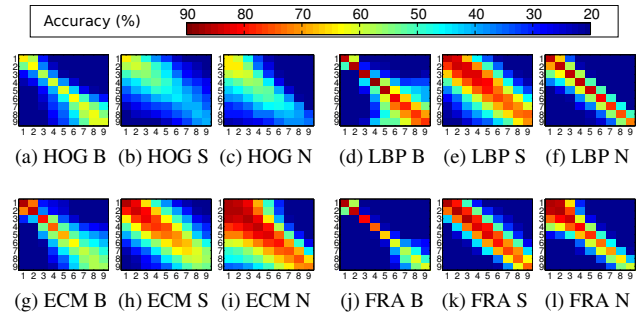


Fig. 1: Achieved classification accuracies with varying levels of degradation in training (horizontal axis) and evaluation set (vertical axis).

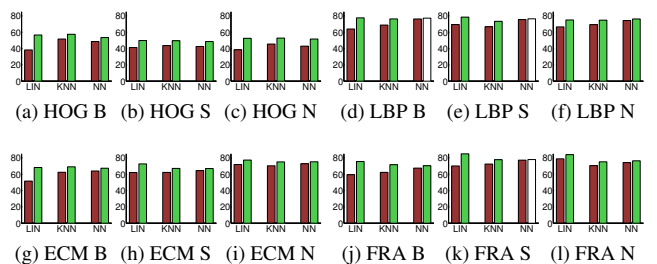


Fig. 2: Comparison of classification accuracies with traditional (left bars) and adaptive classification (right bars), in case of combining all (9) degradation classes within the training and evaluation datasets.

3.2. Results Robustness-Analysis

In Fig. 1, the robustnesses of the investigated features with respect to the three degradations are visualized. Obviously, if the training and the evaluation dataset continuously suffer from similar degradations (referred to as the relative robustness), the accuracy only moderately decreases in most combinations of features and modes (see diagonal axis in the subplots). An extraordinarily high relative robustness is especially achieved with HOG and Mode B. 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. A very distinct behavior is shown by FRA and mode B. We notice that in case of mode B, features generally show high relative robustnesses, but low absolute robustnesses. Obviously blurred images do not strongly affect the accuracy in case of a high blur-similarity, however, a lower similarity leads to significant decreases.

3.3. Results Adaptive Classification

Knowing that a relative robustness can be achieved much more easily than absolute one, we now investigate the effect of the degradation adaptive classification approach proposed in this paper. In Fig. 2, the accuracy improvements for specific features, modes and classifiers are shown if all 9 degradation levels are included in the datasets. Whereas the

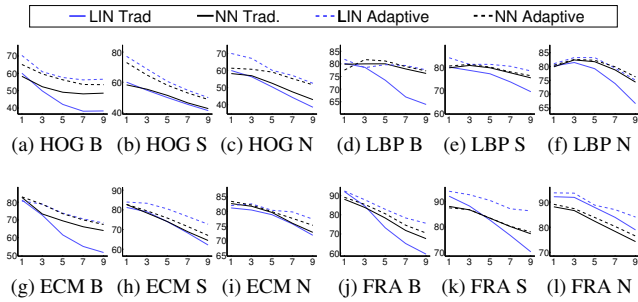


Fig. 3: Classification rates with varying numbers of degradation levels on the horizontal axis. To increase visibility, only the two extreme case classifiers (LIN and NN) are depicted.

left bars show the accuracies achieved with traditional classification, the right bars indicate the rates obtained with the proposed method. If the right bar is green-colored, a significant improvement is observed. We notice that in case of all combinations, improvements are achieved. Whereas these improvements are either small in case of the NN classifier, especially with the LIN classifier high improvements are recognized. Apart from three results with the NN classifier (see white-colored bars), all improvements are significant.

For a more detailed view, in Fig. 3, the impact of varying numbers of degradation levels is shown. Whereas the leftmost values indicate the accuracies achieved with idealistic images (without the simulated degradations), the rightmost values indicate the rates as presented in Fig 2. We notice that the adaptive classification not only profits in case of a high number of degradation levels. Even without any degradation in the images, the proposed method often delivers improved classification accuracies. Another interesting aspect is, that the LIN’s accuracy mostly (in 11 out of 12 experiments) is below the NN’s rate in case of the traditional classification and the strongest degradations. Surprisingly, in case of the new adaptive method in 11 out of 12 cases the LIN classifier outperforms the NN classifier.

Fig. 4 summarizes the results to outline major prominences. In Fig. 4a, we see that the LIN classifier corresponds to the highest average improvements followed by the KNN and the NN classifier. This is especially true in case of a large number of degradation levels. In [5], the authors showed, that the nearest neighbor classifier is implicitly able to compensate scale variations if prevalent in the training and the evaluation set. Due to the evaluated linear decision boundary, the LIN classifier obviously is not able to effectively compensate these variations. Consequently it profits more significantly from the proposed adaptive classification. The KNN classifier is between the others. In Fig. 4b, the rates are summarized, to investigate the three different modes. In case of modes B and N, the improvement is increasing with a larger number of degradation levels. Interestingly, mode S even starts with a high improvement in case of the undegraded datasets. Although this seems to be illogical, there is an explanation

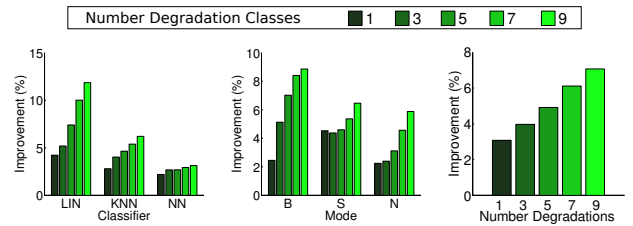


Fig. 4: Summarization of improvements with adaptive classification by averaging the classification accuracies. (a) All features and modes are averaged, to investigate the behavior of the accuracy with respect to the used classifiers and the number of degradation classes. (b) All features and classifiers are averaged, to investigate the behavior of the accuracy with respect to the modes and the number of degradation classes. (c) All features, classifiers and modes are averaged, to investigate the behavior of the accuracy with respect to the number of degradation classes.

Fig. 4: Summarization of improvements with adaptive classification by averaging the classification accuracies.

for this behavior. The adaptive classification method divides the images to similar sets. Even though the images do not suffer from simulated degradation, due to the differing intrinsic scales of textures, scale variations are prevalent. Thereby within one database, occurrences of images with a high intrinsic scale difference are reduced and consequently wrong decisions due to inhomogeneous scales can be circumvented. In case of mode B and mode N a similar but less distinct behavior is observed (if considering the leftmost bars).

In Fig. 4c, the rates are summarized, to investigate the impact of the number of degradation levels. As anticipated, the benefit improves steadily with increasing levels. Once again, we notice that the improvement (on average) does not vanish if considering images without any simulated degradations.

In order to ensure that the accuracy improvements are not only caused by the decreased sizes of the training sets, we finally evaluated the effect of randomly reduced training sets. However, in doing so we did not achieve any significant improvements. Quite the contrary, (as expected) if the sets are heavily reduced, the accuracies are continuously decreasing.

4. CONCLUSION

We have shown, that relative robustness to degradations is rather achieved than an absolute robustness. We have proposed a degradation adaptive classification framework to exploit the achieved knowledge in scenarios with variably degraded images in the datasets. Experimentation has shown that the classification accuracy can be statistically significantly improved by our method in case of all evaluated features and all classifiers and all simulated types of degradations. Surprisingly, enhanced accuracies are not only achieved in case of large degradation differences. Even without any simulated degradation, the classification rate on average can be improved. We especially notice, that the linear classifier stronger benefits from this new technique compared to the highly non-linear nearest neighbor classifier.

5. REFERENCES

- [1] A. Vécsei, G. Amann, S. Hegenbart, M. Liedlgruber, and A. Uhl, "Automated marsh-like classification of celiac disease in children using an optimized local texture operator," *Computers in Biology and Medicine*, vol. 41, no. 6, pp. 313–325, June 2011.
- [2] 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.
- [3] D. J. C. Barbosa, J. Ramos, and C. 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.
- [4] 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 International Conference on Image Processing (ICIP'09)*, 2009, pp. 681–684.
- [5] M. Gadermayr, S. Hegenbart, and A. Uhl, "Scale-adaptive texture classification," Tech. Rep. 2014-01, Department of Computer Sciences, University of Salzburg, Austria, 2014, <http://www.cosy.sbg.ac.at/research>, submitted to ICPR 2014.
- [6] P. Marziliano, F. Dufaux, S. Winkler, T. Ebrahimi, and G. Sa, "A no-reference perceptual blur metric," in *IEEE 2002 International Conference on Image Processing*, Rochester, New York, 2002, pp. 57–60.
- [7] G. 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, September 2011.
- [8] R. P. W. Duin, P. Juszczak, P. Paclík, E. Pekalska, D. de Ridder, D. M. J. Tax, and S. Verzakov, "PR-Tools4.1, a matlab toolbox for pattern recognition," 2007, <http://prtools.org>.
- [9] Q. McNemar, "Note on the sampling error of the difference between correlated proportions of percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, June 1947.
- [10] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, January 1996.
- [11] Y. Xu, H. Ji, and C. Fermüller, "Viewpoint invariant texture description using fractal analysis," *International Journal of Computer Vision*, vol. 83, no. 1, pp. 85–100, 2009.
- [12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, vol. 1, pp. 886–893.
- [13] 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)*, 2004, pp. 753–760.