International Journal of Semantic Computing
Vol. 8, No. 2 (2014) 145–167
©World Scientific Publishing Company
DOI: 10.1142/S1793351X14400029

World Scientific
www.worldscientific.com

Beyond Standard Noise Models: Evaluating Denoising Algorithms with Respect to Realistic Camera Noise

Tamara Seybold
Arnold & Richter Cine Technik, Türkenstraße 89
80799 München, Germany
tseybold@arri.de

Marion Knopp, Christian Keimel and Walter Stechele
Technische Universität München, Arcisstraße 21
80333 München, Germany

The development and tuning of denoising algorithms is usually based on readily processed test images that are artificially degraded with additive white Gaussian noise (AWGN). While AWGN allows us to easily generate test data in a repeatable manner, it does not reflect the noise characteristics in a real digital camera. Realistic camera noise is signal-dependent and spatially correlated due to the demosaicking step required to obtain full-color images. Hence, the noise characteristic is fundamentally different from AWGN. Using such unrealistic data to test, optimize and compare denoising algorithms may lead to incorrect parameter tuning or suboptimal choices in research on denoising algorithms.

In this paper, we therefore propose an approach to evaluate denoising algorithms with respect to realistic camera noise: we describe a new camera noise model that includes the full processing chain of a single sensor camera. We determine the visual quality of noisy and denoised test sequences using a subjective test with 18 participants. We show that the noise characteristics have a significant effect on visual quality. Quality metrics, which are required to compare denoising results, are applied, and we first evaluate the performance of 12 full-reference metrics. As no-reference metrics are especially useful for parameter tuning, we additionally evaluate five no-reference metrics with our realistic test data. We conclude that a more realistic noise model should be used in future research to improve the quality estimation of digital images and videos and to improve the research on denoising algorithms.

Keywords: Denoising; camera noise; quality assessment; quality metrics; no-reference metrics.

1. Introduction

The demand for ever higher resolution has driven an increase in pixel count, resulting in a lower pixel pitch (size of each pixel on the sensor). Thus, the amount of light trapped by a single pixel is lower, and the signal-to-noise ratio decreases. This is especially severe under low light conditions. Hence, algorithmic methods to reduce the noise are key for applications ranging from professional movie shots to smart phone recordings. Denoising has been studied extensively and various methods have been developed [1– 6].

The development and tuning of these algorithms is typically based upon standard test datasets like the Kodak image set [7]. These datasets include a collection of representative reference images. To evaluate denoising algorithms, the reference images are degraded using artificial noise, to obtain pairs of a reference and a noisy image. In the simplest case, the results of the denoising algorithms are subsequently evaluated using the difference to the reference image (PSNR). As this measure does not correlate well with human perception of visual quality, several more sophisticated quality metrics have been proposed [8–15].

Typically, the mentioned noisy images are generated by applying additive white Gaussian noise (AWGN) to the reference images. While AWGN allows us to easily generate noisy images in a repeatable manner, it does not reflect the properties of realistic camera noise. In [16] it is shown that noise in the raw sensor data is signal dependent. Three major steps, shown in Fig. 1, are required to transform the raw sensor data to an image that can be viewed on a display device (display-domain image): The first step is the white balance. Since the sensor data only provides one color value per pixel (Bayer mask) demosaicking is required, as a second step, to obtain a full-color image. In a third step, a color transformation is applied to transform the image into the monitor color space. Previous work showed that the second step, the demosaicking, has a significant effect on denoising results using the Kodak data set, as it introduces a spatial correlation in the noise characteristics [17]. Further, the color transformation changes the noise distribution in a nonlinear manner. Hence, the realistic noise characteristic in the camera data is fundamentally different from AWGN.

Thus, the complete processing pipeline shown in Fig. 1 must be considered to generate test images with realistic camera noise. Some recent denoising methods implement more realistic camera noise models, e.g. [18–20]. To evaluate these algorithms representative result images based on real camera data are shown. To quantitatively compare their results simple models as AWGN are used. When denoising algorithms are evaluated and optimized using unrealistic test data, this, however, may lead to wrong parameter tuning or suboptimal choices in research and development of denoising algorithms.

In this paper, we propose an approach to evaluate denoising algorithms with respect to realistic camera noise. Following our first results in [21], we discuss the individual steps of our approach in detail and extend our work to blind evaluation of denoising results.

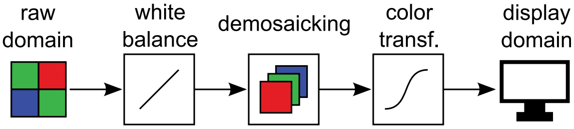


Fig. 1. Image processing pipeline of a single-sensor camera.

We first describe a new camera noise model that includes the complete processing chain. Further, two state-of-the-art denoising algorithms are evaluated with respect to their denoising results on test images with AWGN and images generated using our new realistic camera noise model. To evaluate the visual quality of the denoised images a reliable quality metric is required. Up until now, quality metrics have not been tested with respect to realistic camera noise. Thus, to determine the visual quality of the test images in a reliable way, a subjective test with human observers is performed. Further, the results of the subjective test are compared to a large set of existing quality metrics. We analyze in detail how the individual processing steps influence the performance of these metrics. This allows us to identify the most suitable metrics to evaluate denoising algorithms with respect to realistic camera noise.

Blind quality evaluation, i.e. rating the visual quality without a reference image, is crucial for the automatic parameter tuning of denoising algorithms. Furthermore, these metrics would enable us to directly use test data without reference. Thus, real camera data could be used for denoising evaluation directly. So-called no-reference metrics (NR metrics) are therefore of high interest for the evaluation and parameter tuning of denoising algorithms. As recent research results on NR metrics are very promising [22–25], we especially discuss blind image quality evaluation in this paper.

The remainder of this paper is organized as follows: First, we discuss the camera noise characteristics and describe the processing steps that influence the noise characteristics in Sec. 2. The test setup for the subsequent experiments including the subjective test is outlined in Sec. 3. We discuss the subjective quality results and we compare the performance of full-reference and no-reference quality metrics with respect to realistic camera noise in Sec. 4. In Sec. 5, two state-of-the-art denoising algorithms are evaluated using test images with AWGN and test images with the new realistic camera noise model and quality metrics are compared using the Spearman correlation over all the test sequences. Finally, we conclude in Sec. 6.

2. Camera Noise

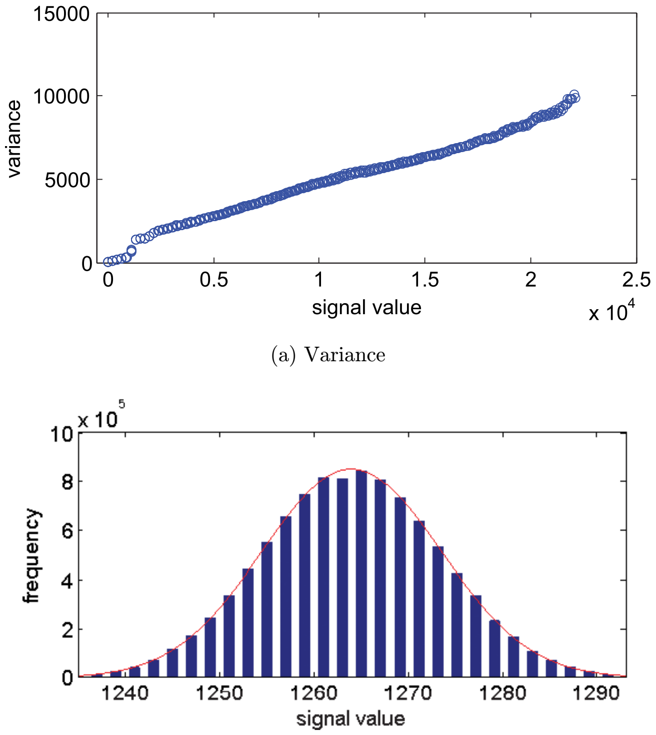
For a realistic denoising evaluation we need a realistic model for the camera noise. To find this model, we first measure the real camera noise in the raw domain. We then discuss the influence of the camera processing pipeline.

2.1. *Camera noise in the raw data*

To measure camera noise in the raw images we take a series of exposures and calculate the noise variance using the photon transfer method [26]. While this measurement can be performed with any camera, we use the ARRI Alexa camera, as it delivers uncompressed raw data in 16 bit precision. Since the data is uncompressed, we can expect unaltered measurement results. Furthermore, the individual camera processing steps are known for this camera [27]. Our method can equivalently be used for other cameras.

The Alexa camera has been developed for motion picture recordings in digital cinema applications. It has a CMOS sensor with a resolution of 2880 × 1620. In front of the sensor, the camera has a filter pack composed of an infrared cut-off filter, an ultraviolet cut-off filter and a low pass filter to reduce aliasing. The color filter array (CFA), which is located between the filter pack and the sensor, is a Bayer mask.

The photon transfer method [26], uses two subsequent frames recorded at constant and homogenous lighting conditions. The noise variance is calculated as the mean of the difference between these two frames, the corresponding signal value is calculated as the mean over all the signal values in these frames. The graph in Fig. 2(a) shows the variance plotted over the respective mean signal value. The variance of the sensor noise can be approximated by a linear model. This finding matches the results reported in [16] where other cameras have been studied. We see, however, one difference in the region around signal value 0.1×10^4. The step in the variance curve is due to a special characteristic of the Alexa sensor, the dual-gain read-out technology. This means, the sensor read-out of the Alexa provides two different paths with different amplification (dual-gain read-out). The low amplified path provides the data for the signal range starting from 0.1×10^4. The high



(a) Variance

(b) Distribution

Fig. 2. Variance and distribution of the noise in the raw domain (signal values in 16 bit precision).

amplified path saturates in the high signal values, but for the low signal values it provides a significantly higher signal-to-noise ratio. The read-out noise (offset of the variance curve) is reduced, thus the dual-gain technology enhances the low light performance of the camera. The two read-out paths are combined in the region around signal value 0.1×10^4, which explains the step in the variance curve.

The distribution is very similar to a Gaussian distribution. In Fig. 2(b) the distribution at signal level 1265 is shown with the Gaussian approximation. That means we can well approximate the sensor noise n in the raw domain using a Gaussian distribution with signal-dependent variance.

$$n \sim \mathcal{N}(0, \sigma(x)) \quad \text{with } \sigma^2(x) = m(x)x + t(x). \tag{1}$$

The variance $\sigma^2(x)$ is approximated as a piecewise linear function depending on the signal x, with the slope $m(x)$ and the intercept $t(x)$ based on the measurement data in Fig. 2(a). Because of the dual-gain read-out the values for $m(x)$ and $t(x)$ are piecewise constant.

Thus we found a model for the camera noise in the raw data. The signal values in the raw data are in linear domain, that means the signal value is proportional to the amount of light collected by the sensor and the raw data provides only one color value per pixel according to the Bayer pattern. Therefore the data must be further processed to obtain a full-color display-domain image that can be used for testing.

2.2. *Camera noise in the processing pipeline*

In the previous section we presented a realistic model for the camera noise in the raw data. As the quality estimation of test images requires images in the display domain, we need to consider all the processing steps to obtain display-domain test images with realistic camera noise. Three main steps are performed:

(1) White Balance,
(2) Demosaicking,
(3) Color Transformations.

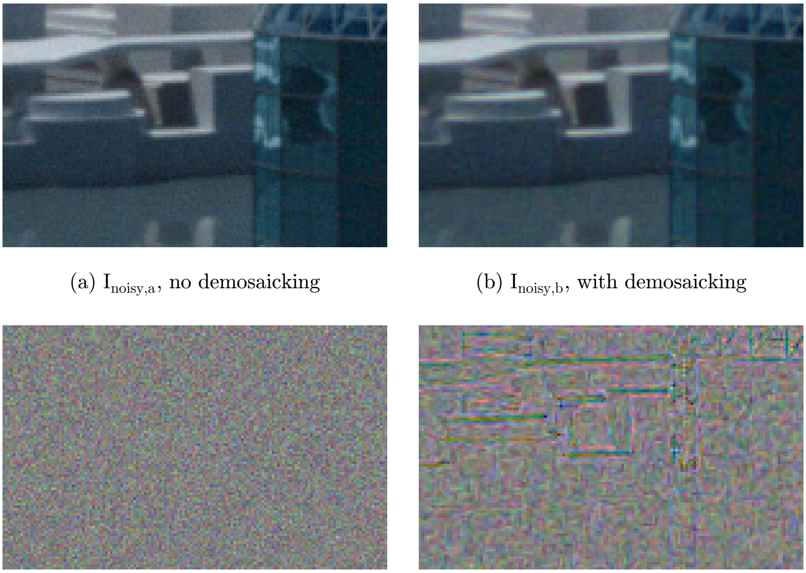
After the three steps a display-domain image is obtained. In the following, we discuss the influence of the individual steps on the camera noise characteristics for a detailed understanding.

White balance is a known gain factor that is different for each color. While the white balance directly influences the noise level in the different color channels, it does not affect the distribution, as it is a linear transformation.

The next step is to create a full-color image with three color values per pixel by demosaicking the white-balanced raw data. Different demosaicking algorithms can be used and the noise characteristic is changed depending on the algorithm. The demosaicking algorithm we use in our test was a standard directional interpolation method using a 5×5 window. The green interpolation is a gradient based decision between horizontal and vertical linear interpolation. The red and blue interpolation additionally uses green high pass information for correcting the chroma values. Like most debayering approaches the algorithm is nonlinear and uses neighboring values, which introduces a spatial correlation between neighboring pixel values and a chromatic correlation between the three color channels.

The influence of the demosaicking step on the noise characteristics is illustrated in Fig. 3. On the left, a test image with uncorrelated noise is shown, which is processed without the demosaicking step. The test image on the right was processed including the demosaicking step, and thus the noise in the image is spatially correlated. The respective difference images are obtained by calculating the difference $I_d = I_{\text{ref}} - I_{\text{noisy}}$ between the reference image I_{ref} and the respective noisy image I_{noisy}. While the difference image $I_{d,a}$ in Fig. 3(c) shows uncorrelated noise, the spatially correlated noise due to demosaicking is shown in Fig. 3(d). Both difference images are scaled the same way to visualize the effect unbiased. We see that the noise after demosaicking is structured and of coarser grain. This might lead to a lower performance of standard denoising algorithms.

The third step, the color transformation, is composed of different steps, usually a nonlinear tone mapping, a color space conversion and a gamma transformation. While the exact color transformation is an individual choice, it is an essential step to map the linear raw values onto displayable signals. The nonlinear tone mapping and the gamma transformation lead to a nonlinear signal-dependence of the noise and to



(a) $\mathrm{I}_{\mathrm{noisy,a}}$, no demosaicking

(b) $\mathrm{I}_{\mathrm{noisy,b}}$, with demosaicking

(c) $\mathrm{I}_{\mathrm{d,a}}$, uncorrelated noise

(d) $\mathrm{I}_{\mathrm{d,b}}$, spatially correlated noise

Fig. 3. Crop of the sequence "City". Noisy image (left) and noisy image with demosaicking (right). In the second row the respective difference image $I_d = I_{\text{ref}} - I_{\text{noisy}}$, scaled for display.

an unknown noise distribution. The color space conversion can strengthen the chromatic and thereby the spatial correlation.

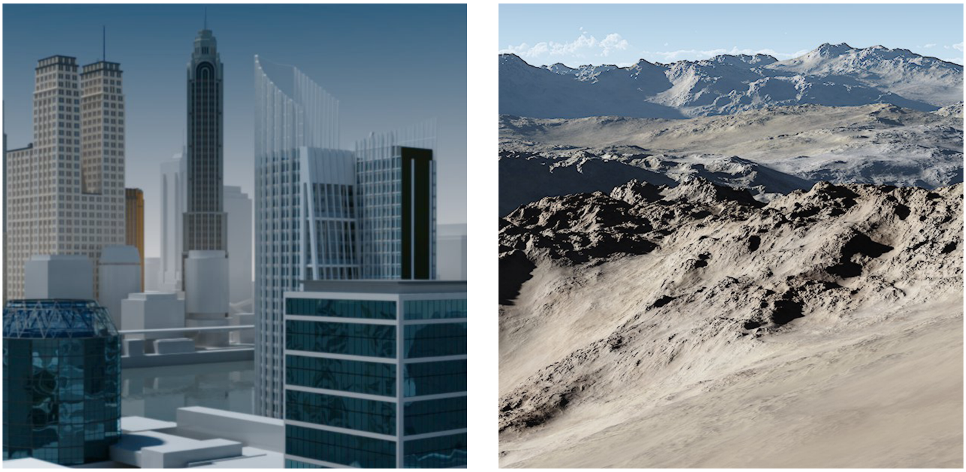
The images in display domain, therefore, have a noise characteristic that is spatially correlated, signal-dependent and with an unknown noise distribution. Hence, the noise characteristic is very different to AWGN in the display domain, which is the adequate domain for human observers and thus the appropriate domain for the evaluation of denoising algorithms. For a realistic evaluation, test data in the display domain is required. We show how to obtain realistic test data taking into account the complete camera processing pipeline in the next section.

3. Test Setup

To evaluate denoising algorithms, our test setup consists of a reference image, a noisy image and a denoised image. While noisy images could be obtained in the form of real camera data, we would lack the corresponding reference imagery. Applying noise to a readily processed image, e.g. an image from the Kodak data set, cannot be used for tests with realistic camera noise, as it is necessary to include all the processing steps described in Sec. 2.2. To obtain realistic test data, a full-color reference in the linear domain is required. We obtained this reference data from a rendered 3D scene.

3.1. *Reference sequences*

For our test we use two different scenarios, one pan over a city, named "city", Fig. 4(a), and a landscape sequence obtained by moving a static image by a few pixels to provide a video sequence, named "landscape", Fig. 4(b). The sequences have been chosen to reflect typical challenges in denoising natural images. The city sequence is



(a) city

(b) landscape

Fig. 4. One frame of the computer-generated test sequence.

dominated by horizontal and vertical edges and squares, whereas the landscape sequence has a lot of fine details that are not part of larger structures. In the landscape sequence, the optical low pass filter of the simulated camera has been adjusted to be less restrictive such that more high-frequency content is left in the images. The sequences are 16 bit data rendered in linear domain. To incorporate the optics of a camera system the images are multiplied in the Fourier domain with the optical transfer function of the camera. This step takes into account the diffraction limited lens, the optical low pass filter and the pixel aperture. For details we refer to [28]. The rendering of the 3D scene has to provide high resolution images to avoid aliasing effects in the camera optics simulation. We use a resolution of 4096^2 pixels to obtain 1024^2-sized images as simulation output. This approach provides realistic reference data in linear domain. Applying the color transformation to this data provides a display-domain reference image.

3.2. *Test sequences*

Besides the reference images, noisy test images are required for denoising evaluation. To generate these noisy images, we use the most usual and most simple model, AWGN in display domain, and our new realistic camera noise model. Additionally we use noise models that enable us to evaluate the impact on visual quality of the two main differences between camera noise and AWGN individually: signal-dependence and spatial correlation through demosaicking.

In Fig. 5 the usual AWGN model corresponds to the simulation path named "noisy AWGN uncorrelated". We use AWGN with a standard deviation of 1100 in 16 bit domain, which is equivalent to approximately 1.7% of the signal range. To

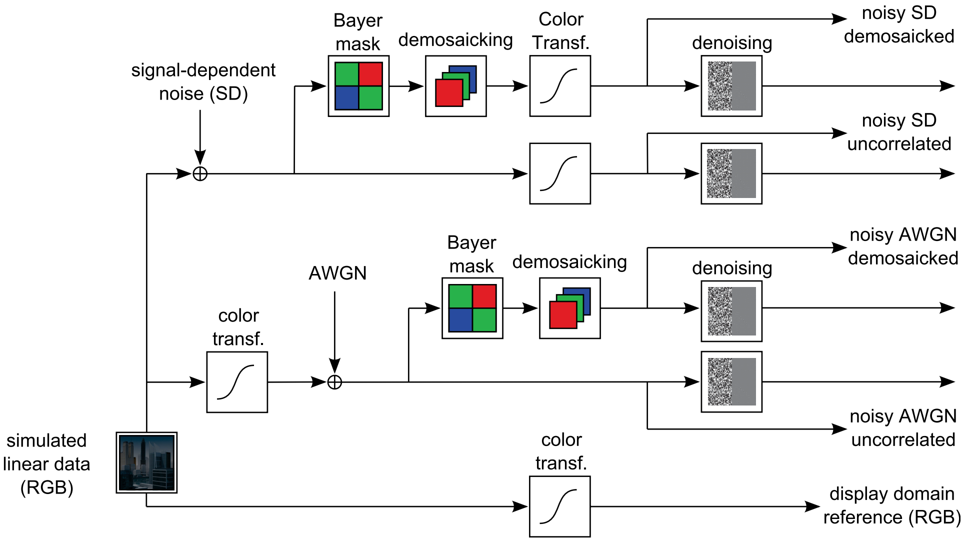


Fig. 5. Processing of simulated sensor data for the test using signal-dependent noise (SD) and AWGN.

evaluate the influence of the demosaicking step on visual quality, we simulate AWGN and apply a Bayer mask with subsequent demosaicking, which is represented by the simulation path "noisy AWGN demosaicked" in Fig. 5. While this does not correspond to real camera data, we can expect more realistic results by including this step into the usual AWGN model. To obtain signal-dependent noise, we replace the AWGN based noise model with the signal-dependent camera noise from Sec. 2 added in linear domain ("noisy SD uncorrelated" in Fig. 5). The noise level of the camera noise in our test sequences corresponds to a sensitivity of 3200 ASA. To obtain realistic camera noise the demosaicking step must be included, thus the simulation path for realistic camera noise is "noisy SD demosaicked" in Fig. 5.

To obtain denoised sequences, all the noisy sequences from Fig. 5 are denoised with two different denoising algorithms, BM3D [4] and BLS-GSM [3].

3.3. *Subjective test design*

With the approach described above, we obtained reference, noisy and denoised images. To evaluate denoising algorithms, the visual quality of the noisy and the respective denoised image must be compared. We thus require a metric, which can reflect the visual quality of the noisy and denoised test images similar to the human perception. However, to the best of our knowledge, the performance of quality metrics has not been validated for realistic camera noise.

The validation of quality metrics is based on databases that contain test images with artificial degradations and respective results of subjective tests, in which the test participants assess the perceived visual quality of these test images. To determine quality metric performance, the metrics are applied to the database test images and the results are compared to the respective subjective test result. While databases as the LIVE [29], the TID2008 [30] and the CSIQ [31] database were used to evaluate state-of-the-art quality metric performance with respect to different noise models, none of them uses realistic camera noise. The performance of the quality metrics is thus unknown for camera noise.

To obtain reliable information on the visual quality of our noisy and denoised test sequences, we conduct a subjective test with our test material. We used the double stimulus DSIS methodology with a undistorted reference and impaired noisy sequence according to ITU-R BT.500. A discrete scale from 1 to 10, representing a impairment range of "very annoying" to "imperceptible", was used. The test participants were 18 students, aged between 20 and 30. The task for the participants was to assess the perceived impairment of the images. Before each test, a training session was performed and the expected distortion, blur and noise, were mentioned in this training. The test was performed in the ITU-R BT.500 compliant video quality evaluation laboratory at the Institute for Data Processing at Technische Universität München. For displaying the videos, a color calibrated Sony BVM-L230 reference LCD display with a screen diagonal of 23 inches was used. To get reliable results, the outlier were removed in the post processing of the subjects' votes. Votes were removed, if they deviated more than 2σ from the mean for a sequence. Using this criterion, 4.6% of all votes were discarded. After outlier removal, the mean opinion score (MOS) was determined for the different test images.

The subjective test results provide reliable values for the subjective quality of all our test images. These results enable us to compare the visual quality of the noisy and denoised test sequences. Furthermore, the test data enables us to determine the performance of quality metrics with respect to realistic camera noise.

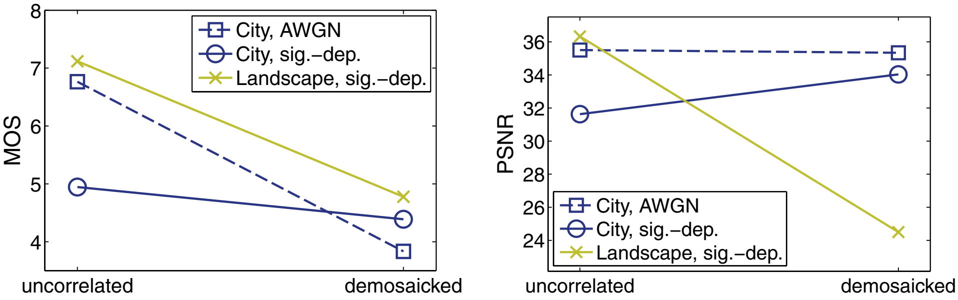
4. Visual Quality of Sequences Degraded by Camera Noise

The performance of denoising algorithms is evaluated by comparing the quality of a denoised image to the quality of the respective noisy image. To determine the performance of denoising algorithms we therefore need to evaluate the quality of the noisy images first.

4.1. *Subjective quality*

As described in Sec 3.2, four different noise models were used in our test: the usual AWGN model, AWGN with demosaicking, signal-dependent noise without demosaicking and finally the realistic camera noise model — signal-dependent noise with demosaicking. In this section, we evaluate the visual quality of the noisy test sequences and analyze the main differences between the realistic camera noise and AWGN: spatial correlation introduced through demosaicking and signal-dependence.

In Fig. 6(a) the MOS results are shown for sequences without the demosaicking step, thus containing uncorrelated noise, on the left, and the results for noisy sequences with demosaicking are shown on the right. We notice that the demosaicking leads to a lower MOS for all three test sequences. The MOS of the city sequence with AWGN is about 3 scores lower when demosaicking is included. Regarding the sequences with signal-dependent noise, the MOS is 0.5 lower for the city



(a)

(b)

Fig. 6. The MOS and PSNR results for the test sequences "City" and "Landscape" using the traditional AWGN model (dashed) and the realistic signal-dependent noise (solid lines). The uncorrelated noise, processed without demosaicking, is shown on the left, the results with demosaicking on the right.

sequence and 2.3 lower for the landscape sequence when demosaicking is included. The demosaicking changes the difference in visual quality of the sequences: While it is noticeable that the MOS for the uncorrelated noisy landscape sequence is quite good (7.1) compared to the respective city sequence (4.9), the difference is much smaller (0.4) when demosaicking is included. This may be explained by the image content: While the fine grain of the uncorrelated noise that is hardly differentiable to the sand, the coarser grain of the spatially correlated noise that is clearly visible in the landscape sequence. For all three test sequences the spatially correlated noise is perceived as more annoying than the uncorrelated noise. Furthermore, the demosaicking changes the difference in visual quality of different noisy images.

To reflect realistic camera noise a signal-dependent noise model is required. The noise level of signal-dependent noise depends on the signal and thus on the image content. Through the color transformation, the signal-dependence is nonlinear in display-domain images, usually noise is most visible in dark regions of the image. The noise level of signal-dependent noise is varying with the image content and thus not directly comparable to AWGN. For the city sequence we simulated both AWGN and the signal-dependent noise. While AWGN is classified as less annoying than signal-dependent noise when it is uncorrelated, AWGN with demosaicking is classified as more annoying than signal-dependent noise with demosaicking. That means, the relative order is changed with the noise model.

The MOS results showed that demosaicking has a significant impact on visual quality, in our test it leads to a lower visual quality of the noisy sequences. The noise model can change the difference and the relative order of the visual quality of the test sequences.

The effect of the noise model on the PSNR results, shown in Fig. 6(b), shows clearly different results than for the MOS. While the decrease with demosaicking is correct for the landscape sequence, very different results are obtained for the city sequences: The PSNR is approximatly constant when AWGN is compared to AWGN with demosaicking, and regarding the signal-dependent noise it is about 2dB higher for the demosaicked city image (33.95 dB). With demosaicking, the quality of the city sequence is rated much higher compared to the landscape sequence, thus the relative order does not correspond to the MOS results.

In this section, we have seen that the noise characteristics have a significant effect on the subjective quality. Interestingly, subjective quality in terms of MOS does not match the PSNR of the test sequences. While the PSNR is still widely used, it is well known that it is not optimal for visual quality estimation. Therefore, we evaluate more sophisticated metrics that correlates better with the visual quality given by the MOS results of our subjective test in the next section.

4.2. *Full-reference quality metrics*

To improve the correlation with the perceived visual quality, adaptions of the mentioned PSNR were proposed, such as the visual signal-to-noise ratio (VSNR) [10] and the human visual system based PSNR (PSNR-HVS/PSNR-HVSM) [11]. The performance of several quality metrics including the above has been evaluated using the TID2008 database in [30] that contains a test setup with different types of noise, including spatially correlated noise. In this test the PSNR-HVS achieved a high correlation with the subjective test results. We therefore can expect good results using this metric for our test data.

Other approaches that showed good results for various degradations, including white noise and Gaussian blur [32], are structural algorithms, as the structural similarity (SSIM) index [8] and the multiscale SSIM (MSSIM) [9], and information-theoretic algorithms, as the visual information fidelity (VIF) [12] and the information fidelity criterion (IFC) [33].

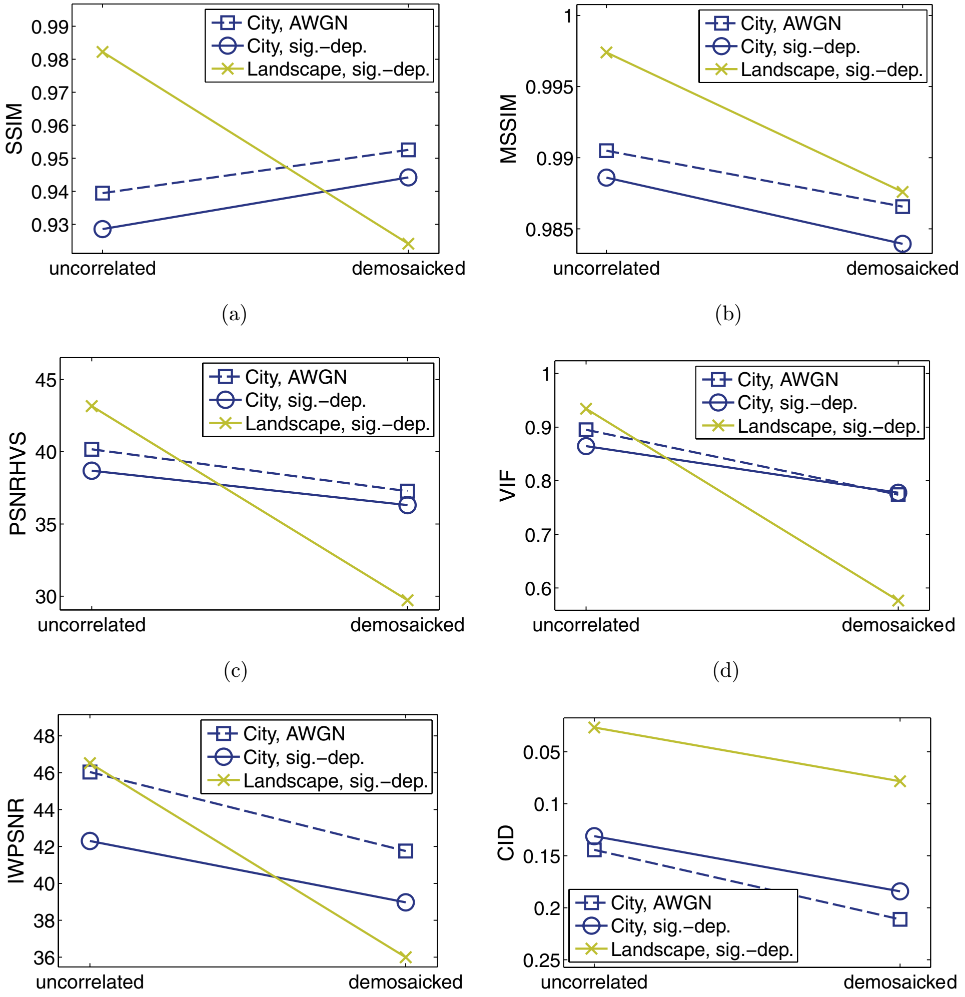
We add three recently proposed metrics that are adaptions of the PSNR and SSIM and showed high correlation based on tests with different standard databases, the PSNR-HMA [14], the "information content weighted PSNR" (IW-PSNR) and the "information content weighted SSIM" (IW-SSIM) [15]. As we work on color image sequences, we add a metric named "color-image-difference" (CID), which is designed for color images specifically [34].

We use the publicly accessible Matlab package "metrix mux" for most full-reference metrics. For PSNR-HVS, PSNR-HMA, IW-PSNR, IW-SSIM and CID we use the Matlab code provided by the authors. While some metrics have parameters that could be tuned, we stick to the default parameters for all the test sequences to achieve comparable results.

An optimal quality metric would show the same tendencies and relative order as the MOS results described in the previous section. One of our main results from the previous section was that including the demosaicking step in the processing pipeline of the noisy test images significantly reduces visual quality. That means, for both noise models the MOS is lower when demosaicking is included. In contrast, the SSIM is higher for the spatially correlated noise with demosaicking (Fig. 7(a)), the relative order of the sequences including demosaicking is thus incorrect.

The MSSIM better matches the subjective quality, as it shows a lower value for the noise after demosaicking, see the MSSIM in Fig. 7(b). IW-SSIM is not shown, as the results are very similar. Also the VIF, IFC, PSNR-HVS, PSNSR-HMA and IW-PSNR and CID rate the noisy images including demosaicking lower and thus reflect well the subjective quality, Figs. 7(c) shows PSNR-HVS and 7(d) the VIF. CID in Fig. 7(f) was included as it specifically works on color images. Despite that the noise seems very colored in the demosaicked landscape sequence, the quality of is rated very high using CID, which is incorrect. Additionally the order of AWGN and signal-dependent noise doesn't match the MOS result.

In summary, none of the tested metric is able to perfectly reflect the visual quality of all the noisy sequences as given by the MOS. However, most of the metrics showed lower results for the noisy images when demosaicking is included compared to the uncorrelated noise, and thus matched this important tendency shown by the MOS.



(a) (b)

(c) (d)

(e) (f)

Fig. 7. Quality metric results for the test sequences "City" and "Landscape" using the traditional AWGN model (dashed) and the realistic signal-dependent noise (solid lines). The uncorrelated noise, processed without demosaicking, is shown on the left, the results with demosaicking on the right.

4.3. *No-reference quality metrics*

A test setup to compare denoising algorithms usually works with pairs of a reference and the corresponding degraded noisy image. However, in real applications that require a denoising step, no reference image is available, which is the reason why real camera data cannot be used directly for a quantitative evaluation. To make use of real-world data that does not provide a reference, e.g. the new image test set proposed in [27], metrics that can evaluate the visual quality without a reference would be necessary. Besides, automatic parameter tuning would require such NR metrics.

The best denoising parameters depend on the image content, and NR metrics would enable to adjust the parameters for each image adaptively. We thus extend our recent work [21] to general no-reference (NR) metrics evaluation.

The recent progress in NR metrics is encouraging. In [13], a metric for parameter tuning, the "MetricQ" is shown to enhance denoising performance. As the metric showed a relatively low correlation in our previous work [21], we try several adaptions in this paper. In 2010, Moorthy and Bovik published the NR metric "blind image quality index" (BIQI) [35]. The metrics are usually evaluated using the LIVE database [32] that provides five different types of degradation: JPEG, JPEG2000 (JP2K), white noise, blur and fast fading. The two most relevant distortions for denoising are white noise and blur; we thus compare and select the methods mainly based on the Spearman correlation for these two types of distortion. For both distortion types a higher correlation than PSNR is reported for BIQI, so we include it in our test.

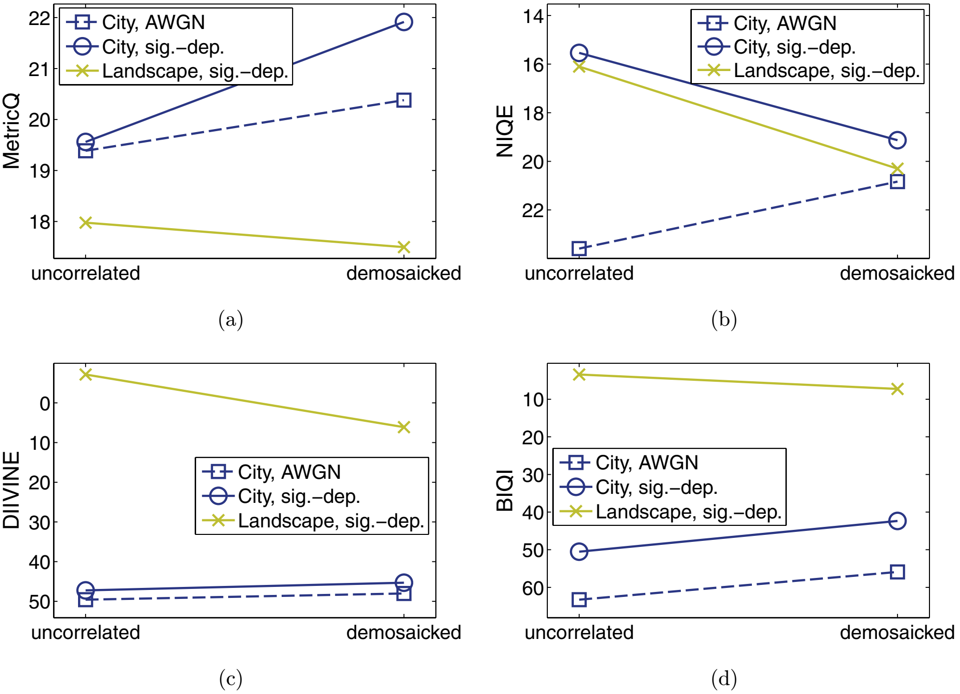
Since then, five other quality metrics have been published. The "general regression neural network (GRNN)", proposed 2011 [36], can provide results highly correlating to the human perception when the right dataset was chosen in the training. For in the most relevant criteria, white noise and blur, the correlation results strongly depends on the dataset: a high correlation for WN comes with a low correlation for blur and vice versa. Therefore we conclude that it may not be appropriate for denoising tests.

The NR metric "distortion identification-based image verity and integrity (DIIVINE)" [23] shows a very high correlation with the mean opinion score of the LIVE database of 0.98 for the white noise distortion and a considerable correlation coefficient of 0.92 for blur. It thus clearly outperforms a NR metric named BLIINDS [37]. An upgraded version thereof, named BLIINDS II [24] was proposed in 2012. However, we do not use it in our test, as it shows an extremely low correlation for white noise (0.1) and thus does not seem to be appropriate for denoising tests.

Two other recently published NR metrics show promising results: the "blind/referenceless image spatial quality evaluator" (BRISQUE) [25] shows a lower correlation for white noise compared to DIIVINE, however, the correlation coefficient is higher for blur. We thus include it in our tests. NIQE [22], which stands for "natural image quality evaluator", outperforms DIIVINE and BRISQUE in both white noise and blur and thus is the most promising method included in our test.

Figure 8 shows the NR-metric results for the noisy test data. MetricQ shows a very high quality for the demosaicked sequences compared to the sequences with uncorrelated noise, whereas the MOS indicates a lower quality for the demosaicked sequences. The same problem we have seen before with the FR metrics SSIM and PSNR. We tried to calculate the anisotropy map differently, however all the MetricQ versions were not able to reflect the decrase when demosaicking is included for the noisy data. As the results of the difference versions are very similar, we show only the original version in Fig. 8(a).

While BIQI shows the same problem (Fig. 8(d)), the metric NIQE in Fig. 8(b) shows for a part of the sequences a lower quality when demosaicking is included, it



(a)

(b)

(c)

(d)

Fig. 8. No-reference quality metric results for the test sequences "City" and "Landscape" using the traditional AWGN model (dashed) and the realistic signal-dependent noise (solid lines). The uncorrelated noise, processed without demosaicking, is shown on the left, the results with demosaicking on the right.

matches the MOS results for the sequences with signal-dependent noise. While DIIVINE in Fig. 8(c) shows the correct tendency for the landscape sequence, it rates the demosaicked sequences with higher quality for the city sequences, which does not match the MOS results. Regarding the relative order of the test sequences, the MOS value only shows a small difference in the absolute quality of the two sequences, while, according to BIQI and DIIVINE, the landscape sequence has a clearly higher quality rating.

While most metrics, except CID, rated the sequences containing signal-dependent noise lower than the sequences with AWGN, all no-reference metrics rate the sequences with signal-dependent noise higher compared to the AWGN sequences. None of the tested metrics, however, evaluated the sequences corresponding to the MOS, which indicates higher quality of the sequence with signal-dependent noise compared to the AWGN sequence when demosaicking is included, but a clearly lower quality for the sequence with uncorrelated signal-dependent noise compared to the sequence with uncorrelated AWGN.

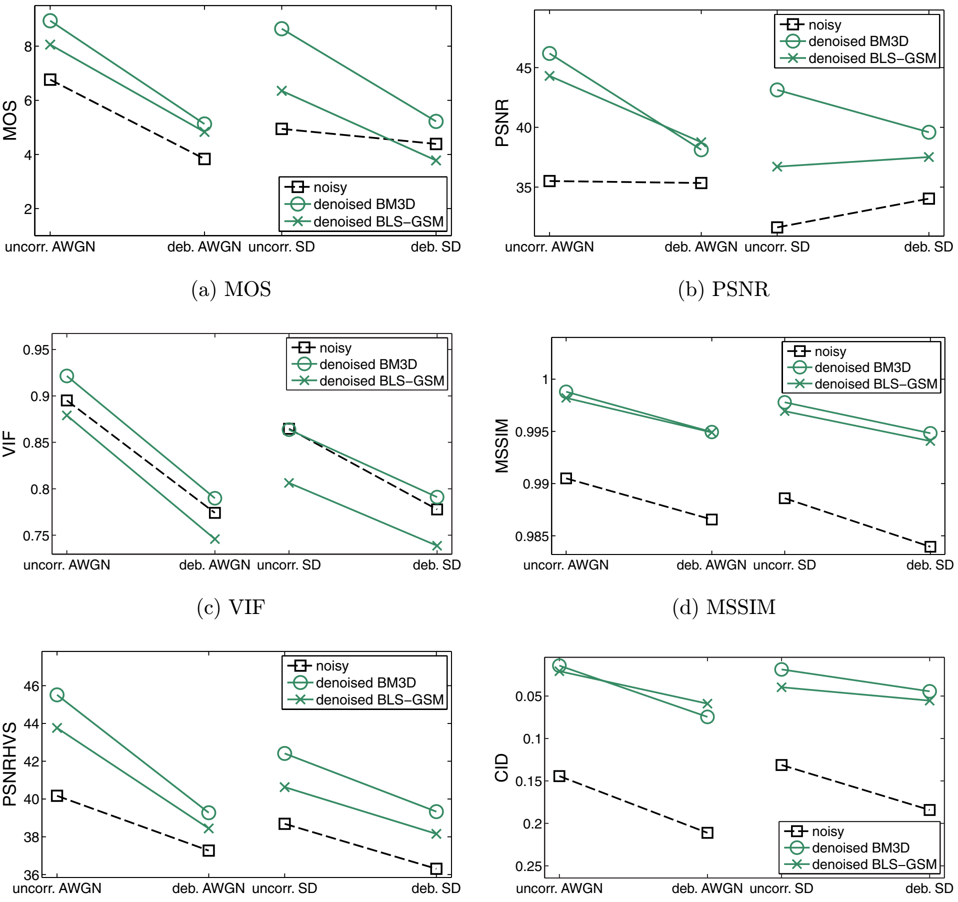
The NR metrics seem to less reflect the MOS result than the FR metrics. Especially the relative order of the sequences with different image content is more difficult to estimate without a reference.

5. Quality Assessment of Denoising Results

In the last section we discussed the subjective quality and the quality metric performance for noisy sequences. In this section, we discuss the visual quality of the denoised test sequences and based on these results we evaluate the performance of the quality metrics for the denoised test images. Finally we identify the most suitable quality metrics for denoising algorithm evaluation.

5.1. *Subjective quality*

We used two different state-of-the-art denoising methods, BM3D [4] and BLS-GSM [3]. In Fig. 9(a) the MOS results for denoised sequences are shown. In the subsequent plots the two left results show the quality of the usual AWGN model and the AWGN



(a) MOS

(b) PSNR

(c) VIF

(d) MSSIM

(e) PSNR-HVS

(f) CID

Fig. 9. MOS and full-reference quality metric results for the denoised test sequences using BM3D [4] and BLS-GSM [3].

model with demosaicking, whereas the two results on the right show the quality for signal-dependent noise and for signal-dependent noise with demosaicking. The last corresponds to the realistic camera noise model.

The noise model significantly affects the visual quality of denoising results. Including the demosaicking step into the noise simulation, leads to a MOS that is up to 3 scores lower. Thus, denoising uncorrelated noise is easier than denoising spatially correlated noise. The signal-dependence also affects the denoising performance: BLS-GSM achieves a 1.7 lower MOS for uncorrelated signal-dependent noise compared to AWGN. When both signal-dependence and demosaicking is included, thus realistic camera noise is used, the quality of the BLS-GSM result is even lower than the quality of the noisy sequence, thus no improvement is achieved with denoising. This shows that it is crucial to use an adequate model in the development of denoising algorithms.

We found that the noise model has a significant effect on the visual quality of denoising results. We have shown that demosaicking decreases denoising performance of both tested algorithms significantly and found that the signal-dependence leads to a lower visual quality of the BLS-GSM results.

5.2. *Full-reference quality metrics*

Optimal quality metrics should show the same tendencies and relative order as the MOS results. As described above, the demosaicking leads to a lower MOS, thus the lines in Fig. 9(a) are decreasing. This is reflected by the MSSIM, VIF, PSNR-HVS and CID, see Figs. 9(c)–9(f). The PSNR in Fig. 9(b) matches this tendency for the denoised sequences except for the sequence with signal-dependent noise denoised with BLS-GSM. The PSNR-HMA shows very similar tendency like the PSNR-HVS and thus is not shown. Regarding the relative order of the sequences with realistic camera noise, the VIF is the only metric in our test that rates the denoised sequences using BLS-GSM lower than the noisy sequence, which reflects the MOS (see "dem. SD" in Fig. 9(c)). However, the VIF rates the sequences denoised with BLS-GSM lower for the other noise types, too, which does not reflect the MOS results.

Comparing our subjective test results to the quality metric results shows that none of the tested metrics reflects the MOS perfectly. To determine the metrics that correlate best with the perceived quality, we evaluate the overall metric performance by means of the Spearman correlation coefficient.

Figure 10 shows the correlation coefficients for the full-reference metrics. The correlation coefficients with and without the Landscape sequences are shown. Most of the metrics show lower correlation coefficients when both test sequences are compared. The most stable results are given by PSNR-HMA and IWPSNR. The highest correlation to the subjective quality over the entire test is achieved by the full-reference metrics IFC, VIF, PSNR-HVS and PSNR-HMA. Thus, among the tested metrics these are the most suitable for the evaluation of denoising algorithms.

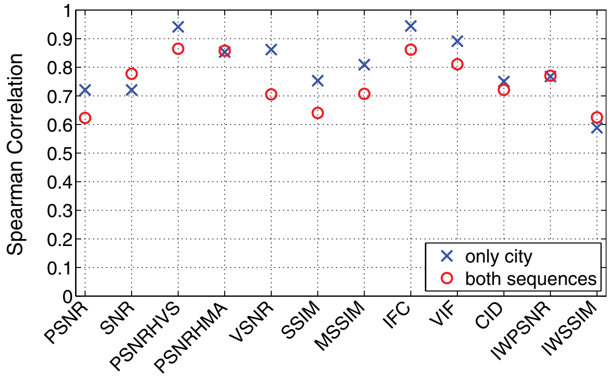
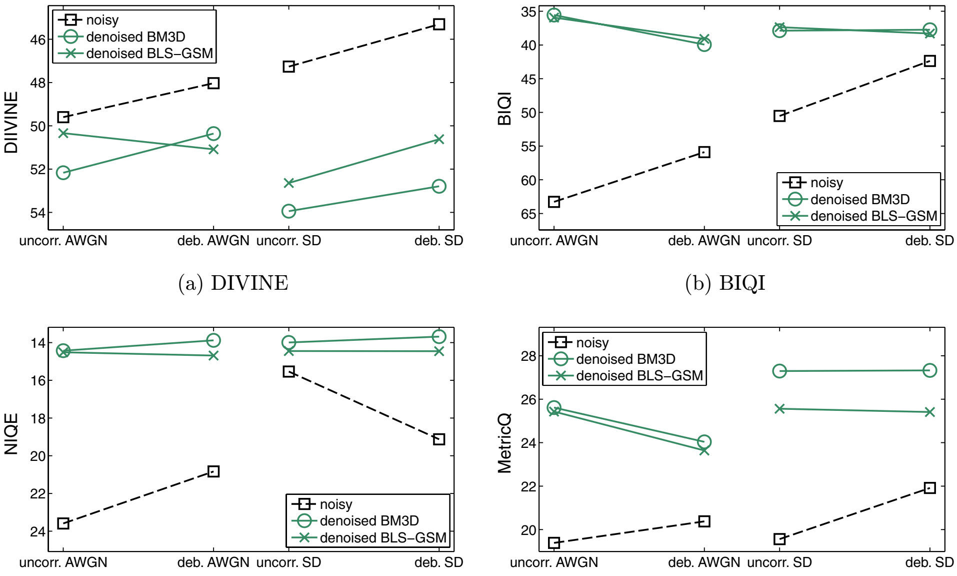


Fig. 10. Spearman correlation coefficients of full-reference metrics for the visual test.

5.3. *No-reference quality metrics*

The visual quality of the denoised sequences is higher than the noisy sequences, except the BLS-GSM algorithm applied to realistic camera noise. BM3D gives a higher MOS rating than BLS-GSM and demosaicking leads to a lower visual quality of the denoising results. An optimal metric would reflect these findings.

While the MetricQ indicates a lower quality for the demosaicked and denoised AWGN test sequences in Fig. 11(d), it does not show this tendency for the



(a) DIVINE

(b) BIQI

(c) NIQE

(d) MetricQ

Fig. 11. No-reference quality metric results for the denoised test sequences using BM3D [4] and BLS-GSM [3].

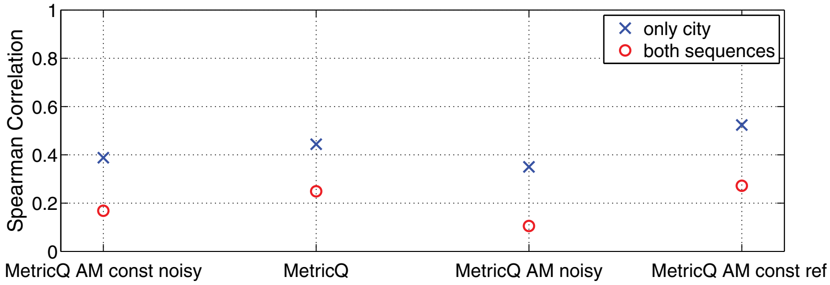


Fig. 12. Spearman correlation coefficients of MetricQ variants.

signal-dependent noise with demosaicking. We tested three versions in addition to the original version: MetricQ with an anisotropy map that was calculated on the same noisy images (MetricQ AM noisy). This is a reasonable setup, as the noisy image is known when denoising is applied. Additionally, we included two versions that are using the reference data partially. In a second new version, the MetricQ with the same anisotropy map (MetricQ AM const noisy) is used for all the different degradation types. As the authors proposed to use the most degraded image in this case, we used the noisy images with debayering included. In the third version the anisotropy map was calculated on the reference image (MetricQ AM const ref). Figure 12 shows the Spearman correlation coefficient over all the test sequences for the different variants of MetricQ. Compared to the original MetricQ, the variants using degraded images correlate less with human perception. The MetricQ with the anisotropy map calculated on the reference image (MetricQ AM const ref) shows a higher correlation.

While DIIVINE rates the denoising results with demosaicking higher compared to the denoised sequences without demosaicking, BIQI, NIQE and MetricQ show a constant or slightly decrease in quality when demosaicking is included before denoising, thus partially match the MOS result. DIIVINE rates all the noisy sequences higher than the denoised sequences. This does not match the MOS result. This may be due to a high sensitivity to blurring in the DIIVINE metric.

The noise model has a significant effect on the visual quality of denoising results and none of the NR metrics in our test fully reflects the MOS. To determine the NR metrics that correlate best with the perceived quality, we evaluate the overall metric performance by means of the Spearman correlation coefficient, shown in Fig. 13. The NR metrics achieve a much lower correlation compared to the FR metrics. The highest correlation coefficient is 0.6 for the BIQI metric, which additionally seems to be most robust to the image content. While NIQE and MetricQ achieve a considerable correlation coefficient of 0.4 on the city test set, BRISQUE practically does not correlate to the MOS results (0.08) and DIIVINE even shows a negative correlation coefficient, as it rates all the denoised sequences with a lower quality than the noisy sequence. We thus did not find a no-reference metric reflecting the MOS in our test.

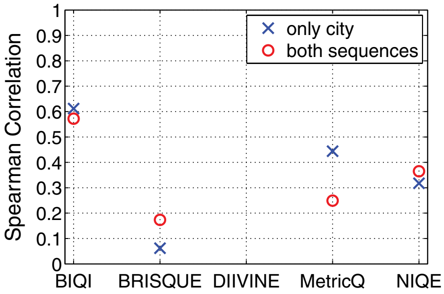


Fig. 13. Spearman correlation coefficients of no-reference metrics for the visual test. DIIVINE shows a negative correlation, which is probably due to the low quality rating for the denoised sequences.

While the correlation coefficients given in the literature usually are around 0.9, we have seen relatively low correlation coefficients between visual quality and metric result. This is mainly due to the different approach of calculating the correlation coefficients: in the literature correlation coefficients are usually given for a single degradation type (e.g. white noise) separately. This means the correlation coefficient represents how well a metric can compare a noisy image to another noisy image. In denoising evaluation, however, the metric must cope with different degradations that occur in one image simultaneously and find the best tradeoff, i.e. the best image. Thus the correlation coefficient has to be calculated over the denoised and the noisy test sequences to obtain useful information about how well a metric is suited for denoising evaluation.

We mentioned two reasons to use NR metrics: First, these metrics would enable us to directly use test data without reference, and thus real camera data could be used for denoising evaluation. However, as the correlation to human perception is much higher using FR metrics, we conclude that future denoising tests should be based on test sets including reference data. The second reason was that parameter tuning could be done automatically using a NR metric. Our test results indicate that BIQI could be a possible candidate for a parameter tuning metric. However, we need to verify this in future tests including more different sequences and different denoising results obtained by varying parameters.

6. Conclusion

In this paper we proposed a realistic camera noise model and showed how to integrate the complete camera processing chain into the test setup to evaluate denoising approaches. We not only showed that the noise characteristic of the typically applied AWGN is fundamentally different from our realistic noise model, but also identified that the signal-dependent noise as well as its spatial correlation has a significant impact on the perceived visual quality of noisy and denoised images. In our subjective test, we found that the spatially correlated noise, introduced in the demosaicking step, is perceived as more disturbing.

Denoising test images degraded by signal-dependent noise leads to results with a lower visual quality than for AWGN. Further, the performance of state-of-the-art denoising algorithms is considerably impaired by spatially correlated noise. Using realistic camera noise, denoising can even reduce the perceived visual quality.

To estimate the visual quality of denoising results without costly subjective tests, reliable quality metrics are required. In this paper, several state-of-the-art quality metrics are evaluated. No-reference metrics as well as full-reference metrics have been tested, however, none of the tested metrics fully reflects the perceived visual quality. While the widely used full-reference metrics PSNR and SSIM show a low correlation with the subjective test results, the highest correlation is obtained using the metrics PSNR-HVS, IFC, VIF and PSNR-HMA. All the tested no-reference metrics show a very low correlation to the test results.

With the significant impact of the chosen noise characteristic on the visual quality of noisy images and of denoising results, we conclude that a realistic noise model should be used in future research. For realistic denoising evaluation, a new extensive test set based on realistic noise, as well as new quality metrics that better reflect the visual quality of camera data, would be required.

References

[1] A. Buades, B. Coll and J. Morel, A review of image denoising algorithms, with a new one, *Multiscale Modeling & Simulation* **4**(2) (2005) 490–530.

[2] E. Vansteenkiste, D. Weken, W. Philips and E. Kerre, Perceived image quality measurement of state-of-the-art noise reduction schemes, in *Advanced Concepts for Intelligent Vision Systems* (Springer, 2006), pp. 114–126.

[3] J. Portilla, V. Strela, M. J. Wainwright and E. P. Simoncelli, Image denoising using scale mixtures of gaussians in the wavelet domain, *IEEE Transactions on Image Processing* **12**(11) (2003) 1338–1351.

[4] K. Dabov, A. Foi, V. Katkovnik and K. Egiazarian, Image denoising by sparse 3-d transform-domain collaborative filtering, *IEEE Transactions on Image Processing* **16**(8) (2007) 2080–2095.

[5] M. Elad and M. Aharon, Image denoising via learned dictionaries and sparse representation, in *CVPR* **1** (2006) 895–900.

[6] J. Mairal, F. Bach, J. Ponce, G. Sapiro and A. Zisserman, Non-local sparse models for image restoration, in *ICCV*, 2009, pp. 2272–2279.

[7] http://r0k.us/graphics/kodak/.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, *IEEE Transactions on Image Processing* **13**(4) (2004) 600–612.

[9] Z. Wang, E. P. Simoncelli and A. C. Bovik, Multiscale structural similarity for image quality assessment, in *Conference Record of the 37th Asilomar Conference on Signals, Systems and Computers* **2** (2003) 1398–1402.

[10] D. M. Chandler and S. S. Hemami, VSNR: A wavelet-based visual signal-to-noise ratio for natural images, *IEEE Transactions on Image Processing* **16**(9) (2007) 2284–2298.

[11] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti and M. Carli, New full-reference quality metrics based on HVS, in *VPQM*, 2006.

[12] H. Sheikh and A. Bovik, Image information and visual quality, *IEEE Transactions on Image Processing* **15**(2) (2006) 430–444.

[13] X. Zhu and P. Milanfar, Automatic parameter selection for denoising algorithms using a no-reference measure of image content, *IEEE Transactions on Image Processing* **19**(12) (2010) 3116–3132.

[14] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian and M. Carli, Modified image visual quality metrics for contrast change and mean shift accounting, in *Proceedings of CADSM*, 2011, pp. 305–311.

[15] Z. Wang and Q. Li, Information content weighting for perceptual image quality assessment, *IEEE Transactions on Image Processing* **20**(5) (2011) 1185–1198.

[16] H. J. Trussell and R. Zhang, The dominance of poisson noise in color digital cameras, *ICIP*, 2012, pp. 329–332.

[17] S. H. Park, H. S. Kim, S. Lansel, M. Parmar and B. Wandell, A case for denoising before demosaicking color filter array data, in *Conference Record of the 43rd Asilomar Conference on Signals, Systems and Computers*, 2009, pp. 860–864.

[18] R. Giryes and M. Elad, Sparsity based poisson denoising, in *IEEE 27th Convention of Electrical Electronics Engineers in Israel*, 2012, pp. 1–5.

[19] A. Foi, Clipped noisy images: Heteroskedastic modeling and practical denoising, *Signal Processing*, 2009, pp. 2609–2629.

[20] C. Liu, R. Szeliski, S. B. Kang, C. Zitnick and W. Freeman, Automatic estimation and removal of noise from a single image, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**(2) (2008) 299–314.

[21] T. Seybold, C. Keimel, M. Knopp and W. Stechele, Towards an evaluation of denoising algorithms with respect to realistic camera noise, in *IEEE International Symposium on Multimedia* **4**(5) (2013).

[22] A. Mittal, R. Soundararajan and A. C. Bovik, Making a "completely blind" image quality analyzer, *IEEE Signal Processing Letters* **20**(3) (2013) 209–212.

[23] A. K. Moorthy and A. C. Bovik, Blind image quality assessment: From natural scene statistics to perceptual quality, *IEEE Transactions on Image Processing* **20**(12) (2011) 3350–3364.

[24] M. A. Saad, A. C. Bovik and C. Charrier, Blind image quality assessment: A natural scene statistics approach in the DCT domain, *IEEE Transactions on Image Processing* **21**(8) (2012) 3339–3352.

[25] A. Mittal, A. K. Moorthy and A. C. Bovik, No-reference image quality assessment in the spatial domain, *IEEE Transactions on Image Processing* **21**(12) (2012) 4695–4708.

[26] *EMVA 1288, Standard for Characterization of Image Sensors and Cameras*, EMVA Std. 1288, 2010.

[27] S. Andriani, H. Brendel, T. Seybold and J. Goldstone, Beyond the kodak image set: A new reference set of color image sequences, in *ICIP*, 2013, pp. 2289–2293.

[28] M. Schöberl, W. Schnurrer, A. Oberdörster, S. Fößel and A. Kaup, Dimensioning of optical birefringent anti-alias filters for digital cameras, in *ICIP*, 2010, pp. 4305–4308.

[29] H. R. Sheikh, Z.Wang, L. Cormack and A. C. Bovik, LIVE image quality assessment database release 2. [Online]. Available: http://live.ece.utexas.edu/research/quality/subjective.htm.

[30] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli and F. Battisti, TID2008-A database for evaluation of full-reference visual quality assessment metrics, *Advances of Modern Radioelectronics* **10**(10) (2009) 30–45.

[31] E. C. Larson and D. M. Chandler, Most apparent distortion: Full-reference image quality assessment and the role of strategy, *Journal of Electronic Imaging* **19**(1) (2010) 011006.

[32] H. Sheikh, M. Sabir and A. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, *IEEE Transactions on Image Processing* **15**(11) (2006) 3440–3451.

[33] H. Sheikh, A. Bovik and G. de Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, *IEEE Transactions on Image Processing* **14**(12) (2005) 2117–2128.

[34] I. Lissner, J. Preiss, P. Urban, M. S. Lichtenauer and P. Zolliker, Image-difference prediction: From grayscale to color, *IEEE Transactions on Image Processing* **22**(2) (2013) 435–446.

[35] A. K. Moorthy and A. C. Bovik, A two-step framework for constructing blind image quality indices, *IEEE Signal Processing Letters* **17**(5) (2010) 513–516.

[36] C. Li, A. C. Bovik and X. Wu, Blind image quality assessment using a general regression neural network, *IEEE Transactions on Neural Networks* **22**(5) (2011) 793–799.

[37] M. A. Saad, A. C. Bovik and C. Charrier, A DCT statistics-based blind image quality index, *IEEE Signal Processing Letters* **17**(6) (2010) 583–586.