Reduced-reference metric design for objective perceptual quality assessment in wireless imaging

Ulrich Engelke [a,*], Maulana Kusuma [b], Hans-Jürgen Zepernick [a], Manora Caldera [c]

[a] *Blekinge Institute of Technology, P.O. Box 520, SE-372 25 Ronneby, Sweden*
[b] *Universitas Gunadarma, Jakarta 12540, Indonesia*
[c] *Gibson Quai-AAS, 30 Richardson Street, Perth, WA 6005, Australia*

Abstract

The rapid growth of third and development of future generation mobile systems has led to an increase in the demand for image and video services. However, the hostile nature of the wireless channel makes the deployment of such services much more challenging, as in the case of a wireline system. In this context, the importance of taking care of user satisfaction with service provisioning as a whole has been recognized. The related user-oriented quality concepts cover end-to-end quality of service and subjective factors such as experiences with the service. To monitor quality and adapt system resources, performance indicators that represent service integrity have to be selected and related to objective measures that correlate well with the quality as perceived by humans. Such objective perceptual quality metrics can then be utilized to optimize quality perception associated with applications in technical systems.

In this paper, we focus on the design of reduced-reference objective perceptual image quality metrics for use in wireless imaging. Specifically, the Normalized Hybrid Image Quality Metric (NHIQM) and a perceptual relevance weighted L_p-norm are designed. The main idea behind both feature-based metrics relates to the fact that the human visual system (HVS) is trained to extract structural information from the viewing area. Accordingly, NHIQM and L_p-norm are designed to account for different structural artifacts that have been observed in our distortion model of a wireless link. The extent by which individual artifacts are present in a given image is obtained by measuring related image features. The overall quality measure is then computed as a weighting sum of the features with the respective perceptual relevance weight obtained from subjective experiments. The proposed metrics differ mainly in the pooling of the features and amount of reduced-reference produced. While NHIQM performs the pooling at the transmitter of the system to produce a single value as reduced-reference, the L_p-norm requires all involved feature values from the transmitted and received image to perform the pooling on the feature differences at the receiver. In addition, non-linear mapping functions are developed that relate the metric values to predicted mean opinion scores (MOS) and account for saturations in the HVS. The evaluation of prediction performance of NHIQM and the L_p-norm reveals their excellent correlation with human perception in terms of accuracy, monotonicity, and consistency. This holds not only for the prediction performance on images taken for the training of the metrics but also for the generalization to unknown images. In addition, it is shown that the NHIQM approach and the perceptual relevance weighted L_p-norm outperform other prominent objective quality metrics in prediction performance.

Key words: Objective perceptual image quality, Normalized hybrid image quality metric, Perceptual relevance weighted L_p-norm, Reduced-reference, Wireless imaging

* Corresponding author.
Email addresses: ulrich.engelke@bth.se (Ulrich Engelke), mkusuma@staff.gunadarma.ac.id (Maulana Kusuma), hans-jurgen.zepernick@bth.se (Hans-Jürgen Zepernick), mcaldera@gqaas.com.au (Manora Caldera).

1. Introduction

The development of advanced transmission techniques for third generation mobile communication systems and their long-term evolution has paved the way for the delivery of mobile multimedia services. Wireless imaging applications are among those services that are offered on modern mobile devices to support communication options beyond the traditional voice services. As the bandwidth resources allocated to mobile communication systems are scarce and expensive, digital images and videos are compressed prior to their transmission. In addition, the time-varying nature of the wireless channel caused by multipath propagation, the changing interference conditions within the system, and other factors cause the channel to be relatively unreliable. As a consequence, the quality of wireless imaging services are impaired not only by the lossy compression technique adapted but also by the burst error mechanisms induced by the wireless channel.

The performance evaluation of mobile multimedia systems has conventionally been based on link layer metrics such as the signal-to-noise ratio (SNR) and the bit error rate (BER) [25]. Similarly, performance of image compression techniques is often quantified by fidelity metrics such as the mean squared error (MSE) and the peak signal-to-noise ratio (PSNR) [46]. In the case of communicating visual content, however, it has been shown that these metrics do not necessarily correlate well with the quality as perceived by the human observer [13,45]. As a result, user-oriented assessment methods that can measure the overall perceived quality have gained increased interest in recent years. It is expected that these methods will facilitate more efficient designs of mobile multimedia systems by establishing trade-offs between the allocation of system resources and Quality of Service (QoS) [27,33]. In other words, not only metrics associated with the underlying technical system are considered but also quality indicators that can accurately predict the visual quality as perceived by human observers.

1.1. *Visual quality assessment*

A wide range of approaches has been followed in the design of such visual quality metrics ranging from simple numerical measures [8] to highly complex models incorporating those characteristics of the human visual system (HVS) that are considered as being crucial for visual quality perception [22,30,37]. Specifically, the phenomenon that the HVS is adapted to extraction of structural information has received strong attention for metric design [1,3,40]. These psychophysical approaches, which are based on modeling various aspects of the HVS, correlate well with human visual perception and are usable over a wide range of applications. However, these benefits often come at the expense of high computational complexity. In contrast, methods following an engineering inspired approach are mainly based on image or video analysis and feature extraction, which does not exclude that certain aspects of the HVS are considered in the metric design.

Most of the proposed HVS based metrics are following the full-reference (FR) approach [6,20,34,43], meaning, that they rely on the reference image being available for the quality assessment. Clearly, this limits their applicability to wireless imaging as a reference image would generally not be available at the receiver where quality assessment takes place. Thus, a no-reference (NR) metric may be more appropriate since it measures the quality solely based on the received image. Although it is easy for humans to judge the quality of an image without any reference, it is extremely difficult for an automated algorithm to execute.

As a consequence, metrics following the NR approach such as [11,24,36] usually provide inferior quality prediction performance as compared to metrics that take into account some amount of reference information from the transmitted image, or process the whole original image itself as in case of FR metrics. Furthermore, as NR metrics provide an absolute measure about the quality of a received image, it may be difficult to distinguish quality degradations that have been induced during image transmission from those that have already been present in the image prior to transmission. Hence, there would be strong limitations to execute link adaptation and resource management procedures based upon this type of metrics.

In this respect, a good compromise between the FR and NR methods are the reduced-reference (RR) metrics. These metrics rely only on a set of image features, the reduced-reference, instead of the entire reference image. These features are simply extracted from an image prior to its transmission and used at the receiver for detecting quality degradations. The reduced-reference may then be transmitted over an ancillary channel, piggy backed with the image, or embedded into the image using data hiding tech-

niques [42].

Wang et al. [41] have proposed a RR metric based on a natural image statistic model in the wavelet domain and Carnec et al. [2] define the C4 criterion which is an RR metric based on an elaborate model of the HVS. Both metrics have been shown to correlate well with human perception, which comes at the cost of a high computational complexity. This may restrict their application in the context of wireless imaging where computational resources are very limited, in particular in the mobile device. Yamada et al. [47] and Chono et al. [5] propose RR metrics that can accurately predict PSNR. The former metric is based on a selection of representative luminance values whereas the latter metric utilizes distributed source coding to communicate the RR signal. These metrics may be applicable for usage in an image communication context due to their low computational complexity. However, the ability of these metrics to accurately predict perceived visual quality is doubtful due to the poor quality prediction performance of PSNR.

1.2. *Overview of the proposed metric design*

In view of the above, this paper focuses on the development of RR objective perceptual image quality metrics that are applicable in a wireless imaging context. As such, image impairments representative for a wireless imaging system are produced to constitute the basis of the design framework. In addition, particular care has been taken to limit the overhead needed for communicating reduced-reference information and hence conserve the scarce bandwidth resources allocated to wireless systems. Furthermore, feature extraction algorithms are selected to have small computation complexity in order not to drain battery power at the wireless handheld device and in turn support longer service time.

Specifically, images in the widely adopted Joint Photographic Experts Group (JPEG) format are examined with typical impacts of a mobile communication system included through a simulated wireless link. This system under test enabled us to produce artifacts beyond those inflicted purely by lossy source encoding but to account also for end-to-end degradations caused by a transmission system. In particular, the artifacts of blocking, blur, ringing, masking, and lost blocks have been observed ranging from extreme to almost invisible presence.

The information about the individual artifacts in an image can be deduced from related image features such as edges, image activity and histogram statistics. The extent by which the considered artifacts exist in a given image can therefore be quantified by using selected image feature extraction algorithms. As some artifacts influence the perceived quality stronger than others, perceptual relevance weights are given to the associated image features. Clearly, subjective experiments and their analysis are not only instrumental but critical in the process of revealing the specific values of perceptual relevance weights. For this reason, we conducted subjective image quality experiments in two independent laboratories. The particular values of the weights were deduced as Pearson linear correlation coefficients between the related features and the Mean Opinion Scores (MOS) from the subjective experiments. In this respect, the perceptual relevance weights obtained from analyzing the subjective data constitute a key component in the transition from subjective quality prediction methods to an automated quality assessment that would be suitable for real-time applications. Given these perceptual relevance weights, an objective perceptual image quality metric may then be designed to exploit image feature values and their weights within a suitable pooling process. In this paper, we consider two feature-based objective perceptual quality metrics that mainly differ in the pooling process and the amount of reduced-reference as follows.

Firstly, the Normalized Hybrid Image Quality Metric (NHIQM) is designed. It operates on extreme value normalized image features from which it produces a weighted sum with respect to the relevance of the involved features. The result is a single value that can be communicated from transmitter to receiver where it is utilized as reduced-reference information. The same processing is performed on the received image resulting in the related NHIQM value. The absolute difference between the NHIQM values of the transmitted and received image constitutes the objective perceptual quality metric and is used to detect distortions.

Secondly, we consider a perceptual relevance weighted L_p-norm as a means of pooling the image features. Specifically, the L_p-norm is applied here to detect differences between features [7,10]. In this case, the pooling at the transmitter is omitted but requires the features being transmitted over the channel to the receiver. At the receiver, the difference between the transmitted and received features are combined to an overall quality metric. This ap-

proach allows to track degradations for each of the involved features. On the other hand, the amount of reduced-reference overhead is increased compared to the NHIQM-based approach.

The design of both feature-based RR metrics, NHIQM and L_p-norm, follows the same methodology. It comprises of the selection of suitable feature extraction algorithms, the feature extraction for image samples of a training set, normalization of the calculated feature values, and the acquisition of the perceptual relevance weights from the subjective experiments. A non-linear mapping function is derived in a final step that relates the objective perceptual quality metric to predicted MOS. In this way, non-linearities in the HVS with respect to the processing of quality degradations can be accounted for. The non-linear mapping function is derived using curve fitting methods where, again, the MOS from the subjective experiments are essential in deriving the parameters of the mapping functions.

A comprehensive evaluation of the prediction performance of NHIQM and the L_p-norm is provided in terms of accuracy, monotonicity, and consistency [35]. These performance measures are given for the metric design on a training set of images and the generalization to unknown images. It turns out that the proposed feature-based metrics outperform other considered RR and FR metrics in the context of wireless imaging distortions and with respect to the above prediction performance measures.

1.3. *Contributions of this work*

Considering the above, this paper contributes a framework for image quality metric design in a wireless communication system. As such, the metrics proposed in this paper have been designed to be able to measure quality degradation during image transmission using a RR approach. Unlike other RR metrics from the literature, the metrics in this paper are designed based on a set of test images that take into account the complex nature of a wireless communication system, rather than just accounting for source coding artifacts or additional noise. Furthermore, low computational complexity and low overhead in terms of reduced-reference have been major design issues in order to put low burdens on the communication system.

A statistical analysis of experiments that we conducted in two independent laboratories reveals insight into the subjectively perceived quality of wireless imaging distortions. In addition, a statistical and correlation analysis of objective feature metrics provides further insight into the artifacts observed in wireless imaging and the performance of the feature metrics that were used to quantify the related artifacts. Comparison of the proposed RR quality metrics to other contemporary quality metrics reveals the ability of the proposed metrics to predict perceived quality in the context of wireless imaging.

This paper is organized as follows. Section 2 provides an overview of RR objective quality assessment in wireless imaging and the particular system under test as considered in this paper. A detailed description of the conducted subjective quality experiments is contained in Section 3 along with a statistical analysis of the experiment outcomes. The objective feature extraction metrics, which build the very basis of the metric design, are discussed in Section 4. An additional analysis of the feature metrics provides insight into their performance to measure artifacts in the images. On the basis of both the subjective and objective data, the RR metric design for objective perceptual quality assessment is then described in detail in Section 5. In Section 6, the prediction performance of NHIQM and L_p-norm is evaluated and compared to other prominent objective quality metrics. Finally, conclusions are drawn in Section 7.

2. Reduced-reference objective perceptual quality assessment in wireless imaging

A typical link layer of a wireless communication system is shown in Fig. 1. Here, the functional blocks in shaded boxes relate to the components that would need to be included for performing the operations related to RR objective perceptual quality assessment. As such, the system is able to monitor quality degradations that are incurred during transmission unlike in the case of deploying a NR quality assessment method, where an absolute quality of the received image would be obtained. Given the strict limitations on system resources such as bandwidth, the overhead induced by the reduced-reference becomes a critical metric design issue. It is therefore beneficial to extract and pool representative features of an image I_t at the transmitter (t) in order to condense the image content and structure to a few numerical values. The transmission of the source encoded image may then be accompanied by the reduced-reference, which could be communi-

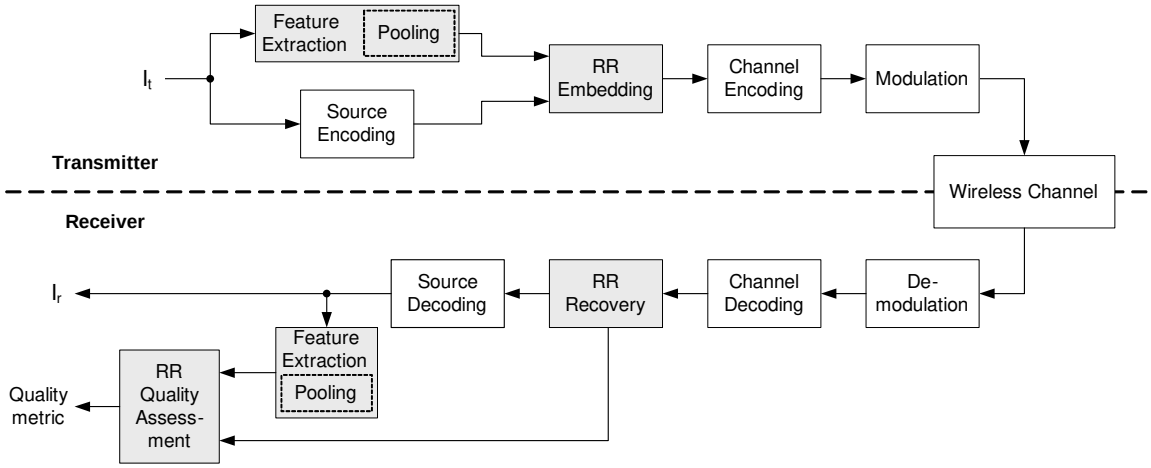


Fig. 1. Overview of reduced-reference objective perceptual quality assessment deployed in a wireless imaging system.

cated either in-band as an additional header or in a dedicated control channel. Subsequently, channel encoding, modulation and other wireless transmission functions are performed on the source encoded image and the reduced-reference. At the receiving side, the inverse functions are performed including demodulation, channel decoding, and source decoding. The reduced-reference features are recovered from the received data and the related features of the reconstructed image I_r at the receiver (r) are extracted and pooled to produce the related metric value. The difference between metric values for the images I_t and I_r can then be explored for end-to-end image quality assessment. The outcome of the RR quality assessment may drive, for instance, link adaption techniques such as adaptive coding and modulation, power control, or automatic repeat request strategies provided a feedback link would be available.

2.1. *System under test*

In the scope of this paper we consider a particular setup of the wireless link model as shown in Fig. 1 which turned out to results in a set of test images covering a broad range of artifact types and severities. In particular, the JPEG format has been chosen to source encode the images prior to transmission. It is noted that JPEG is a lossy image coding technique using a block discrete cosine transform (DCT) based algorithm, thus, facilitating an easy transition to state-of-the-art DCT based video codecs, such as H.264. Due to the quantization of DCT coefficients, artifacts may already be introduced during source encoding. A $(31, 21)$ Bose-Chaudhuri-Hocquenghem (BCH) code was then used for error protection purposes and binary phase shift keying (BPSK) for modulation. An uncorrelated Rayleigh flat fading channel in the presence of additive white Gaussian noise (AWGN) was implemented as a simple model of the wireless channel. Severe fading conditions may cause bit errors or burst errors in the transmitted signal which are beyond the correction capabilities of the channel decoder and as a result, artifacts may be induced in the decoded image in addition to the ones purely caused by the source encoding. To produce severe transmission conditions, the average bit energy to noise power spectral density ratio E_b/N_0 was chosen as 5 dB.

It should be noted, that the RR objective quality metric design is based upon this particular setup. However, the proposed metric design framework can be easily adopted to other specific system components, given that the objective data (test images) and subjective data (MOS) sets are available that are crucial for the metric design. This may for instance include an extension from JPEG to JPEG2000 or to measuring spatial artifacts in video, such as H.264.

2.2. *Artifacts in wireless imaging*

The system under test as outlined in Section 2.1 turned out to be beneficial with respect to generating impaired images ranging from extreme artifacts to images with almost invisible artifacts. Specifically, the range of artifacts spanned beyond those typically induced by source encoding such as block-

Fig. 2. Distorted image samples showing different artifacts: "Lena" with blocking, "Goldhill" with blur in 8×8 blocks (top); "Pepper" with ringing and intensity masking, "Barbara" with extreme artifacts (bottom).

ing and blur but also comprised of ringing, intensity masking, lost blocks, and combinations thereof. These artifacts will be briefly discussed in the following sections. In addition, some example images are shown in Fig. 2 to illustrate the observed artifacts.

2.2.1. *Blocking*

Blocking artifacts are inherent with block-based image compression techniques such as JPEG or H.264. Blocking or blockiness can be observed as surface discontinuity at block boundaries and is a direct consequence of the independent quantization of the individual blocks of pixels. In particular, in JPEG compressed images blocking is present on the 8×8 block borders due to independent quantization of the DCT coefficients.

2.2.2. *Blur*

Blur relates to the loss of spatial detail and is observed as texture blur. In addition, blur may be observed due to a loss of semantic information that is carried by the shapes of objects in an image. In this case, edge smoothness relates to a reduction of edge sharpness and contributes to blur. In relation to compression, blur is a consequence of the coarse quantization of frequency components and the associated suppression of high-frequency coefficients. In case of JPEG compression blur is usually observed within the 8×8 blocks rather than on a global scale.

2.2.3. *Ringing*

The artifact of ringing appears to the human observer as periodic pseudo edges around the original edges of the objects in an image. Ringing is caused by improper truncation of high-frequency components, which in turn can be noticed as high-frequency irregularities in the reconstruction. Ringing is usually more evident along high contrast edges, especially if these edges are located in areas of smooth textures.

2.2.4. *Intensity masking and lost blocks*

In general, masking occurs when the visibility of a stimulus is reduced due to the presence of another stimulus [45]. In this context, intensity shifts in parts of an image, or the whole image, may result in either a darker or brighter appearance of the area as compared to the original image and thus cause such a reduction in visibility. This phenomenon, which we refer to as intensity masking, is a typical artifact in wireless image communication appearing in the presence of strong multipath fading. In the worst case, entire image blocks are lost resulting in parts of the image being black.

3. Subjective image quality experiments

The methodology used for the subjective assessment of image quality is described hereafter. In particular, the laboratory environment, the test material, the panels of viewers, and the test procedure adapted in the subjective experiments are given in detail. According to the guidelines outlined in Recommendation BT.500-11 [17] of the radio communication sector of the International Telecommunication Union (ITU-R), subjective experiments were conducted in two independent laboratories. The first subjective experiment (SE 1) took place at the Western Australian Telecommunications Research Institute (WATRI) in Perth, Australia and the second subjective experiment (SE 2) was conducted at the Blekinge Institute of Technology (BIT) in Ronneby, Sweden.

3.1. *Laboratory environment*

The general viewing conditions were arranged as specified in the ITU-R Recommendation BT.500-11 [17] for a laboratory environment.

The subjective experiments were conducted in a room equipped with two 17" cathode ray tube (CRT) monitors of type Sony CPD-E200 (SE 1)

and a pair of 17" CRT monitors of type DELL and Samtron 75E (SE 2). The ratio of luminance of inactive screen to peak luminance was kept below a value of 0.02. The ratio of the luminance of the screen given it displays only black level in a dark room to the luminance when displaying peak white was approximately 0.01. The display brightness and contrast was set up with picture line-up generation equipment (PLUGE) according to Recommendations ITU-R BT.814 [14] and ITU-R BT.815 [15]. The calibration of the screens was performed with the calibration equipment ColorCAL from Cambridge Research System Ltd., England, while the DisplayMate software was used as pattern generator.

Due to its large impact on the artifact perceivability, the viewing distance must be taken into consideration when conducting a subjective experiment. The viewing distance is in the range of four times (4H) to six times (6H) the height H of the CRT monitors, as stated in Recommendation ITU-R BT.1129-2 [16]. The distance of 4H was selected here in order to provide better image details to the viewers.

3.2. *Test material*

Seven reference images of dimension 512×512 pixels and represented in gray scale have been chosen to cover a variety of textures, complexities, and arrangements. The images are shown in Fig. 3 and Fig. 4 where the images in Fig. 3 represent humans and human faces and the images in Fig. 4 represent more complex structures and natural scenes. The wireless link simulation model as explained in Section 2.1 has then been utilized to create test images that exhibit the wide variety of distortions as discussed in Section 2.2. In particular, two sets of forty images each, $\mathcal{I}_1$ and $\mathcal{I}_2$, were created to be used in the two subjective experiments SE 1 and SE 2, respectively. The images were chosen such as to cover a wide variety of artifacts and also a broad range of severities for each of the artifacts, from almost invisible to highly distorted. Thus, the metric design is based on a set of test images that incorporates distortions near the just noticeable differences regime to artifacts widely covering the suprathreshold regime.

Fig. 3. Reference images showing low texture human faces: "Lena", "Elaine" (top); "Tiffany", "Barbara" (bottom).

Fig. 4. Reference images showing complex textures: "Goldhill", "Pepper", and "Mandrill" (left to right).

3.3. *Viewers*

The viewers are the respondents in the experiment. Experienced viewers, i.e. individuals that are professionally involved in image quality evaluation/assessment at their work, are not eligible to participate in the subjective experiments. As such, only inexperienced (or non-expert) viewers were allowed to take part in the conducted subjective experiments. In order to support generalization of results and statistical significance of the collected subjective data, the experiments were conducted in two different laboratories involving 30 non-expert viewers in each experiment. Thus, the minimum requirement of at least 15 viewers, as recommended in [17], is well satisfied. In order to support consistency and eliminate systematic differences among results at the different testing laboratories, similar panels of test subjects in terms of occupational category, gender, and age were established. In particular, 25 males and 5 females, participated in SE 1. They were all university staff and students and their ages were distributed in the range of 21 to 39 years with the average age being 27 years. In the second experiment, SE 2, 24 males and 6 females participated. Again, they were all university staff and students and their ages were distributed in the range of 20 to 53 years with the average age being 27 years.

3.4. *Test procedure*

3.4.1. *Selection of test method*

Different test methodologies are provided in detail in [17] to best match the objectives and circumstances of the assessment problem. The methodologies are mainly classified into two categories, as double-stimulus and single-stimulus. In double-stimulus, the reference image is presented to the viewer along with the test image. On the other hand, in single-stimulus, the reference image is not explicitly presented and may be shown transparently to the subject for judgement consistency observation purpose. As we consider RR metric design in this paper, where partial information related to the reference image is available, we chose to deploy a double-stimulus method, the double-stimulus continuous quality scale (DSCQS). Moreover, DSCQS has been shown to have low sensitivity to contextual effects [17,35]. Contextual effects occur when the subjective rating of an image is influenced by presentation order and severity of impairments. This relates to the phenomenon that test subjects may tend to give an image a lower score than it might have normally been given if its presentation was scheduled after a less distorted image.

3.4.2. *Presentation of test material*

The test sessions were divided into two sections. Each section lasted up to 30 minutes and consisted of a stabilization and a test trial. The stabilization trials were used as a warm-up to the actual test trial in each section. In addition, one training trial was conducted at the very beginning of the test session to demonstrate the test procedure to the viewer and allow them to familiarize themselves with the test mechanism. Clearly, the scores obtained during the training and stabilization trials are not processed but only the scores given during the test trials are analyzed. In order to reduce the viewer's fatigue, a 15 minutes break was given between sections.

Given the DSCQS method, pairs of images A and B are presented in alternating order to the viewers for assessment, with one image being the original, undistorted image and the other being the distorted test image. As the DSCQS method is quite sensitive to small quality differences, it is well suited to not just cope with highly distorted test images but also with cases where the quality of original and distorted image is very similar.

3.4.3. *Grading scale*

The grading is performed with reference to a five-point quality scale (Excellent, Good, Fair, Poor, Bad), which is used to divide the continuous grading scale into five partitions of equal length. Given the pair of images A and B, the viewer is requested to assess their quality by placing a mark on each quality scale. As the reference and distorted image appear in pseudo random order, A and B may refer to either the reference image or the distorted image, depending on the actual arrangement of images in an assessment pair.

3.5. *Subjective data analysis*

The outcomes of the subjective experiments are discussed in the following by means of a statistical analysis. In this respect, a concise representation of the subjective data can be achieved by calculating conventional statistics such as the mean, variance, skewness, and kurtosis of the related distribution of opinion scores. The statistical analysis of this data reflects the fact that perceived quality is a subjective measure and hence may be described statistically.

3.5.1. *Statistical measures*

Let the MOS value for the k^{th} image in a set $\mathcal{K}$ of size K be denoted here as μ_k. Then, we have

$$\mu_k = \frac{1}{N}\sum_{j=1}^{N} u_{j,k} \tag{1}$$

where $u_{j,k}$ denotes the opinion score given by the j^{th} viewer to the k^{th} image and N is the number of viewers. The confidence interval associated with the MOS of each examined image is given by

$$[\mu_k - \epsilon_k, \mu_k + \epsilon_k] \tag{2}$$

The deviation term ϵ_k can be derived from the standard deviation σ_k and the number N of viewers and is given for a 95% confidence interval according to [17] by

$$\epsilon_k = 1.96\frac{\sigma_k}{\sqrt{N}} \tag{3}$$

where the standard deviation σ_k for the k^{th} image is defined as the square root of the variance

$$\sigma_k^2 = \frac{1}{N-1}\sum_{j=1}^{N}(u_{j,k} - \mu_k)^2 \tag{4}$$

The skewness measures the degree of asymmetry of data around the mean value of a distribution of samples and is defined by the second and third central moments m_2 and m_3, respectively, as

$$\beta = \frac{m_3}{m_2^{3/2}} \tag{5}$$

where the l^{th} central moment m_l is defined as

$$m_l = \frac{1}{N}\sum_{j=1}^{N}(u_j - \mu)^l \tag{6}$$

The peakedness of a distribution can be quantified by the kurtosis, which measures how outlier-prone a distribution is. The kurtosis is defined by the second and fourth central moments m_2 and m_4, respectively, as

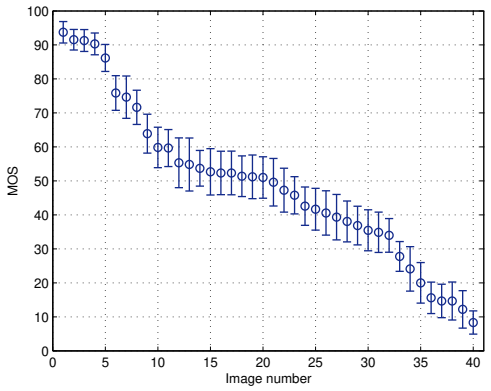
$$\gamma = \frac{m_4}{m_2^2} \tag{7}$$

It should be mentioned that the kurtosis of the normal distribution is obtained as 3. If the considered distribution is more outlier-prone than the normal distribution, it results in a kurtosis greater than 3. On the other hand, if it is less outlier-prone than the normal distribution, it gives a kurtosis less than 3. A distribution of scores is usually considered as normal if the kurtosis is between 2 and 4.

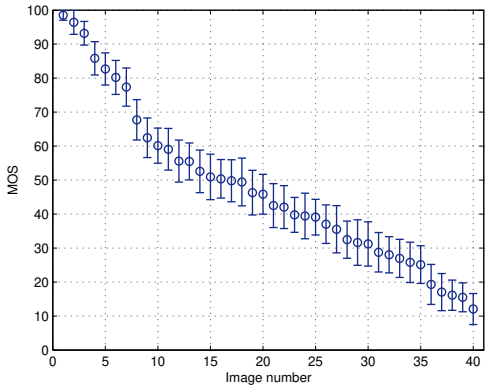
3.5.2. *Statistical analysis*

Figs. 5(a)-(b) show the scatter plots of MOS for SE 1 and SE 2, respectively. The forty images in each experiment are ordered with respect to decreasing subjective ratings in MOS. It can be seen from the figures that the material presented to the viewers resulted in a wide range of perceptual quality ratings indeed for both subjective experiments. As such, both experiments contained the extreme cases of excellent and bad image quality while the intermediate quality decreases approximately linearly in between. It is also observed that the spread of ratings around the MOS in terms of the 95% confidence interval is generally narrower for the images at the upper and lower end of the perceptual quality scale. Thus, the viewers seemed to be more confident with giving their quality ratings in case that the quality of the presented images was either of very high or very low quality. On the other hand, in the middle ranges of quality the confidence of viewers on the quality of an image was significantly lower.

Figs. 6(a)-(d) show the MOS, variance, skewness, and kurtosis, respectively, for each image sample



(a)



(b)

Fig. 5. Perceived quality ordered according to decreasing MOS with error bars indicating the 95% confidence intervals: (a) SE 1, (b) SE 2.

that was rated in the two subjective experiments. The image samples in all four figures are, as in Fig. 5, ordered with respect to decreasing MOS. In addition to the image samples the figures depict the related fits to these statistics, which reveal good agreement among the data for the two subjective experiments as the fits progress closely in the same manner over the ordered image samples. This indicates that the two experiments have been very well aligned with each other and also that the two viewer panels, even though originating from different countries, seem to have given similar quality scores for the test images they have been shown.

Fig. 6(a) depicts the impaired image samples with respect to decreasing MOS along with the linear fit through this data. It can be seen from the figure, that the linear fit for both experiments are very close indicating that the set of image samples used in the two independent experiments at WATRI and BIT

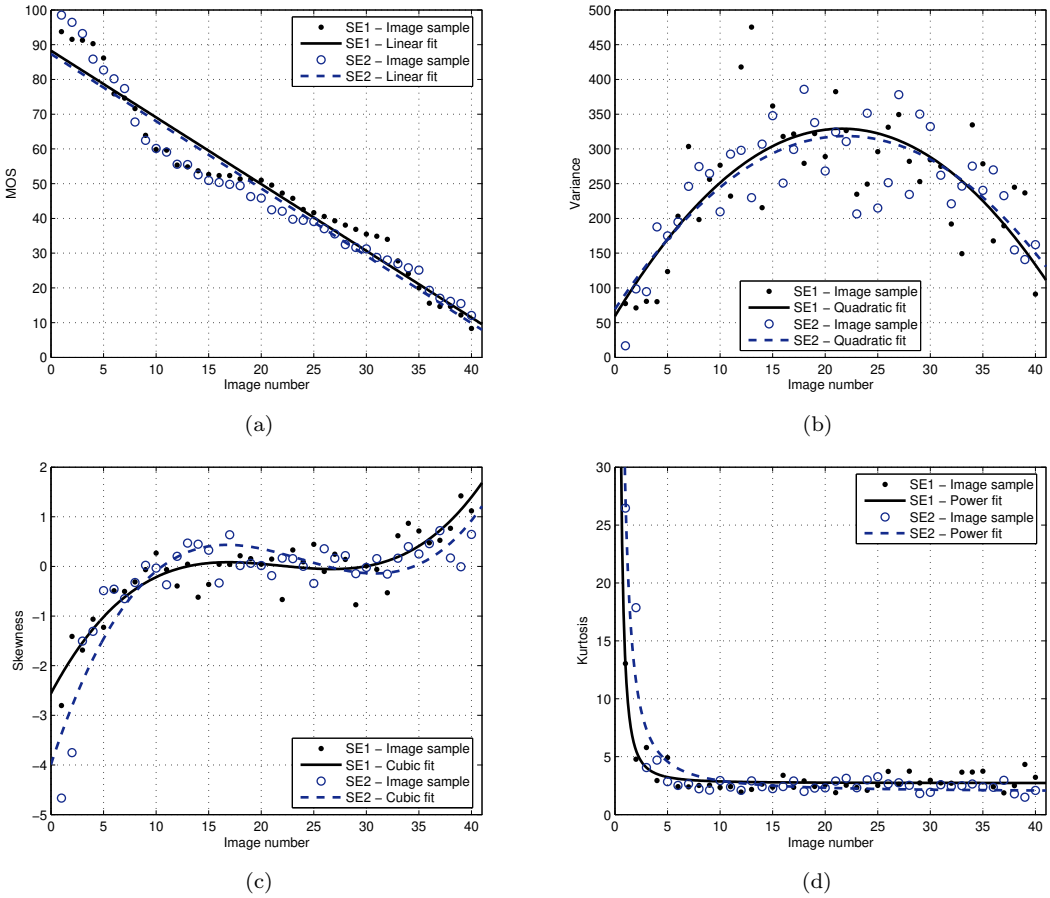


Fig. 6. Statistics of opinion scores for the impaired image samples: (a) MOS, (b) Variance, (c) Skewness, and (d) Kurtosis.

comprised of a similar range of quality impairments.

Fig. 6(b) shows the variance of all opinion scores for each image sample. The variance can be regarded as a measure of how much the viewers agree on the perceived quality of a certain image sample. In other words, the smaller the variance, the more pronounced the agreement between all viewers. It can clearly be seen from the figure that the variance is relatively small for images that have obtained either excellent or bad subjective quality ratings. In contrast, in the region where perceptual quality of the impaired images ranges between good and poor, the variance tends to be larger with the peak at about the middle of the quality range. This is an interesting result since it indicates that the viewers appear to be rather sure whether an image sample is of excellent or bad quality while opinions about images of average quality differ to a wider extent. These conclusions are supported by the confidence intervals shown in Fig. 5(a)-(b), which are narrower for images rated as being excellent and bad.

Fig. 6(c) shows the skewness of the opinion scores distribution for each image sample. In the context of the subjective ratings of image quality, a negative or positive skewness translate to the subjective scores being more spread towards lower or higher values than the MOS, respectively. For the images that were perceived as being of high quality, the negative skewness indicates that subjective scores tend to be asymmetrically spread around the MOS towards lower opinion scores and thus, that a number of viewers gave significantly lower quality scores as compared to the MOS. In the other extreme of image quality being perceived as bad, the positive skewness points to an asymmetrically spread around the MOS towards higher opinion scores. However, the positive skewness is not as distinct as the negative skewness at the high quality end, indicating that the agreement of low quality was higher as compared to the agreement about high quality. The asym-

Table 1
Image features, feature extraction algorithms, and related artifacts

Feature		Algorithm	Related artifact
$\tilde{f}_1$	Block boundary differences	Wang et al. [39]	Blocking
$\tilde{f}_2$	Edge smoothness	Marzilliano et al. [23]	Blur
$\tilde{f}_3$	Edge-based image activity	Saha et al. [31]	Ringing
$\tilde{f}_4$	Gradient-based image activity	Saha et al. [31]	Ringing
$\tilde{f}_5$	Image histogram statistics	Kusuma et al. [19]	Intensity masking, lost blocks

metry in subjective scores for the extreme cases of excellent and bad quality is thought to be due to the rating scale being limited to 100 and 0, respectively. As such, subjective scores have to approach the maximal and minimal possible rating from below or above, respectively. The skewness of around zero for the middle range of qualities reveals that the subjective scores seem to be symmetrically distributed with respect to MOS, even though the variance for images of average quality is larger.

Fig. 6(d) provides the kurtosis for each impaired image sample. It can be seen from the figure, that the distribution of subjective scores for some of the images scoring high MOS values in both experiments give kurtosis values much greater than of a normal distribution. This is a strong indication for outliers, meaning, that a few of the viewers gave the image quality a low rating whereas the majority of viewers agreed on a high image quality. With the progression of images towards decreasing MOS, the associated kurtosis fits quickly level out around the value 3, pointing to a normal distribution of the opinion scores around MOS. It is interesting to point out, that the high kurtosis in the high quality end does not occur at the bad quality end. This means that the entire viewer panel agreed on the bad quality images with no outlier scores being present. This result is also evident in the skewness distribution where the decline towards lower values at the high quality end is much more pronounced as compared to the incline of the skewness at the low quality end.

4. Objective structural degradation metrics

The design of the RR metrics proposed in this paper is based on the extraction of structural information from the images. In this section we will discuss the objective feature metrics that were deployed to measure the artifacts as observed in the test images (see Section 2.2). An analysis of the objective measures provides further insight into the feature metrics performance of quantifying the artifacts.

4.1. *Feature metrics*

Given the set of artifacts as observed in the test images, algorithms for feature extraction can be deployed to capture the amount by which each of the artifacts is present in the images. The selection of the algorithms to be used is driven by three constraints, namely, a reasonable accuracy in capturing the characteristics of the associated artifact, a representation of the feature that incurs low overhead in terms of reduced-reference (conserve bandwidth), and computational inexpensiveness (conserve battery power). The features and feature extraction algorithms deployed here to measure and quantify the presence of the related artifacts are listed in Table 1 and will be described in the following sections.

4.1.1. *Feature $\tilde{f}_1$: Block boundary differences*

The first feature metric $\tilde{f}_1$ is based on the algorithm by Wang et. al. [39] and comprises of three measures. The first measure, B, estimates blocking as average differences between block boundaries. Two image activity measures (IAM), A and Z, are applied as indirect means of quantifying blur. The former IAM computes absolute differences between in-block image samples and the latter IAM computes a zero-crossing rate. All three measures are computed in both horizontal and vertical direction and combined in a pooling stage as follows

$$\tilde{f}_1 = \alpha + \beta B^{\gamma_1} A^{\gamma_2} Z^{\gamma_3} \tag{8}$$

where the parameters α, β, γ_1, γ_2, and γ_3 were estimated in [39] using MOS from subjective experiments. Despite the two IAM incorporated in $\tilde{f}_1$, we found that this metric accounts particularly well for blocking artifacts in JPEG compressed images. This might be due to the magnitude of γ_1, being reported in [39] as relatively large compared to γ_2 and γ_3,

giving the blocking measure a higher impact on the metric $\tilde{f}_1$.

4.1.2. *Feature $\tilde{f}_2$: Edge smoothness*

The extraction of feature metric $\tilde{f}_2$ relates purely to measuring blur artifacts and follows the work of Marziliano et. al. [23]. It accounts for the smoothing effect of blur by measuring the distance between edges. It was found that it is sufficient to measure the blur along vertical edges, which allows for saving computational complexity as compared to computation on all edges. Therefore, a Sobel filter is applied to detect vertical edges in the image. The edge image is then horizontally scanned. For pixels that correspond to an edge point, the local extrema in the corresponding image are used to compute the edge width. The edge width then defines a local measure of blur. Finally, a global blur measure is obtained by averaging the local blur values over all edge locations. This metric was chosen to complement the IAM in $\tilde{f}_1$ since it does not just account for in-block blur but rather contributes a global blur measure.

4.1.3. *Features $\tilde{f}_3$ and $\tilde{f}_4$: Image activity*

Ringing artifacts are observed as periodic pseudo-edges around original edges, thus increasing the activity within an image. The feature metrics $\tilde{f}_3$ and $\tilde{f}_4$ provide an indirect means of measuring ringing artifacts and are based on two IAM by Saha and Vemuri [31].

Here, $\tilde{f}_3$ quantifies image activity (IA) based on normalized magnitudes of edges in an edge image $B(i)$ as follows

$$\tilde{f}_3 = \left(\frac{1}{M \times N} \sum_{i=1}^{M \times N} B(i) \right) \times 100 \qquad (9)$$

where M and N denote the image dimensions. Since $\tilde{f}_3$ does not depend on the direction of the edge, it also very well complements the blocking measure in $\tilde{f}_1$, which is purely designed to measure on the 8×8 block boundaries in JPEG coded images.

On the other hand, $\tilde{f}_4$ measures IA in an image $I(i,j)$ based on local gradients in both vertical and horizontal direction as follows

$$\tilde{f}_4 = \frac{1}{M \times N} \left(\sum_{i=1}^{M-1} \sum_{j=1}^{N} |I(i,j) - I(i+1,j)| + \sum_{i=1}^{M} \sum_{j=1}^{N-1} |I(i,j) - I(i,j+1)| \right) \qquad (10)$$

In [31], the IAM were evaluated and in particular $\tilde{f}_4$ has been found to quantify IA very accurately. We have further identified that both $\tilde{f}_3$ and $\tilde{f}_4$ account well for measuring ringing artifacts and also other high frequency changes within the image.

4.1.4. *Feature $\tilde{f}_5$: Image histogram statistics*

Finally, feature metric $\tilde{f}_5$ accounts for intensity masking and lost blocks using an original algorithm [19]. Both these artifacts cause an intensity shift in parts of an image or the whole image, which may result in either a darker or brighter appearance of the area as compared to the original image. As such we found that a simple computation of the standard deviation in the first-order image histogram provides an adequate measure of both intensity masking and lost blocks. We have thus adapted feature metric $\tilde{f}_5$ as follows

$$\tilde{f}_5 = \sqrt{\frac{1}{L} \sum_{i=0}^{L} (h_i - \overline{h})^2} \qquad (11)$$

where h_i denotes the number of pixels at grey level i, $\overline{h}$ denotes the mean grey level, and L is the maximum grey level of 255 when using 8 bits per pixel.

4.2. *Feature normalization*

The proposed NHIQM follows the design philosophy of our previous work that resulted in the Hybrid Image Quality Metric (HIQM) [18,19]. Although HIQM inherently uses feature relevance weights, the actual feature values $\tilde{f}_i$ have generally different meaning and different value ranges. As a consequence, it may be difficult to explore the resulting feature space for classification purposes and quality assessment if only relevance weighting was used as with HIQM. It is therefore suggested here to perform also an extreme value normalization to the features. This allows for a more convenient and meaningful comparison of the contribution of each normalized feature f_i to the overall metric, as they are then taken from the same value range as

$$0 \leq f_i \leq 1 \qquad (12)$$

Specifically, let us distinguish among I different image features. The related feature values $\tilde{f}_i$, $i = 1, \ldots, I$, shall be normalized as follows [26]:

$$f_i = \frac{\tilde{f}_i - \min\limits_{k=1,\ldots,K}(\tilde{f}_{i,k})}{\delta_i}, \qquad i = 1, \ldots, I \tag{13}$$

where the feature values $\tilde{f}_{i,k}, k = 1, \ldots, K$ are taken from a set $\mathcal{K}$ of size K. In our case, these features were extracted from the images used in the subjective experiments, including all reference images and test images. Furthermore, the normalization factor δ_i in (13) is given by

$$\delta_i = \max_{k=1,\ldots,K}(\tilde{f}_{i,k}) - \min_{k=1,\ldots,K}(\tilde{f}_{i,k}) \tag{14}$$

As far as the extreme value normalized features defined by (13) are concerned, it should be mentioned that the boundary conditions apply to those normalized feature values $f_{i,k}$ which are associated with the feature values $\tilde{f}_{i,k} \in \mathcal{K}$ of the images used in the experiments. In a practical system, it may also be beneficial to clip the normalized feature values that are actually calculated in a real-time wireless imaging application to fall in the interval $[0, 1]$ as well. For instance, severe signal fading in a wireless channel can result in significant image impairments at particular times causing the user-perceived quality to fall in a region where the HVS is saturated to notice further degradation.

4.3. *Feature metrics performance analysis*

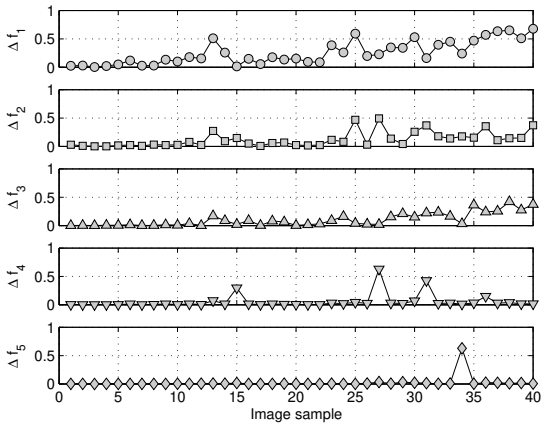
In order to gain deeper knowledge and understanding about the feature extraction, it is of interest to examine the extent to which different features are present in the stimuli and to quantify a relationship between the feature metrics and MOS. Given the context of RR metric design in wireless imaging, where we are interested in the difference between the quality of the received image as compared to the quality of the transmitted image, let us in the following consider the magnitude of normalized feature differences

$$\Delta f_i = |f_{t,i} - f_{r,i}|, \qquad i = 1, \ldots, 5 \tag{15}$$

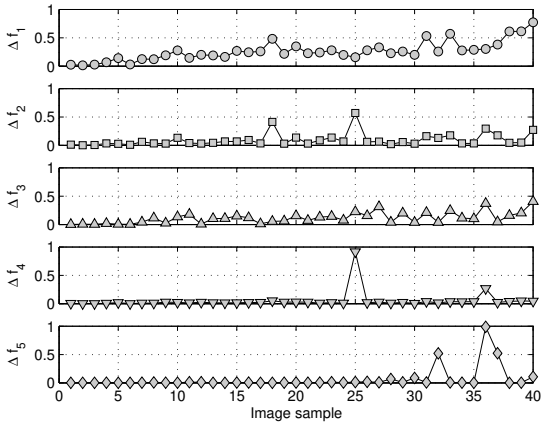
where $f_{t,i}$ and $f_{r,i}$ denote the i^{th} feature value of the transmitted and received image, respectively.

4.3.1. *Feature magnitudes over MOS*

Figs. 7(a)-(b) show the magnitudes of the normalized feature differences Δf_i for the image samples that were presented in SE 1 and SE 2. For each experiment, the related forty feature differences are ranked with respect to image samples of decreasing



(a)



(b)

Fig. 7. Magnitude of differences between normalized feature values for the considered image samples ranked according to decreasing MOS: (a) SE 1, (b) SE 2.

MOS. It can be seen from these figures that the wireless link scenario indeed inflicted all five features but with different degrees of severity. While for the image samples of high perceptual quality ratings, feature differences are almost absent, the feature differences tend to increase with decreasing MOS. Especially, the level of Δf_1, relating to blocking, contained in the image samples shows the widest spread and becomes more pronounced when progressing from images of excellent to bad perceptual quality. A similar behavior is observed for edge-based image activity Δf_3 but appears not as pronounced as for Δf_1. As far as the remaining three features are concerned, these become less prevalent for most of the images but large for some of the stimuli. In particular, gradient-based image activity Δf_4 and intensity

masking Δf_5 occur very distinctively with selected image samples while being almost absent from the majority of image samples.

4.3.2. *Feature statistics*

As with the MOS gathered from the subjective experiments, the statistical analysis may be extended to the actual feature differences in order to obtain a better understanding of the underlying objective quality degradations. However, overall statistics for the whole set of data, instead of image dedicated statistics, shall be presented hereafter. Accordingly, for all five feature differences Δf_i the mean, variance, skewness, and kurtosis have been computed over all images that have been shown in experiments SE 1 and SE 2 (see Fig. 7). The results of all statistics are presented for both experiments in Tables 2 and 3.

From comparison of the two tables one can observe that for all four statistics and for all five feature differences, the magnitudes of the values are very much in alignment between the two experiments SE 1 and SE 2. This indicates that the stimuli, in terms of the distorted test images, had similar characteristics in both experiments. Thus, not only subjective data is in alignment but also the composition of objective features among the test material. In particular, it can be seen from both tables that the mean of the blocking differences dominates over the other features. This is a direct result of the JPEG source encoding of which it is well known that blocking artifacts are dominant over other artifacts such as blur. The mean values of feature differences Δf_4 and Δf_5 are particularly small, however, these features exhibit instead a very high skewness and kurtosis as compared to the other features. Clearly, this quantifies the progression of feature differences in the stimuli as shown in Figs. 7(a)-(b) Δf_4 and Δf_5 being either negligibly small or distinctively developed.

4.3.3. *Feature cross-correlations*

Even though the feature metrics were selected to account for a particular artifact, one may expect some overlap in quantifying the different artifacts. To further understand the performance of the feature metrics in comparison to each other, Tables 4 and 5 show the Pearson linear correlation coefficient between each of the feature metrics for both SE 1 and SE 2. In this context, the cross-correlation measures the degree to which two features are simultaneously affected by a certain type and severity of an artifact. As expected, the correlation of a feature with itself exhibits the maximum magnitude of 1.

Table 2
Statistics of magnitudes of feature differences Δf_i for SE 1

	Δf_1	Δf_2	Δf_3	Δf_4	Δf_5
Mean	0.253	0.120	0.102	0.053	0.022
Variance	0.043	0.017	0.014	0.015	0.009
Skewness	0.627	1.425	1.124	3.518	6.015
Kurtosis	2.082	4.120	3.241	15.010	37.466

Table 3
Statistics of magnitudes of feature differences Δf_i for SE 2

	Δf_1	Δf_2	Δf_3	Δf_4	Δf_5
Mean	0.263	0.094	0.115	0.049	0.061
Variance	0.029	0.013	0.010	0.021	0.035
Skewness	1.066	2.495	1.072	5.461	3.785
Kurtosis	4.056	9.531	3.843	32.434	17.063

Table 4
Correlations between feature differences for SE 1

	Δf_1	Δf_2	Δf_3	Δf_4	Δf_5
Δf_1	1.000	0.625	0.821	0.016	0.027
Δf_2		1.000	0.440	0.649	0.112
Δf_3			1.000	0.056	−0.061
Δf_4				1.000	0.000
Δf_5					1.000

Table 5
Correlations between feature differences for SE 2

	Δf_1	Δf_2	Δf_3	Δf_4	Δf_5
Δf_1	1.000	0.376	0.640	−0.014	0.115
Δf_2		1.000	0.486	0.753	0.316
Δf_3			1.000	0.323	−0.272
Δf_4				1.000	0.170
Δf_5					1.000

It can be seen from the tables that the cross-correlations between the features vary strongly in their magnitudes. A particularly pronounced cross-correlation can be observed between feature metrics Δf_1 (block boundary differences) and Δf_3 (edge-based IA) for both SE 1 and SE 2. This is thought to be due to both metrics being based on measuring edges of an image. However, it should be noted again that feature metric Δf_1 only considers the 8×8 block borders of the JPEG encoding whereas fea-

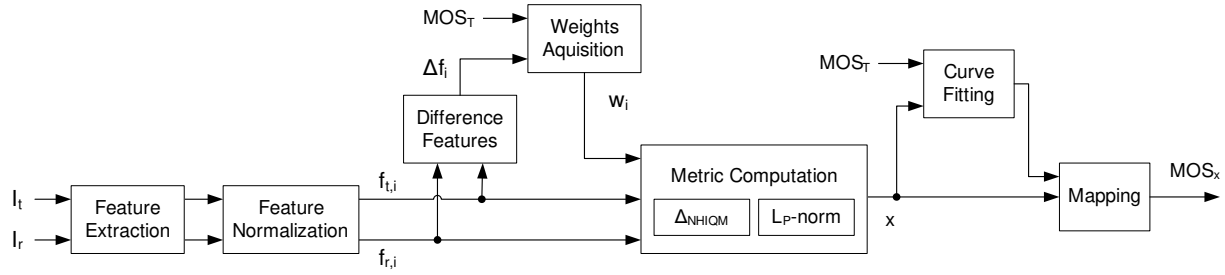


Fig. 8. Framework for designing feature-based objective perceptual image quality metrics.

ture metric Δf_3 quantifies image activity based on edges in all spatial locations and directions. Furthermore, feature metrics Δf_2 (edge smoothness) and Δf_4 (gradient-based IA) exhibit pronounced cross-correlations in the test sets of both experiments which may be a result of both metrics being designed to quantify smoothness in images based on gradient information. As for feature metric Δf_5 (image histogram statistics), it can be seen that this metric is only negligibly correlated to any of the other feature metrics. This is a highly desired result since the feature metrics other than Δf_5 should be widely unaffected by intensity shifts.

5. Objective perceptual metric design

In this section we will in detail describe the RR objective quality metric design. In this respect, the quality ratings obtained in the subjective experiments are instrumental for the transition from subjective to objective quality assessment.

5.1. *Metric training and validation*

As foundation of the metric design, the 80 images in $\mathcal{I}_1$ (SE 1) and $\mathcal{I}_2$ (SE 1) from the two experiments were organized into a training set $\mathcal{I}_T$ containing 60 images and a validation set $\mathcal{I}_V$ containing 20 images. For this purpose, 30 images were taken from $\mathcal{I}_1$ and 30 images from $\mathcal{I}_2$ to form $\mathcal{I}_T$ while the remaining 10 images of each set compose $\mathcal{I}_V$. Accordingly, a training set and a validation set were established with the corresponding MOS, here referred to as MOS_T and MOS_V. The training sets, $\mathcal{I}_T$ and MOS_T, are then used for the actual metric design. The validation sets, $\mathcal{I}_V$ and MOS_V, are used to evaluate the metrics ability to generalize to unknown images.

5.2. *Metric design framework*

A block diagram of the framework used in this paper to design RR objective perceptual image quality metrics is shown in Fig. 8. A brief overview of the design process is given in the sequel with reference to this figure.

The first key operation in the transition from subjective to objective perceptual image quality assessment is executed within the process of feature weights acquisition. As a prerequisite of weights acquisition, the different features of the transmitted and received image are extreme value normalized to allow for a meaningful weight association. As the RR design is focused on detecting distortions between related features, the weights acquisition is performed with respect to feature differences Δf_i, $i = 1, \ldots, 5$. Given the MOS values MOS_T for the images in the training set $\mathcal{I}_T$ and the related feature differences Δf_i for each image, correlation coefficients between subjective ratings and feature differences are computed as weights w_i, $i = 1, \ldots, 5$ to reveal the feature relevance to the subjectively perceived quality. It is then straightforward to compute a feature-based objective quality metric by applying a pooling function to condense the information to a single value x. Here, two metrics are proposed, namely Δ_{NHIQM} and the relevance weighted L_p-norm.

The second essential component in moving from subjective to objective quality assessment relates to the curve fitting block as shown in Fig. 8. Its inputs are the MOS values MOS_T for the images in the training set $\mathcal{I}_T$ and the values of the objective perceptual quality metric x for each of these images. The relationship between subjective quality given by MOS_T and objective quality represented by x, is then modeled by a suitable mapping function. The parameters of potential mapping functions can be obtained by using standard curve fitting techniques.

The selection of suitable mapping functions is typically based on both goodness of fit measures and visual inspection of the fitted curve. The obtained mapping function $f(x)$ can then be used to calculate predicted MOS values, MOS_x, for given values of the quality metric value x.

5.3. *Perceptual relevance of features*

The Pearson linear correlation coefficient r_P has been chosen to reveal the extent by which the individual feature differences contribute to the overall perception of image quality. In this sense, it captures prediction accuracy referring here to the ability of a feature difference to predict the subjective ratings with minimum average error. Given a set of K data pairs (u_k, v_k), this ability can be quantified by

$$r_P = \frac{\sum_{k=1}^{K}(u_k - \bar{u})(v_k - \bar{v})}{\sqrt{\sum_{k=1}^{K}(u_k - \bar{u})^2}\sqrt{\sum_{k=1}^{K}(v_k - \bar{v})^2}} \qquad (16)$$

where u_k and v_k are the feature difference and the subjective rating related to the k^{th} image, respectively, and $\bar{u}$ and $\bar{v}$ are the means of the respective data sets.

This choice is motivated by the fact that the correlation coefficient explicitly characterizes the association between two variables, which are given here by pairs of ratings and difference feature metrics. The sign of the correlation value may be neglected as it only represents the direction (increase/decrease) in which one variable changes with the change of the other variable. In view of the above, the absolute values of the Pearson linear correlation coefficients r_P are computed as the perceptual weights w_i of the related features. A higher correlation coefficient then corresponds to a feature that more significantly contributes to the overall quality as perceived by the viewer, while a lower correlation coefficient means less perceptual significance. Also, if the correlation coefficient approaches to the zero value, the relationship between the perceptual quality and the examined feature is not strongly developed.

Table 6 shows the values of the Pearson linear correlation coefficients, or feature weights, that were obtained for each of the five feature differences Δf_i, $i = 1, \ldots, 5$ for the training set when correlated to the associated MOS_T values. Accordingly, block boundary differences (Δf_1) appear to be the most relevant feature followed by edge-based image activity (Δf_3), edge smoothness (Δf_2), image histogram statistics (Δf_5), and gradient-based image activity (Δf_4). This relates to blocking being the most annoying artifact followed by ringing due to edge-based image activity, blur, intensity masking, and ringing due to gradient-based image activity. Similar findings have also been made by Farias et al. [9] who observed that blocking is more annoying than blur. The same group also found [10] that ringing is the least annoying artifact. This agrees with our feature metric Δf_4 which also received the smallest weight. On the other hand, the feature metric Δf_3 deployed in our paper measures ringing as well but received a higher weight. We believe that this outcome can be related to Δf_3 having a strong correlation with Δf_1 (see Tables 4 and 5), thus not only accounting for ringing but also for blocking artifacts.

Table 6
Perceptual relevance weights of feature differences Δf_i for the images in the training set

Metric	Weight	Value
Δf_1	w_1	0.819
Δf_2	w_2	0.413
Δf_3	w_3	0.751
Δf_4	w_4	0.182
Δf_5	w_5	0.385

It should be noted here that the relevance weights in Table 6 were obtained for the particular case of JPEG source encoding where blocking artifacts are predominant over other artifacts such as blur. This may also contribute to the higher correlation weights for the edge based features Δf_1 and Δf_3 as compared to the gradient based features Δf_2 and Δf_4. Hence, the relevance weights may not be purely related to the perceptual relevance but also to the particular artifacts that are observed in the visual content. As such, one may obtain different relevance weights in case of other source encoders, such as JPEG2000.

5.4. *RR objective metric computation*

In the following two sections, we will consider two different pooling functions that are based on weighted combinations of the feature metrics. Firstly, we introduce NHIQM, which linearly combines extreme value normalized image features to

a single quality value. Secondly, a perceptual relevance weighted version of the L_p-norm is proposed, which calculates a weighted sum of image feature differences between original and impaired image. In both cases, the respective image features are extracted with the metrics as summarized in Section 4.1, while the actual weights used for feature combination have been deduced as discussed in Section 5.3.

5.4.1. *Normalized hybrid image quality metric*

The proposed NHIQM is defined as a weighted sum of the extreme value normalized features as

$$NHIQM = \sum_{i=1}^{I} w_i f_i \tag{17}$$

where w_i denotes the relevance weight of the associated feature f_i. Clearly, this RR metric is particularly beneficial for objective perceptual quality assessment in wireless imaging, as the reduced-reference is represented by only one single value for a given image. Accordingly, NHIQM can be communicated from the transmitter to the receiver whilst imposing very little stress on the bandwidth resources.

Regarding applications in wireless imaging, NHIQM can be calculated for the transmitted image I_t and received image I_r, resulting in the corresponding values $NHIQM_t$ and $NHIQM_r$ at the transmitter and receiver, respectively. Provided that the $NHIQM_t$ value is communicated to the receiver, structural differences between the images at both ends may simply be represented by the absolute difference

$$\Delta_{NHIQM} = |NHIQM_t - NHIQM_r| \tag{18}$$

5.4.2. *Perceptual relevance weighted L_p-norm*

The L_p-norm, also referred to as Minkowski metric, is a distance measure commonly used to quantify similarity between two signals or vectors. In image processing it has been applied, for instance, with the percentage scaling method [29] and the combining of impairments in digital image coding [28].

In this paper, we incorporate the relevance weighting for the extreme value normalized features into the calculation of the L_p-norm. This modification of the L_p-norm shall be defined as follows:

$$L_p = \left[\sum_{i=1}^{I} w_i^p |f_{t,i} - f_{r,i}|^p \right]^{\frac{1}{p}} \tag{19}$$

where $f_{t,i}$ and $f_{r,i}$ denote the i^{th} feature value of the transmitted and the received image, respectively.

The Minkowski exponent p may be determined experimentally [29]. Alternatively, the Minkowski exponent p may be assigned a fixed value. In both cases, a higher value of p increases the impact of the dominant features on the overall metric. In the limit of p approaching infinity, we obtain

$$L_\infty = \max_{i=1,\cdots,I} |f_{t,i} - f_{r,i}| \tag{20}$$

meaning that the largest absolute feature value difference solely dominates the norm. We have found [7] that values beyond $p = 2$ do not improve the quality prediction performance of the modified L_p-norm given in (19). We believe that this characteristic is because of the perceptual relevance weights obtained for each feature inherently accounting for the dominance of the particular features. In the sequel, we therefore consider the modified L_p-norm for Minkowski exponents of $p = 1$ and $p = 2$ only.

Although the L_p-norm belongs to the class of RR metrics, it requires more transmission resources compared to Δ_{NHIQM}, as all feature values need to be communicated from the transmitter to the receiver. On the other hand, the information about each of the feature degradations may provide further insights into the channel induced distortions. Hence, overhead may be traded off at the expense of a reduction about structural degradation information by neglecting feature metrics that received low perceptual relevance weights.

5.5. *Mapping functions*

Due to non-linear quality processing in the HVS, artifacts and quality do not follow a linear relationship. To account for this phenomenon, a mapping function is applied to the quality metrics. In general, an objective quality metric x may be mapped using a non-linear mapping function $f(x)$. The mapping function may then be used to determine the predicted mean opinion score MOS_x for a given x as

$$MOS_x = f(x) \tag{21}$$

Specifically, we will consider the following three classes of mapping functions:

$$MOS_x \triangleq \begin{cases} \sum_{j=0}^{m} p_j x^j & \text{Polynomial} \\ \sum_{j=0}^{m} a_j\, e^{b_j x} & \text{Exponential} \\ \frac{100}{1+e^{-l_1(x-l_2)}} & \text{Logistic} \end{cases} \quad (22)$$

where the coefficients $p_0, \ldots, p_m$ of the polynomial function of degree m, the initial value $a_1 \ldots, a_m$ and growth/decay $b_1 \ldots, b_m$ of the exponential function of order m, and the parameters l_1, l_2 of the logistic function are to be determined through curve fitting based on the given experimental data from the training set.

These three classes of mapping functions have been chosen as candidates for quality prediction due to the following reasons:

- **Polynomial functions** provide sufficient flexibility to support simple empirical prediction.
- **Exponential functions** are imposed to enable a good fit to experimental data over the middle-to-upper range of the quality impairment measure [21] and may be less prone to overfitting compared to functions with many parameters.
- **Logistic functions** facilitate the mapping of quality impairment measures into a finite interval. They produce scale compressions at the high and low extremes of quality while progressing approximately linear in the range between these extremes.

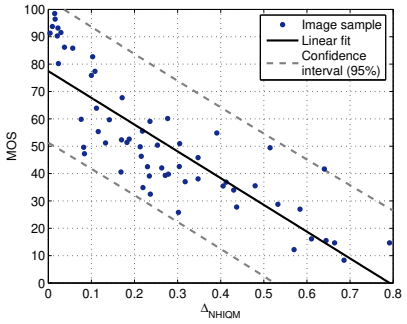
Standard curve fitting techniques have been used to deduce the parameters of the mapping functions that mathematically describe best the relationship between subjective ratings and objective perceptual quality metric with respect to a given goodness of fit measure. A mapping function obtained in this way translates an objective perceptual quality metric x into predicted MOS, MOS_x. The goodness of fit between MOS and predicted MOS, can be specified by either of the following statistics:

- $\mathbf{R^2}$ captures the degree by which variations in the MOS values are accounted for by the fit. It can assume any value in the interval $[0, 1]$ with a good fit being close to 1.
- **Root mean squared error (RMSE)** is referred to as the standard error of the fit with a good fit indicated by an RMSE value close to 0.
- **Sum of squares due to error (SSE)** represents the total deviation between predicted MOS and MOS from the experiments. The smaller the SSE value, the better the fit.

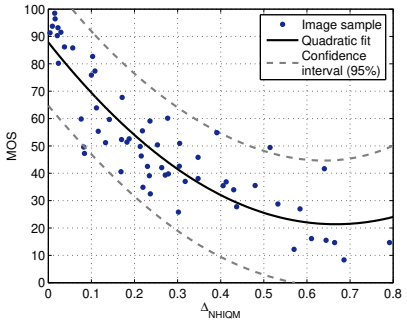
Table 7
Mapping functions $f(x) = MOS_{NHIQM}$, $x = \Delta_{NHIQM}$ and their goodness of fit.

Polynomial	Parameters	R^2	RMSE	SSE
p_1x+p_0	$p_1 = -97.8$	0.71	12.78	9472
	$p_0 = 77.45$			
$p_2x^2+p_1x+p_0$	$p_2 = 149.5$	0.79	11.07	6982
	$p_1 = -199.4$			
	$p_0 = 87.88$			
$p_3x^3+p_2x^2+p_1x+p_0$	$p_3 = -493.9$	0.82	10.17	5792
	$p_2 = 672.2$			
	$p_1 = -338.3$			
	$p_0 = 94.87$			
Exponential function				
$a_1e^{b_1x}$	$a_1 = 88.79$	0.79	10.76	6714
	$b_1 = -2.484$			
$a_1e^{b_1x}+a_2e^{b_2x}$	$a_1 = 69.76$	0.83	10.01	5612
	$b_1 = -1.719$			
	$a_2 = 32.05$			
	$b_2 = -17.39$			
$a_1e^{b_1x}+a_2e^{b_2x}+a_3e^{b_3x}$	$a_1 = 63.18$	0.80	11.12	6678
	$b_1 = -3.056$			
	$a_2 = -175$			
	$b_2 = 0.1434$			
	$a_3 = 198.2$			
	$b_3 = 0.041$			
Logistic function				
$100/[1+e^{-l_1(x-l_2)}]$	$l_1 = -4.613$	0.72	12.63	9263
	$l_2 = 0.262$			

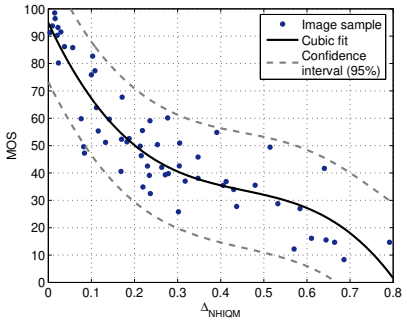
The Matlab Curve Fitting Toolbox was used to find the parameters of the considered mapping functions. The mapping functions have been derived for both Δ_{NHIQM} and the relevance weighted L_p-norm, however, only the results for Δ_{NHIQM} will be presented in the following. The results are provided in Table 7 along with the different goodness of fit measures. A visual examination of the fitted mapping functions is supported by the Figs. 9-11, which also show the 95% confidence interval for each fit.



(a)

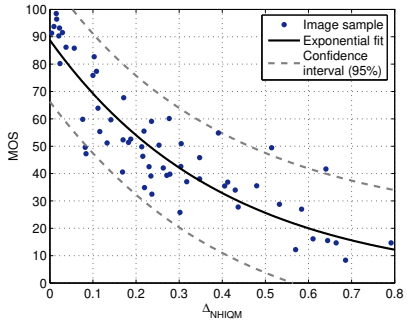


(b)

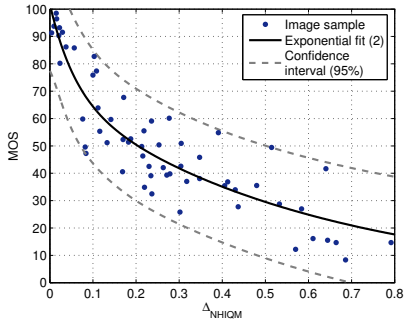


(c)

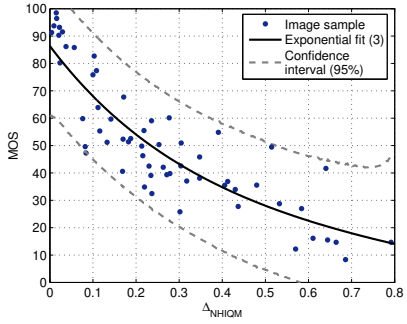
Fig. 9. Polynomial mapping functions: (a) Linear, (b) Quadratic, (c) Cubic.



(a)



(b)



(c)

Fig. 10. Exponential mapping functions: (a) Exponential, (b) Double exponential, (c) Triple exponential.

As far as the polynomial functions are concerned, it could be concluded at first sight from looking only at the goodness of fit statistics that the cubic polynomial would perform similarly favorable in perceptual quality prediction as the exponential functions. However, visual inspection of Fig. 9 suggests the opposite as the good fit applies only for the given data range but tends to diverge outside this range. For example, an increase of the objective perceptual quality metric beyond the value of 0.8 would actually predict "negative" MOS values (see Fig. 9 (c)). As higher-degree polynomials may even result in more severe overfitting, the class of polynomials has little to offer for use in objective perceptual quality assessment.

In contrast to the polynomial functions, favorable fitting has been obtained for all three considered exponential mapping functions, not only in terms of goodness of fit measures but also confirmed by visual inspection (see Fig. 10). However, it can be observed that the triple exponential function performs similarly to the exponential function but at the price of a larger computational complexity due to its more involved analytical expression. As such, the triple exponential function may not be considered further.

As for the logistic mapping function, the good-

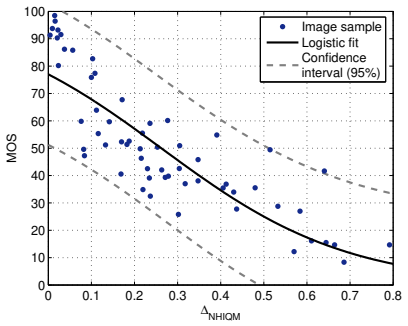


Fig. 11. Logistic mapping function.

ness of fit measures indicate a rather poor fit to the data from the subjective experiments. Especially, the compression at the high end of the quality scale produces disagreement with MOS (see Fig. 11).

In light of the above findings, both the exponential and double exponential mapping are selected for further consideration in the metric design.

6. Evaluation of image quality metrics

With the exponential and double exponential mapping being identified as suitable models for objectively predicting perceptual image quality, an evaluation of the prediction performance of Δ_{NHIQM} and L_p-norm on the training set (60 images in $\mathcal{I}_T$ and related MOS_T) and its generalization to the validation set (20 images in $\mathcal{I}_V$ and related MOS_V) is given in this section.

6.1. *Other objective quality metrics for comparison*

We have selected contemporary quality metrics that have been proposed in recent years to allow for a performance comparison with the proposed feature-based Δ_{NHIQM} and the L_p-norm. Specifically, the reduced-reference image quality assessment (RRIQA) technique proposed in [41] is chosen as a prominent member of the class of RR metrics. In addition, the structural similarity (SSIM) index [40], the visual information fidelity (VIF) criterion [32], the visual signal-to-noise ratio (VSNR) [4], and the peak signal-to-noise ratio (PSNR) [26] are chosen as the FR metrics. It is noted that FR metrics would not be suitable for the considered wireless imaging scenario but rather serve to benchmark prediction performance, which can be expected to be high due to the utilization of the reference image.

- **RRIQA:** This metric [41] is based on a natural image statistic model in the wavelet domain. The image distortion measure is obtained from the estimation of the Kullback-Leibler distance between the marginal probability densities of wavelet coefficients in the subbands of the reference and distorted images as follows

$$\mathrm{D} = \log_2\left(1 + \frac{1}{D_0}\sum_{k=1}^{K}|\hat{d}^k(p^k\|q^k)|\right) \tag{23}$$

  where the constant D_0 is used as a scaler of the distortion measure, $\hat{d}^k(p^k\|q^k)$ denotes the estimation of the Kullback-Leibler distance between the probability density functions p^k and q^k of the k^{th} subband in the transmitted and received image, and K is the number of subbands. The overhead needed to represent the reduced-reference is given as 162 bits [41].
- **SSIM:** The SSIM index [40] is based on the assumption that the HVS is highly adapted to the extraction of structural information from the visual scene. As such, SSIM predicts structural degradations between two images based on simple intensity and contrast measures. The final SSIM index is given by

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{24}$$

  where μ_x, μ_y and σ_x, σ_y denote the mean intensity and contrast of image signals x and y, respectively. The constants C_1 and C_2 are used to avoid instabilities in the structural similarity comparison that may occur for certain mean intensity and contrast combinations ($\mu_x^2+\mu_y^2 = 0, \sigma_x^2+\sigma_y^2 = 0$).
- **VIF:** The VIF criterion [32] approaches the image quality assessment problem from an information theoretical point of view. In particular, the degradation of visual quality due to a distortion process is measured by quantifying the information available in a reference image and the amount of this reference information that can be still extracted from the test image. As such, the VIF criterion measures the loss of information between two images. For this purpose, natural scene statistics and, in particular, Gaussian scale mixtures (GSM) in the wavelet domain, are used to model the images. The proposed VIF metric is given by

$$\mathrm{VIF} = \frac{\sum_{j \in subbands} I(\overrightarrow{C}^{N,j}; \overrightarrow{F}^{N,j} | s^{N,j})}{\sum_{j \in subbands} I(\overrightarrow{C}^{N,j}; \overrightarrow{E}^{N,j} | s^{N,j})} \quad (25)$$

where $\overrightarrow{C}$ denotes the GSM, N denotes the number of GSM used, and $\overrightarrow{E}$ and $\overrightarrow{F}$ denote the visual output of a HVS model, respectively, for the reference and test image.

- **VSNR:** The VSNR [4] metric deploys a two-stage approach based on near-threshold and suprathreshold properties of the HVS to quantify image fidelity. The first stage determines whether distortions are visible in an image. For this purpose, contrast thresholds for distortion detection are determined using wavelet-based models of visual masking. If the distortions are below the threshold, the quality of the image is assumed to be perfect and the algorithm is terminated. If the distortions are visible, a second stage implements perceived contrast and global precedence properties of the HVS to determine the impact of the distortions on perceived quality. The final VSNR metric is then given as

$$\mathrm{VSNR} = 20 \log_{10} \left(\frac{C(\mathbf{I})}{\alpha \, d_{pc} + (1-\alpha) \frac{d_{gp}}{\sqrt{2}}} \right) \quad (26)$$

where $C(\mathbf{I})$ denotes the root-mean-squared contrast of the original image $\mathbf{I}$, d_{pc} and d_{gp} are, respectively, measures of perceived contrast and global precedence disruption, and α is a weight regulating the relative contributions of d_{pc} and d_{gp}.

- **PSNR:** Image fidelity is an indication about the similarity between the reference and distorted images and measures pixel-by-pixel closeness between those pairs. The PSNR [26] is the most commonly used fidelity metric. It measures the fidelity difference of two image signals $I_R(x, y)$ and $I_D(x, y)$ on a pixel-by-pixel basis as

$$\mathrm{PSNR} = 10 \log \frac{\eta^2}{\mathrm{MSE}} \quad (27)$$

where η is the maximum pixel value, here 255. The mean square error is given as

$$\mathrm{MSE} = \frac{1}{XY} \sum_{x=1}^{X} \sum_{y=1}^{Y} [I_R(x, y) - I_D(x, y)]^2 \quad (28)$$

where X and Y denote horizontal and vertical image dimensions, respectively. Despite being an FR metric, PSNR usually does not correlate well with the visual quality as perceived by a human observer [38].

Table 8
Computational complexity of the metrics and amount of reference information needed.

Metric		Computation	Reference
Type	Name	time/image	information
RR	Δ_{NHIQM}	1.55 sec	17 bits
	L_p-norm	1.55 sec	85 bits
	RRIQA	7.12 sec	162 bits
FR	SSIM	0.37 sec	Full image
	VIF	0.92 sec	Full image
	VSNR	0.33 sec	Full image
	PSNR	0.05 sec	Full image

6.2. *Computational complexity and amount of reference information*

In the following, we will discuss the computational complexity of the considered metrics and the amount of reference information that is needed in order to assess the quality of a test image. The details are summarized in Table 8.

The computational complexity is measured in terms of the time that each of the metrics needs to assess the quality of a single image in our sets $\mathcal{I}_1$ and $\mathcal{I}_2$. Here, we have computed each metric over all 80 images and then determined the average time. The metrics were run on a laptop computer containing an Intel T2600 Dual Core processor with 2.16GHz and 4GB of RAM. In order to allow for a fair comparison, the publicly available Matlab implementation of each metric was used even though there may be other implementations available for some of the metrics. It can be seen from Table 8 that the computational complexity of all FR metrics is lower as compared to the RR metrics. Amongst the FR metrics, PSNR outperforms by far the other considered metrics in terms of computational complexity. Regarding the RR metrics, it is observed that both Δ_{NHIQM} and L_p-norm are significantly less complex than RRIQA.

In the context of wireless imaging, the amount of reference information needed for quality assessment determines the overhead of data that needs to be transmitted over the channel along with the actual image. From Table 8 one can see that the reference information is significantly lower for both Δ_{NHIQM}

and L_p-norm as compared to RRIQA. The particularly small reference information for Δ_{NHIQM} results from the fact that only a single value $NHIQM_t$ needs to be transmitted. On the other hand, with the L_p-norm five features need to be transmitted resulting in a five times higher overhead. However, as discussed in Section 5.4.2, the number of features used may be traded off with the transmission overhead by neglecting features of low perceptual relevance. As for the FR metrics, the reference image is needed for the quality assessment and as such, the size of the image determines the amount of reference information. Independent of the image size, however, the amount of reference information would be magnitudes higher as compared to the RR metrics.

6.3. *Prediction performance measures*

The quality prediction performance of the considered objective metrics will be quantified in terms of accuracy, monotonicity, and consistency as recommended by the Video Quality Experts Group (VQEG) [35].

The prediction accuracy of each objective quality metric will be quantified using the Pearson linear correlation coefficient as defined in (16). The prediction monotonicity will be measured by the non-parametric Spearman rank order coefficient

$$r_S = \frac{\sum_{k=1}^{K}(\chi_k - \bar{\chi})(\gamma_k - \bar{\gamma})}{\sqrt{\sum_{k=1}^{K}(\chi_k - \bar{\chi})^2}\sqrt{\sum_{k=1}^{K}(\gamma_k - \bar{\gamma})^2}} \tag{29}$$

where χ_k and γ_k denote the ranks of the predicted scores and the subjective scores, respectively, and $\bar{\chi}$ and $\bar{\gamma}$ are the midranks of the respective data sets. This measure is used to quantify if changes (increase or decrease) in one variable is followed by changes (increase or decrease) in another variable, irrespective of the magnitude of the changes.

The prediction consistency is identified by the outlier ratio. A data pair (u_k, v_k) may be declared as an outlier when the absolute difference between u_k and v_k is greater than a certain threshold. As suggested in [35], the threshold shall be chosen at least twice as much as the MOS standard deviation σ_{v_k} such that

$$|u_k - v_k| > 2\sigma_{v_k} \tag{30}$$

Then, the outlier ratio can be calculated as

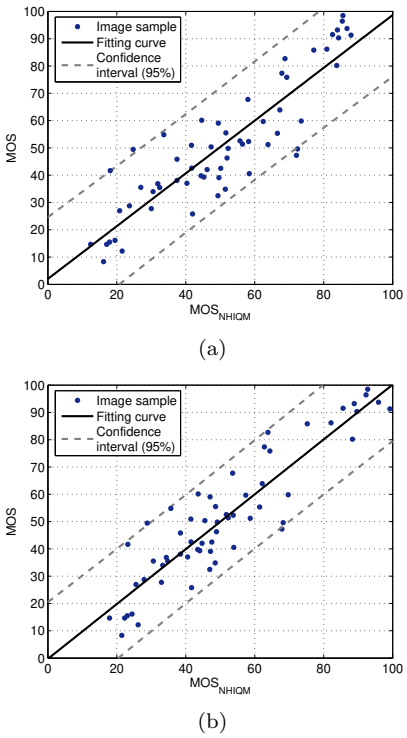


Fig. 12. MOS versus predicted MOS, MOS_{NHIQM}: (a) Exponential mapping, (b) Double exponential mapping.

$$r_O = R_O/R \tag{31}$$

where R_O denotes the total number of outliers and R is the size of the data set.

6.4. *Linear regression*

Prior to the evaluation of prediction performance for the considered objective image quality metrics, the favorable mapping functions will be used to relate the predicted MOS values to the actual MOS values from the subjective experiments. The predicted scores, MOS_T and MOS_V, respectively, are calculated for each image in the training set $\mathcal{I}_T$ and the validation set $\mathcal{I}_V$.

As an example, Fig. 12 shows the result for Δ_{NHIQM} using the exponential and double exponential mapping functions. Here, the MOS values from the subjective experiments are plotted versus the predicted MOS values, MOS_{NHIQM}, for the images in the training set $\mathcal{I}_T$. In addition, a linear function has been fitted to the data set and is presented along with the 95% confidence interval. It should be mentioned that the fitting curves for both exponential and double exponential mapping

Table 9
Parameters of prediction functions for objective quality metrics using exponential and double exponential mapping

Metric		Exponential		Double exponential			
Type	Name	a_1	b_1	a_1	b_1	a_2	b_2
RR	Δ_{NHIQM}	88.79	-2.484	69.76	-1.720	32.04	-17.39
	L_1-norm	87.63	-1.840	68.15	-1.251	34.31	-13.46
	L_2-norm	90.20	-2.820	69.83	-1.950	32.28	-16.24
	RRIQA	102.1	-0.160	101.3	-0.157	$-3.486 \cdot 10^{-14}$	3.701
FR	SSIM	13.91	1.715	31.34	0.446	$3.964 \cdot 10^{-7}$	18.58
	VIF	4.291	2.886	14.06	1.366	$7.721 \cdot 10^{-14}$	33.98
	VSNR	25.662	0.033	21.964	0.041	$-1.388 \cdot 10^{-6}$	0.387
	PSNR	18.33	0.036	22.68	0.029	0	0.029

produce the desired linear relationship between predicted MOS and MOS. Specifically, the range between 0 and 100 is nicely captured for predicted MOS and MOS. The prediction performance measures will be calculated for these post-mapped relationships in addition to the actual metric values.

6.5. *Analysis of mapping parameters*

The evaluation of the prediction performance of Δ_{NHIQM}, L_1-norm, and L_2-norm will be presented here and compared to RRIQA, SSIM, VIF, VSNR, and PSNR. For this purpose, the parameters of the exponential and double exponential mapping functions have been derived for all of these metrics, following the methodology as outlined in Section 5.5.

Table 9 presents the parameters of the prediction functions deduced from curve fitting of the considered quality metrics to the MOS values in the training set of images using the exponential and double exponential mapping. It can be seen from the numerical values of the parameters that distinct exponential mapping functions are produced in terms of growth and decay. The negative decay parameter for the feature-based objective perceptual quality metrics Δ_{NHIQM}, L_1-norm, L_2-norm, as well as RRIQA relate to the fact that these RR metrics represent image degradation. Thus, larger values of these metrics correspond to lower perceptual quality. In contrast, the FR metrics SSIM, VIF, VSNR, and PSNR measure image similarity of some sort, which is represented by the positive decay parameter in their exponential mapping functions. In these cases, a larger metric value corresponds to higher perceptual quality. As for the double exponential mapping functions, these are pronounced only for the feature-based objective perceptual quality metrics of Δ_{NHIQM} and the L_1- and L_2-norm. Specifically, the growth and decay parameters for both involved exponential functions are substantially different to zero. This is not the case for the other considered quality metrics, RRIQA SSIM, VIF, and VSNR, with the double exponential mapping functions degenerating to an exponential function for small metric values. Due to the initial value a_2 being close to zero, the second exponential function can contribute to the prediction only for extremely large metric values although this may still be insignificant to the first exponential function involved. In the case of PSNR, the initial value of the second exponential function is actually obtained from the curve fitting as being zero. Accordingly, the double exponential mapping function in fact degenerates to an exponential mapping function.

6.6. *Evaluation of quality prediction performance*

Given the parameters of the prediction functions for the examined quality metrics, the prediction performance of these metrics is presented in Table 10. In particular, the prediction accuracy is quantified by the Pearson linear correlation coefficient. It has been calculated on the basis of the 60 images in the training set $\mathcal{I}_T$ and the 20 images of the validation set $\mathcal{I}_V$. Moreover, prediction accuracy has been calculated for the relationship between MOS and the pure metric as well as for the relationship between MOS and predicted MOS using exponential mapping and double exponential mapping.

Table 10
Prediction performance of objective quality metrics, predicted MOS using exponential mapping, and predicted MOS using double exponential mapping

Metric		Accuracy						Monotonicity		Consistency			
Type	Name	Metric		Exponential		2-exponential		Metric, Mapping		Exponential		2-exponential	
		$r_{P,T}$	$r_{P,V}$	$r_{P,T}$	$r_{P,V}$	$r_{P,T}$	$r_{P,V}$	$r_{S,T}$	$r_{S,V}$	$r_{O,T}$	$r_{O,V}$	$r_{O,T}$	$r_{O,V}$
RR	Δ_{NHIQM}	0.843	0.840	0.892	0.888	0.910	0.860	0.867	0.892	0.017	0	0	0.050
	L_1-norm	0.833	0.841	0.873	0.897	0.895	0.893	0.854	0.901	0.017	0	0.017	0
	L_2-norm	0.845	0.846	0.888	0.884	0.903	0.878	0.875	0.890	0.017	0	0	0
	RRIQA	0.821	0.772	0.829	0.749	0.831	0.752	0.786	0.758	0.050	0.050	0.050	0.050
FR	SSIM	0.582	0.434	0.632	0.511	0.701	0.605	0.558	0.347	0.117	0.050	0.100	0.050
	VIF	0.713	0.737	0.789	0.788	0.877	0.795	0.813	0.729	0.083	0	0.033	0
	VSNR	0.766	0.696	0.758	0.686	0.783	0.686	0.686	0.510	0.083	0	0.050	0.050
	PSNR	0.742	0.712	0.738	0.709	0.741	0.711	0.638	0.615	0.100	0	0.150	0

As can be seen from the numerical results in Table 10 for the metric training, the prediction accuracy of the feature-based metrics, Δ_{NHIQM}, L_1-norm, and L_2-norm, outperform the other considered metrics, RRIQA, SSIM, VIF, VSNR, and PSNR. This applies for the training with respect to all three cases, i.e. the pure metric prior to mapping and after mapping with exponential and double exponential functions. The comparison between the feature-based quality metrics indicate the comparable performance of Δ_{NHIQM} and the L_p-norms.

Similar observations about accuracy can be made for metric validation. In terms of metric generalization to these unknown images from the validation set, the feature-based quality metrics significantly outperform the other considered metrics in accuracy. While Δ_{NHIQM} and the L_p-norms provide an accuracy over 80% and in some cases close to 90%, all other considered metrics fall below the 80% threshold of generalization accuracy. It is also observed that the largest accuracy being $r_p = 0.91$ for Δ_{NHIQM} on the training set using double exponential mapping does not generalize as well as for the pure metric or exponential mapping. This indicates that fitting Δ_{NHIQM} to a double exponential mapping may already produce some degree of overfitting. Similar trends to overfitting using double exponential mapping can be observed with the L_2-norm and VIF. In view of this and the degeneration of double exponential mapping to exponential mapping with some metrics, the exponential function may in fact constitute the most preferred mapping in the considered context of wireless imaging.

Let us now compare the prediction monotonicity of the proposed image quality metrics with the other state of the art image quality metrics. As all relationships follow strictly decreasing or increasing functions, differentiation between metric, exponential, and double exponential mapping is not required as ranks are kept the same for all three cases. The results shown in Table 10 reveal that the feature-based Δ_{NHIQM} approach and the L_p-norms perform favorably over the remaining four metrics with prediction monotonicity well above 80% for both metric training and validation. From the other metrics, only VIF shows a satisfactory prediction monotonicity of $r_S = 0.813$ for the training but does not generalize well to the unknown images.

Finally, the prediction consistency for the training of both feature-based metrics, Δ_{NHIQM} and L_p-norms, is superior compared to the other four metrics. It is also observed that the prediction consistency for the validation of Δ_{NHIQM} is better when using the exponential mapping compared to the double exponential mapping.

7. Conclusions

In this paper, the design of RR objective perceptual image quality metrics for wireless imaging has been presented. Instead of focusing only on artifacts due to source encoding, the design follows an end-to-end quality approach that accounts for the complex nature of artifacts that may be induced by a wireless communication system. As such, the proposed image quality metrics constitute alternatives to tra-

ditional link layer metrics and may readily be utilized for in-service quality monitoring and resource management purposes. Specifically, both Δ_{NHIQM} and the perceptual relevance weighted L_p-norm are designed with respect to low computational complexity and low overhead, to measure quality degradations in a wireless communication system, and to account for different structural artifacts that have been observed in our distortion model of a wireless link. Here, structural artifacts are detected by related feature metrics.

The general framework for the design of RR objective perceptual image quality metrics is outlined. It comprises of feature extraction, feature normalization, calculation of difference features, relevance weight acquisition, and feature pooling. In addition, curve fitting techniques are used to find the parameters of suitable mapping functions that can translate objective quality metrics into predicted MOS. The transition from subjective to objective perceptual quality is executed in the process of relevance weight acquisition and the derivation of the mapping functions. In both these parts of the design framework, the results of subjective experiments are engaged to train our feature-based quality metrics. Moreover, a detailed description and statistical analysis of the subjective data gathered in these experiments and related objective feature data is provided.

The evaluation of the quality prediction performance reveals that Δ_{NHIQM} and the perceptual relevance weighted L_p-norm both correlate similarly well to human perception on images. This holds not only for the training of the metrics but also for the generalization to unknown images. Furthermore, the numerical results show that both feature-based RR metrics outperform even the considered state of the art FR metrics in prediction performance. As the reduced-reference overhead associated with the calculation of Δ_{NHIQM} is condensed to only a single number, this approach may be the more favorable choice for use in wireless imaging applications compared to the perceptual relevance weighted L_p-norm, which requires all involved features to be communicated from the transmitter to the receiver.

Acknowledgements

The authors wish to thank staff and students of the Western Australian Telecommunications Research Institute, Perth, Australia and the School of Engineering at the Blekinge Institute of Technology, Ronneby, Sweden for participating in the subjective experiments. We would also like to thank the anonymous reviewers for their highly valuable comments which helped to significantly improve the quality of this article.

References

[1] A. C. Brooks and T. N. Pappas, "Structural similarty quality metrics in a coding context: Exploring the space of realistic distortions," IS&T/SPIE Human Vision and Electronic Imaging XI, vol. 6057, pp. 60570U-1–12, Jan. 2006.

[2] M. Carnec, P. Le Callet, and D. Barba, "Objective Quality Assessment of Color Images Based on a Generic Perceptual Reduced Reference," ELSEVIER Image Communication, vol. 23, no. 4, pp. 239–256, Apr. 2008.

[3] D. M. Chandler, K. H. Lim, and S. S. Hemami, "Effects of spatial correlations and global precedence on the visual fidelity of distorted images," IS&T/SPIE Human Vision and Electronic Imaging XI, vol. 6057, pp. 131–145, Jan. 2006.

[4] D. M. Chandler and S. S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," IEEE Trans. on Image Processing, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.

[5] K. Chono, Y.-C. Lin, D. Varodayan, Y. Miyamoto, and B. Girod, "Reduced-Reference Image Quality Assessment Using Distributed Source Coding," IEEE Int. Conf. on Multimedia and Expo, pp. 609-612, Jun. 2008.

[6] S. Daly, "Visible differences predictor: An algorithm for the assessment of image fidelity," IS&T/SPIE Human Vision, Visual Processing, and Digital Display III, vol. 1666, pp. 2–15, Aug. 1992.

[7] U. Engelke and H.-J. Zepernick, "Quality evaluation in wireless imaging using feature-based objective metrics," IEEE Int. Symp. on Wireless Pervasive Comp., San Juan, Puerto Rico, Feb. 2007, pp. 367–372.

[8] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," IEEE Trans. on Commun., vol. 43, no. 12, pp. 2959–2965, Dec. 1995.

[9] M. C. Q. Farias, M. S. Moore, J. M. Foley, and S. K. Mitra, "Perceptual Contributions of Blocky, Blurry,and Fuzzy Impairments to Overall Annoyance," Proc. IS&T/SPIE Human Vision and Electronic Imaging IX, vol. 5292, 2004, pp. 109–120.

[10] M. C. Q. Farias, J. M. Foley, and S. K. Mitra, "Perceptual analysis of video impairments that combine blocky, blurry, noisy, and ringing sytnthetic artifacts," IS&T/SPIE Human Vision and Electronic Imaging X, vol. 5666, 2005, pp. 107–118.

[11] R. Ferzli, L. J. Karam, and J. Caviedes, "A robust image sharpness metric based on kurtosis measurement of wavelet coefficients," Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics, Jan. 2005.

[12] R. Ferzli and L. J. Karam, "Human visual system based no-reference objective image sharpness metric," IEEE

Int. Conf. on Image Processing, Oct. 2006, pp. 2949–2952.

[13] B. Girod, "What's wrong with mean-squared error," in Digital Images and Human Vision, A. B. Watson, ed., pp. 207-220, MIT Press, 1993.

[14] ITU-R Recommendation BT.814, "Specifications and alignment procedures for setting of brightness and contrast of displays," Geneva, Switzerland, 1994.

[15] ITU-R Recommendation BT.815, "Specification of a signal for measurement of the contrast ratio of displays," Geneva, Switzerland, 1994.

[16] ITU-R Recommendation BT.1129-2, "Subjective assessment of standard definition digital television (SDTV) systems," Geneva, Switzerland, 1998.

[17] ITU-R Recommendation BT.500-10, "Methodology for the subjective assessment of the quality of television pictures," Geneva, Switzerland, 2002.

[18] M. Kusuma and H.-J. Zepernick, "Objective hybrid image quality metric for in-service quality assessment," in *Signal Processing for Telecommunications and Multimedia*, Springer, 2005, pp. 44–55.

[19] T. M. Kusuma, "A perceptual-based objective quality metric for wireless imaging," Ph.D. thesis, Curtin University of Technology, Perth, Australia, 2005.

[20] J. Lubin, "A visual discrimination model for imaging system design and evaluation," in Vision Models for Target Detection and Recognition, World Scientific, E. Peli (Ed.), pp. 245–283, 1995.

[21] F. X. J. Lukas and Z. L. Budrikis, "Picture quality prediction based on a visual model," IEEE Trans. on Commun., vol. 30, no. 6, pp. 1679–1692, Jul. 1982.

[22] J. B. Martens, "Multidimensional modeling of image quality," Proc. of the IEEE, vol. 90, no. 1, pp. 133–153, Jan. 2002.

[23] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," IEEE Int. Conf. on Image Processing, pp. 57–60, Sep. 2002.

[24] L. Meesters and J. B. Martens, "A single ended blockiness measure for JPEG coded images," Signal Processing, vol. 82, pp. 369–387, Mar. 2002.

[25] A. F. Molisch, "Wireless communications," Wiley-IEEE Press, 2005.

[26] J.-R. Ohm, "Multimedia communication technology: Representation, transmission and identification of multimedia signals," Springer, 2004.

[27] F. Pereira, "Sensations, perceptions and emotions towards quality of experience evaluation for consumer electronics video adaptations," Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics, Jan. 2005.

[28] H. de Ridder, "Minkowski-metrics as a combination rule for digital-image-coding impairments," IS&T/SPIE Human Vision, Visual Processing, and Digital Display III, vol. 1666, Jan. 1992, pp. 16–26.

[29] H. de Ridder, "Percentage scaling: A new method for evaluating multiple impaired images," Proc. of SPIE, vol. 3959, 2000, pp. 68–77.

[30] A. W. Rix, A. Bourret, and M. P. Hollier, "Models of human perception," J. of BT Tech., vol. 17, no. 1, pp. 24-34, Jan. 1999.

[31] S. Saha and R. Vemuri, "An analysis on the effect of image features on lossy coding performance," IEEE Signal Processing Letters, vol. 7, no. 5, pp. 104–107, May 2000.

[32] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," IEEE Trans. on Image Processing, vol. 15, no. 2, pp. 430-444, Feb. 2006.

[33] D. Soldani, M. Li, and R. Cuny (Ed.), "QoS and QoE management in UMTS cellular systems," Wiley, 2006.

[34] C. J. van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatio-temporal model of the human visual system," SPIE Digital Video Compression: Algorithms and Technologies, vol. 2668, pp. 450–460, Jan. 1996.

[35] Video Quality Experts Group, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment phase II," Aug. 2003, [Online]. Available: http://www.vqeg.org.

[36] T. Vlachos, "Detection of blocking artifacts in compressed video," IEE Electronics Letters, vol. 36, no. 13, pp. 1106–1108, June 2000.

[37] B. A. Wandell, "Foundations of vision," Sinauer Associates, 1995.

[38] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, vol. 4, pp. 3313–3316, May 2000.

[39] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," IEEE Int. Conf. on Image Processing, pp. 477–480, Sep. 2002.

[40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. on Image Processing, vol. 13, no. 4, pp. 600–612, April 2004.

[41] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," IS&T/SPIE Human Vision and Electronic Imaging X, vol. 5666, Mar. 2005, pp. 149–159.

[42] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E. H. Yang, and A. C. Bovik, "Quality aware images," IEEE Trans. on Image Processing, vol. 15, no. 6, pp. 1680–1689, June 2006.

[43] A. B. Watson, J. Hu, and J. F. McGowan, "Digital video quality metric based on human vision," SPIE Journal of Electronic Imaging, vol. 10, no. 1, pp. 20–29, Jan. 2001.

[44] A. R. Weeks, "Fundamentals of electronic image processing," SPIE/IEEE Series on Imaging Science and Engineering, 1998.

[45] S. Winkler, "Digital video quality," Wiley, 2005.

[46] H. R. Wu and K. R. Rao (Ed.), "Digital video image quality and perceptual coding," CRC Press, 2006.

[47] T. Yamada, Y. Miyamoto, M. Serizawa, and H. Harasaki, "Reduced-Reference Based Video Quality Metrics Using Representative Luminance Values," Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics, Jan. 2007.