Quality Evaluation in Wireless Imaging Using Feature-Based Objective Metrics

Ulrich Engelke, Hans-Jürgen Zepernick

ISWPC

2007 San Juan, Puerto Rico

Quality Evaluation in Wireless Imaging Using Feature-Based Objective Metrics

Ulrich Engelke and Hans-Jürgen Zepernick
Blekinge Institute of Technology
PO Box 520, SE–372 25 Ronneby, Sweden
E-mail: {ulrich.engelke, hans-jurgen.zepernick}@bth.se

***Abstract*— This paper addresses the evaluation of image quality in the context of wireless systems using feature-based objective metrics. The considered metrics comprise of a weighted combination of feature values that are used to quantify the extend by which the related artifacts are present in a processed image. In view of imaging applications in mobile radio and wireless communication systems, reduced-reference objective quality metrics are investigated for quantifying user-perceived quality. The examined feature-based objective metrics provide options for controlling the processing complexity and implementation resources required for performing the quality evaluation. Furthermore, the metrics support the scaling of overhead needed for communicating the image features over the wireless link and the accuracy for predicting the image quality.**

I. Introduction

With the deployment of third generation mobile radio networks and the widespread use of wireless networks on business and home premises, delivery of multimedia services to subscribers has become a reality. As a consequence, real-time evaluation of Quality of Service (QoS) and Quality of Experience (QoE) [1] has gained a significant amount of attention. In order to assess the user's satisfaction with a particular service, suitable quality measures need to be identified that are able to quantify service quality as experienced by the end-user. The related quality metrics should preferably not cause impairments to the service but only extract certain features that are relevant for quality assessment. In addition, reduced-reference metrics are favorable in a communication context as these need only to transfer a small set of service features from the transmitter to the receiver. As such, these metrics do not rely on having the complete multimedia signal present at the receiver as reference. This property is particularly useful in wireless applications where bandwidth is usually scarce and expensive.

In this paper, we focus on the evaluation of image quality in wireless systems using objective metrics that comprise of a weighted combination of feature values related to artifacts. These artifacts include blocking, blur, ringing, image activity, and intensity masking [2]. As radio channels would typically have rather limited transmission bandwidth, reduced-reference objective quality metrics are investigated for quantifying user-perceived quality. The actual choice of image features depends on the potential artifacts associated with a particular wireless imaging application, the desired accuracy for predicting the quality, and the available resources for implementing the related feature extraction algorithms. Accordingly, the considered objective metrics provide options for scaling processing complexity by way of selecting the desirable combination of features that suit a particular application. Similarly, the overhead needed for communicating the relevant image features over the wireless link can be adopted to the aforementioned constraints.

This paper is organized as follows. In Section II, feature-based objective image quality metrics are presented and their characteristics are discussed. Section III contains numerical quality results for a wireless imaging scenario along with a discussion of major findings. The conclusions are drawn in Section IV.

II. Feature-based objective image quality metrics

The objective metrics considered in this paper have the benefit of representing structural information of an image by only a small number of bits and hence are favorably for applications in wireless networks where bandwidth resources are usually scarce. It is noted that the attribute of metrics being objective relates to the fact that these estimate or predict subjective scores as opposed to subjective test approaches which require actual test subjects rating the quality. The latter experimental approach does not suit real-time communications but would only be used in the training or derivation of prediction functions for objective metrics.

A. Hybrid Image Quality Metric

The hybrid image quality metric (HIQM) has been proposed in [3], [4]. It belongs to the class of reduced-reference metrics as the original image is not needed to assess the quality of the impaired image. Basically, HIQM measures five image artifacts or features and then represents image quality by a weighted sum of the feature values of the considered image. As such, the quality of an image is concisely characterized by a single value calculated as

$$HIQM = \sum_{i=1}^{I} w_i f_i^* \tag{1}$$

where f_i^* and w_i, $i = 1, \cdots, I$ denote the feature values and their perceptual weights, respectively.

In view of the wireless imaging scenarios examined in this work, we consider five features f_i^*, $i = 1, \cdots, 5$ relating to

blocking f_1^*, blur f_2^*, edge-based image activity f_3^*, gradient-based image activity f_4^*, and intensity masking f_5^*. The related relevance weights $w_i \in [0,1]$, $i = 1, \cdots, 5$ were deduced from subjective quality tests using concepts of statistics such as the Pearson correlation coefficient [5] as has been comprehensively reported in [4]. The similarity between a transmitted and the related received image can then be measured by the absolute difference

$$\Delta_{HIQM} = |HIQM_t - HIQM_r| \quad (2)$$

where $HIQM_t$ and $HIQM_r$ refer to the HIQM values calculated for the transmitted and received image, respectively.

B. Extreme Value Normalized Hybrid Image Quality Metric

Although HIQM uses feature value normalization, namely relevance normalization, an extreme value normalization would be more convenient in view of the intended comparison with other distance metrics. In this case, we normalize the feature values f_i^* as follows [2]:

$$f_i = \frac{f_i^* - \min\limits_{j=1,\cdots,J}(f_{i,j}^*)}{c_i}, \qquad i = 1, \cdots, I \quad (3)$$

where the feature values $f_{i,j}^*$, $j = 1, \cdots, J$ are taken from a training set $\mathcal{J}$ of size J, which in our case were extracted from the images used in the aforementioned subjective quality tests, and

$$c_i = \max_{j=1,\cdots,J}(f_{i,j}^*) - \min_{j=1,\cdots,J}(f_{i,j}^*) \quad (4)$$

In addition to the normalization of (3), let us also account for the special case where these normalized features are simply accumulated to produce the NHIQM value

$$NHIQM = \sum_{i=1}^{I} f_i \quad (5)$$

and then the related absolute difference

$$\Delta_{NHIQM} = |NHIQM_t - NHIQM_r| \quad (6)$$

between the NHIQM values $NHIQM_t$ and $NHIQM_r$ of the transmitted and the received image, respectively.

Alternatively, (1) can be adopted to produce a weighted version of the NHIQM value as

$$NHIQM_W = \sum_{i=1}^{I} w_i f_i \quad (7)$$

by replacing the features f_i^* by the extreme value normalized features f_i given in (3). Accordingly, the weighted version of (6) can be written as

$$\Delta_{NHIQM_W} = |NHIQM_{W,t} - NHIQM_{W,r}| \quad (8)$$

It should be mentioned that the relationship

$$0 \leq f_{i,j} \leq 1 \quad (9)$$

applies to the normalized feature values $f_{i,j}$ associated to the feature values $f_{i,j}^* \in \mathcal{J}$. It can 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 $[0,1]$. For example, severe signal fading due to multipath propagation could result in significant image impairments at certain times such that the user-perceived quality is in a region where the human visual system is saturated, that is where it cannot differentiate anymore among quality degradation levels. Given this clipping is applied such that $f_i \in [0,1]$, $i = 1, \cdots, I$, then the following relationship holds for the NHIQM$_W$ values:

$$0 \leq NHIQM_W \leq \sum_{i=1}^{I} w_i \quad (10)$$

C. L_P-norm

The L_P-norm, also referred to as Minkowski metric, is a distance measure that can be applied to quantify the similarity between two signals or vectors. In image processing it has been applied, for example, with the percentage scaling method [6] and the combining of impairments in digital image coding [7]. In general, the L_P-norm is defined as

$$L_P = \left[\sum_{i=1}^{I} |f_{t,i} - f_{r,i}|^P\right]^{\frac{1}{P}} \quad (11)$$

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

For the two special cases of $P=1$ and $P=2$, respectively, we obtain the Manhattan distance and the Euclidian distance. For $P \to \infty$, we have

$$L_\infty = \max_{i=1,\cdots,I} |f_{t,i} - f_{r,i}| \quad (12)$$

meaning that the largest absolute feature value difference solely dominates the norm. This property may be explored in wireless imaging as fading may cause significant degradation of one particular image feature. In this case, only the associated feature extraction algorithm would need to be run on the mobile terminal resulting in conservation of battery power.

As we have user-perceived relevance of the considered five artifacts on the quality available from our data base of subjective tests, the following weighted version of the L_P-norm can be applied:

$$L_{P,W} = \left[\sum_{i=1}^{I} w_i^P |f_{t,i} - f_{r,i}|^P\right]^{\frac{1}{P}} \quad (13)$$

Presuming normalized features and clipping of values outside the range of the training set similar as with NHIQM, the following bounds apply to the values of the $L_{P,W}$-norm:

$$0 \leq L_{P,W} \leq \left[\sum_{i=1}^{I} w_i^P\right]^{\frac{1}{P}} \quad (14)$$

TABLE I

PREDICTION ACCURACY OF IMAGE QUALITY FOR DIFFERENT FEATURE COMBINATIONS AND IMAGE QUALITY METRICS

Number of features	Features used					NHIQM		Order P of $L_{P,W}$- and L_P-norm			
						$NHIQM_W$	$NHIQM$	$P=1$		$P=2$	
	f_1	f_2	f_3	f_4	f_5	r_W	r	$r_{P,W}$	r_P	$r_{P,W}$	r_P
1	x					0.870	0.870	0.870	0.870	0.870	0.870
		x				0.471	0.471	0.471	0.471	0.471	0.471
			x			0.653	0.653	0.653	0.653	0.653	0.653
				x		0.649	0.649	0.649	0.649	0.649	0.649
					x	0.499	0.499	0.499	0.499	0.499	0.499
2	x	x				0.826	0.702	0.862	0.829	0.879	0.854
	x		x			0.884	0.880	0.877	0.873	0.882	0.883
	x			x		0.884	0.860	0.883	0.858	0.879	0.864
	x				x	0.896	0.892	0.896	0.894	0.886	0.888
		x	x			0.632	0.249	0.655	0.636	0.661	0.636
		x		x		0.452	0.390	0.529	0.584	0.481	0.524
		x			x	0.474	0.474	0.507	0.507	0.495	0.495
			x	x		0.635	0.608	0.643	0.658	0.643	0.630
			x		x	0.779	0.789	0.780	0.786	0.766	0.777
				x	x	0.414	0.689	0.794	0.811	0.805	0.801
3	x	x	x			0.795	0.815	0.858	0.826	0.881	0.855
	x	x		x		0.788	0.796	0.856	0.814	0.878	0.844
	x	x			x	0.882	0.861	0.879	0.843	0.892	0.865
	x		x	x		0.883	0.853	0.877	0.849	0.885	0.864
	x		x		x	0.911	**0.914**	0.904	**0.906**	0.894	**0.899**
	x			x	x	0.907	0.885	**0.906**	0.904	0.895	0.864
		x	x	x		0.660	0.501	0.655	0.658	0.658	0.642
		x	x		x	0.712	0.449	0.722	0.694	0.716	0.678
		x		x	x	0.464	0.422	0.606	0.672	0.498	0.571
			x	x	x	0.777	0.809	0.774	0.804	0.758	0.772
4	x	x	x	x		0.852	0.808	0.852	0.812	0.880	0.844
	x	x	x		x	0.843	0.899	0.880	0.851	0.892	0.868
	x	x		x	x	0.842	0.873	0.872	0.830	0.891	0.853
	x		x	x	x	**0.912**	0.897	0.904	0.887	**0.897**	0.882
		x	x	x	x	0.735	0.584	0.723	0.720	0.713	0.686
5	x	x	x	x	x	0.894	0.899	0.874	0.838	0.891	0.856

Correlation: r_W - relevance weighted NHIQM ($\forall i: 0 \leq w_i \leq 1$), r - non-weighted NHIQM ($\forall i: w_i = 0$ or 1)
$r_{P,W}$ - relevance weighted L_P-norm ($\forall i: 0 \leq w_i \leq 1$), r_P - non-weighted L_P-norm ($\forall i: w_i = 0$ or 1)

III. NUMERICAL RESULTS

The results presented in this section are based on subjective tests that were conducted at the Department of Signal Processing of the Blekinge Institute of Technology adopting the test methodology as outlined in ITU-R Recommendation BT.500-11 [8]. In particular, a Rayleigh fading channel in the presence of Additive White Gaussian Noise (AWGN) along with error control coding was used as a simple simulation model of a wireless link. This scenario was used to generate a wide range of impaired images for the subjective test including images with extreme artifacts. A total of forty images in Joint Photographic Experts Group (JPEG) format and represented in gray scale were used. The subjective test, involving 30 non-expert viewers to provide their quality ratings, resulted in a data base of Mean Opinion Scores (MOS) for each image. This enabled us to deduce the feature weights needed in the calculation of the considered image quality metrics following the approach reported in [9]. In view of applying extreme value normalized metrics, the weights for incorporating normalized feature values were obtained from statistical analysis as

$$w_1 = 0.77,\ w_2 = 0.35,\ w_3 = 0.61,\ w_4 = 0.16,\ w_5 = 0.35.$$

Blocking turns out to be the most annoying artifact in the considered wireless imaging test scenario as it produces the highest relevance weight of $w_1 = 0.77$. On the other hand, impairments related to the gradient-based image activity result in the lowest response as seen from the weight $w_4 = 0.16$. It is noted that the above feature weights w_i for the related extreme

TABLE II

FIRST AND SECOND ORDER STATISTICS FOR THE CONSIDERED DIFFERENCE FEATURES

	Mean μ and variance σ^2 of difference features Δf				
	Δf_1	Δf_2	Δf_3	Δf_4	Δf_5
μ	0.314	0.126	0.115	0.049	0.061
σ^2	0.041	0.023	0.010	0.021	0.036

	Cross-correlation between difference features Δf				
	Δf_1	Δf_2	Δf_3	Δf_4	Δf_5
Δf_1	1.000	0.377	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

Δf_1: Blocking, Δf_2: Blur, Δf_3: Edge, Δf_4: Gradient, Δf_5: Masking

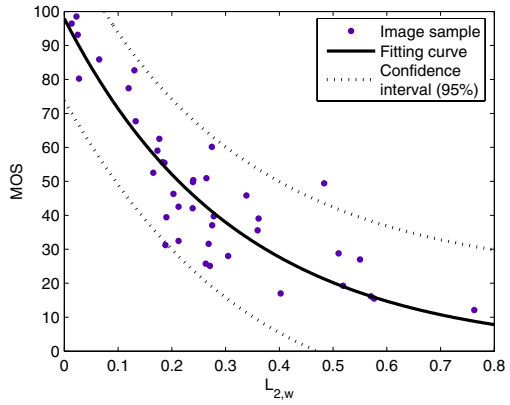

Fig. 1. Curve fitting of subjective scores, MOS, versus relevance weighted L_P-norm, $L_{2,W}$ when using all five features and their related weights. The obtained fitting curve is used as prediction function to obtain predicted MOS for a given value of the L_P-norm. Here, curve fitting results in the prediction function $MOS_{L_{2,W}} = c\,e^{-d\,L_{P,2}}$ with $c = 97.94$ and $d = -3.157$.

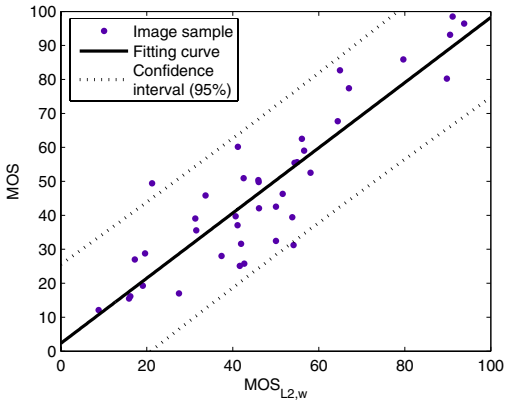

Fig. 2. Curve fitting of subjective scores, MOS, versus predicted scores, $\text{MOS}_{L_{2,W}}$, using the prediction function for the relevance weighted L_P-norm $L_{2,W}$. The correlation of the fitting represents the prediction accuracy and is obtained here as $r_{2,W} = 0.891$. In other words, it quantifies the degree to which the metric captures the image quality as experienced by a user.

value normalized features f_i in (3) are equal to the weights for the non-normalized features f_i^* in [4]. This is because the normalization does not effect the statistical relationship between the image impairments and the subjective scores, that is the correlation between features and MOS values.

Given these relevance weights, an exponential prediction function has been suggested in [4] to translate difference HIQM values to predicted MOS values. In the sequel, we therefore comply with this approach and use the following relationships to calculate predicted MOS values for difference NHIQM values as well as for the L_P-norm:

$$MOS_{NHIQM_W} = a\,e^{-b\,\Delta_{NHIQM_W}} \quad (15)$$

$$MOS_{L_{P,W}} = c\,e^{-d\,L_{P,W}} \quad (16)$$

where the parameters a, b, c, and d are obtained from the curve fitting of the respective prediction function to the results obtained from the subjective tests (see example fittings in Figs. 1 and 2 for the L_2-norm). Note that for each metric, the related curve fitting has to be performed in order to derive the corresponding prediction function.

Apart from this particular weighting, other choices of combining the impact of different artifacts on the user-perceived quality may be considered. For example, one or the other feature may be removed from the evaluation. The related feature extraction algorithm would not have to be run on the mobile terminal which in turn would reduce the implementation complexity. In order to provide some guidelines for combining features, a number of selections have been considered and the results in terms of their prediction accuracy of user-perceived quality are shown in Table I. The table shows the respective Pearson correlation coefficients for NHIQM and for the L_P-norm with and without relevance weighting. The higher the value in the table, the higher is the accuracy by which the feature combination predicts the user-perceived image quality. The maximum achieved correlation is shown in bold for each of the examined metrics. Some more generic characteristics of the considered metrics for quality evaluation in wireless imaging may be deduced from Table I as follows:

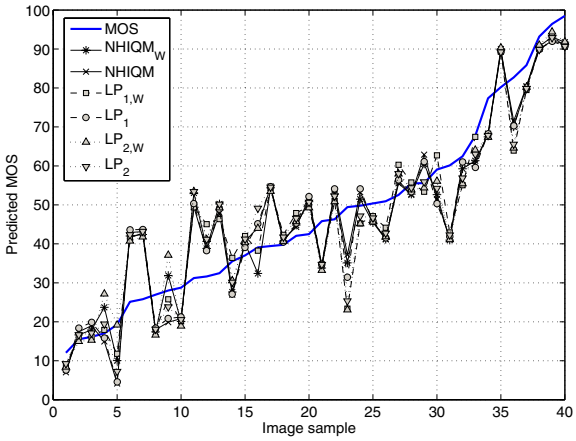

Fig. 3. Predicted MOS for image samples that are ranked with respect to increasing MOS values (unmarked solid line).

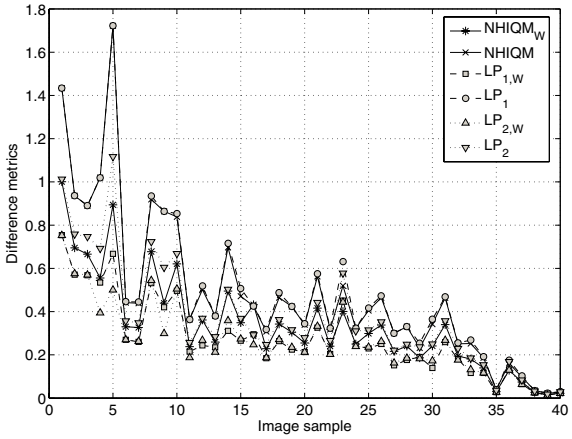

Fig. 4. Values of the difference metrics related to image samples that are ranked with respect to increasing MOS values (see Fig. 3).

- The spread of quality prediction accuracy for different combinations of features appears to decrease the more features are used. This indicates an increase in robustness with respect to changes in the application scenario when several features are accounted for in the quality evaluation. It is also noted that this characteristic is more developed for the L_P-norm.
- For many feature combinations, an increase in quality prediction accuracy can be achieved with the L_p-norm by simply increasing the parameter P from $P = 1$ to $P = 2$. Further improvements in accuracy by increasing the parameter P to $3, 4, 5$ have not been observed.
- It has been observed that high quality prediction accuracy is always connected to blocking being included into the combination of features.
- By allowing the relevance weights to take on values other than those derived from curve fitting the difference features to the MOS results from the subjective tests, a better prediction accuracy has been achieved for some feature combinations.

Table II provides first and second order statistics for the

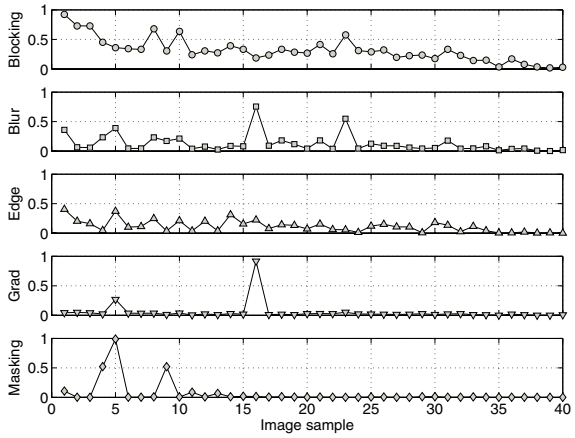

Fig. 5. Values of the difference features related to image samples that are ranked with respect to increasing MOS values (see Fig. 3).

difference features

$$\Delta f_i = |f_{t,i} - f_{r,i}|, \qquad i = 1, \cdots, 5 \tag{17}$$

in terms of their mean μ and variance σ^2 as well as the cross-correlation between these difference features. It should be mentioned that these statistics are based on the set of images used in the subjective tests. It can be seen from the table that blocking appears in the mean as the dominant feature, which can be somewhat expected given the JPEG format of the images. The variance is relatively small for blocking, blur, and edge-based image activity but larger for gradient-based image activity and masking with reference to their mean. This is as either very small or large values of the latter two features were measured for the considered set of images (see also Fig. 5).

The cross-correlation between the difference features indicate strong similarity of blocking and edge-based image activity as well as blur and gradient-based image activity. It is also interesting that blocking and edge-based image activity are paired as two artifacts that may cause rather abrupt changes in the image structure. On the other hand, blur and gradient-based image activity relate to smoother impairments to the image.

In the sequel, we will consider more detailed those particular feature combinations that have produced the maximum prediction accuracy as represented by the largest correlation value in each column of the related objective metric in Table I. These correlation values shall be summarized for convenience:

$$\begin{aligned} \text{NHIQM}_W\text{:} \quad & r_W = 0.912 \\ \text{NHIQM:} \quad & r = 0.914 \\ L_{1,W}\text{:} \quad & r_{1,W} = 0.906 \\ L_1\text{:} \quad & r_1 = 0.906 \\ L_{2,W}\text{:} \quad & r_{2,W} = 0.897 \\ L_2\text{:} \quad & r_2 = 0.899 \end{aligned}$$

Figure 3 shows the progression of the predicted MOS related to the forty image samples that are ranked with respect to the MOS results from the subjective tests for the

aforementioned special weighted and non-weighted feature combinations. It can be seen from the figure that the predicted MOS follows the increasing trend given by the actual MOS.

The difference metrics that produce the predicted MOS values are depicted in Fig. 4. It can be seen that for the image sample 5 the largest spread of difference metrics is generated with the L_1-norm and NHIQM indicating the largest impairments. In order to reveal how the individual features contribute to the difference metrics for the considered images, Fig. 5 provides the progression of the values of the difference features. It can be seen that for image sample 5, significant masking is detected while blocking appears as additional impairment.

Figures 6 and 7 show the two image samples with index 5 and 7, respectively, with the first producing extreme difference features (see Fig. 5). The image sample 'Barbara' in Fig. 6 appears severely impaired by masking with about a third of the viewing area in the bottom part of the image turned dark while the upper area is shifted towards brighter intensity. Furthermore, the four remaining features reach difference values of almost 0.5 and hence pose additional stress on the image quality. It is also noted that the predicted MOS value for 'Barbara' is lower then the actual MOS (see Fig. 3) indicating a conservative prediction value. This may potentially be attributed to saturation effects in the human visual system.

As far as the image sample 'Mandrill' in Fig. 7 is concerned, the dominant impairments that were detected are blocking and edge-based image activity. This impaired image seems to be somewhat overrated by all considered difference metrics giving an overall predicted MOS of around 40. This behavior is thought to be attributed to the artifacts being most visible in the central region of the image while the outer region appears undistorted due to masking effects. This may be used in future work for quality assessment with respect to region of interest.

Fig. 6. Image sample number 5: 'Barbara'.

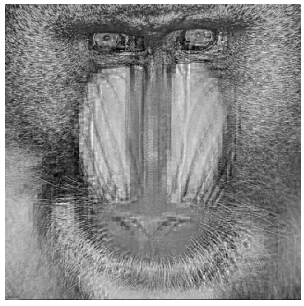

Fig. 7. Image sample number 7: 'Mandrill'.

IV. Conclusions

In this paper, we have focused on the evaluation of image quality for wireless systems using feature-based objective metrics. In particular, the extreme value normalized hybrid image quality metric and the L_P-norm have been considered. Both types of metrics support reduced-reference concepts in which only structural information of the images have to be communicated over the wireless link. The considered feature-based objective image quality metrics have the potential to be deployed in wireless imaging systems due to favorable characteristics such as the following:

- The NHIQM-based quality evaluation requires the lowest overhead for carrying the reference information about the transmitted image to the receiver as only a single number has to be communicated. As such, this overhead does not depend on the number of features considered. Also, NHIQM provides the best prediction accuracy.
- Quality evaluation using the L_P-norm provides the mobile with information about each considered feature, which may be used to adapt the system such that particular impairments are controlled and preferably suppressed.
- The overhead needed for the L_P-norm reduces with each feature that is removed from the metric calculation.
- The considered image quality metrics may be used to support explicit and implicit link adaptation in wireless multimedia systems. This includes techniques such as power control, adaptive modulation and coding as well as automatic repeat request and soft-combining schemes.
- Adaption of the examined metrics to wireless video applications that use inter-frame coded formats such as Motion JPEG2000 can easily be performed.

References

[1] D. Soldani, M. Li, and R. Cuny, (Co-Editors), "QoS and QoE Management in UMTS Cellular Systems," Chichester: John Wiley & Sons, 2006.

[2] J.-R. Ohm, "Multimedia Communication Technology: Representation, Transmission and Identification of Multimedia Signals," Berlin: Springer, 2004.

[3] T. M. Kusuma and H.-J. Zepernick, "A Reduced-reference Perceptual Quality Metric for In-service Image Quality Assessment," IEEE Symp. on Trends in Commun., Bratislava, Slovakia, Oct. 2003, pp. 71-74.

[4] T. M. Kusuma, "A Perceptual-based Objective Quality Metric for Wireless Imaging," Ph.D. thesis, Curtin University of Technology, Perth, Australia, 2005.

[5] S. Winkler, "Digital Video Quality - Vision Models and Metrics," Chichester: John Wiley & Sons, 2005.

[6] H. de Ridder and M. C. Willemsen, "Percentage Scaling: A New Method for Evaluting Multiply Impaired Images," Proc. of SPIE Human Vision and Electonic Imaging V, vol. 3959, pp. 68-77, 2000.

[7] H. de Ridder, "Minkowski-metrics as a Combination Rule for Digital-image-coding Impairments," Proc. of SPIE Human Vision, Visual Processing, and Digital Display III, vol. 1666, pp. 16-26, 1992.

[8] ITU-R Recommendation BT.500-11, "Methodology for the Subjective Assessment of the Quality of Television Pictures," 2002.

[9] M. Kusuma, H.-J. Zepernick, and M. Caldera, "On the Development of a Reduced-Reference Perceptual Image Quality Metric," Int. Conf. on Multimedia Commun. Systems, Montreal, Canada, Aug. 2005, pp. 178-184.