VISUAL QUALITY INDICES AND LOW QUALITY IMAGES

Heinz Hofbauer and Andreas Uhl

Department of Computer Sciences
University of Salzburg
{hhofbaue, uhl}@cosy.sbg.ac.at

ABSTRACT

Visual quality indices are frequently used instead of human evaluation for the quality assessment of impaired images (or video material). These visual quality indices are in turn evaluated on databases containing impaired images in conjunction with a score given by evaluation with human observers. The fitness of these indices are judged on the entire quality scale of the respective database. However, this leads to the incorrect assumption that these quality indices perform well over the whole range possible qualities. This is unfortunately not true, especially towards the low quality range of images these quality indices often show little actual correlation to human judgement. In this paper a number of visual quality indices will be evaluated with regard to the lower quality spectrum of impairments and it will be shown that the overall fitness of a quality index is not generally related to its performance regarding high impairment.

Index Terms— Image analysis, Quality control

1. INTRODUCTION

The assessment of image and video quality is important whenever image and videos are transmitted (gauging of transmission errors), encoded (compression vs. quality) etc. Optimally a jury of humans would judge the impact of the impairment, however the high time and cost required to do this are prohibitive. Thus, visual quality indices(VQI) are used to simulate the assessment that should be made by humans. In order to judge the correlation of VQI to the average human judgement a number of databases have been created containing distorted images along with a mean opinion score (MOS) of human observers, for example LIVE [1] or TID [2]. These databases typically contain different testsets which correspond to typical application scenarios, e.g. JPEG or JPEG2000 compression, distortion scenarios, e.g. transmission errors or denoising, and operations on images, e.g. gaussian blur or masking.

The evaluation of a VQI is usually done over the whole range of impairments in a given database. This is reasonable to estimate the overall fitness of VQIs but there are certain shortcomings in this approach. For example in high compression scenarios a VQI which does well overall is less useful than one which performs best for the given low quality scale (and the same is essentially true for a high quality range). The underlying problem is that VQIs overall performance does not correlate to performance for low quality scenarios or even, though less frequently, for high quality scenarios. Typical low quality scenarios are low bitrate videos [3, 4], video streaming [5, 6], assessment of transmission errors [7] or quality control for transparent encryption [8].

While there is some previous work regarding low quality video sequences [9], there is, to the extent of the authors knowledge, no information available regarding low quality image assessment over the range of recent VQIs. To rectify this shortcoming, state of the art VQIs will be evaluated on known databases where the focus is on low and high quality subsets rather than the whole database. This will also show that the databases which are already in existence are sufficient to properly evaluate the performance of VQIs and point out that it would be good practice to evaluating performance in a more discerning way.

In order to facilitate the reproducibility of the research we restricted ourself to publicly available data and implementations.

In section 2 a brief overview of the evaluated VQIs will be given, in section 3 the evaluation based on the LIVE and selected subsets of the TID database will be given.

2. OVERVIEW OF VISUAL QUALITY INDICES

Modern VQIs often use a sophisticated approach on quality which heavily relies on knowledge about the human visual system (HVS). In the following we will give a short review of the VQIs which will be evaluated.

The most widely used VQI today is the peak signal-to-noise ratio (PSNR) since it is easy to implement and fast to compute. It is also well known that the PSNR does not reflect human judgement very well. In [10] Huynh-Thu and Ghanbari showed that as long as the content is unchanged the PSNR reasonably well reflects the human observer.

The luminance and edge similarity score (LSS and ESS) was introduced by Mao and Wu [11]. They used the informa-

tion on the HVS to find criteria how observers judge images. The edge information reflects the assessment of humans regarding the shape or contour of objects and the luminance score reflects changes in the color space. Both algorithm use 8×8 windows to assess the edge direction and mean luminance of a region in the image.

The visual signal-to-noise ratio (VSNR) [12] uses a two stage method of quality assessment. In the first stage contrast detection thresholds are calculated by using wavelet (DWT) based models of visual masking and summation to assess if the errors are perceivable by the HVS. If the errors are judged to be below the detection threshold the image is considered pristine. When the errors are above the threshold of detection a score is calculated by using the ratio of the RMS contrast to the weighted values of the perceived contrast and global precedence.

The structural similarity index measure SSIM [13] extracts three separate scores from the image and combines them into the final score. First the visual influence is calculated locally then luminance, contrast and structural scores are calculated globally. These separate scores are then combined with equal weight to form the SSIM score.

The multi-scale structural similarity index measure MS-SSIM [14] is an extension of the SSIM to take into account that the perceivability of image impairments is different depending on the sampling density of the image signal, e.g. as influenced by viewing distance. To take this into account the similarity scores are calculated at different spatial scales. The core operation is similar to SSIM, contrast and structural scores are calculated at each scale and the luminance score is calculated at the lowest scale. The factors for combining these scores where found by experiments with human observers.

Criterion v4.0 C4 [15] uses a detailed model of the HVS, and information regarding the score is extracted from a transformation of the image in the perceptual space. The transformation include compensation for display device gamma, perceptual colorspace, luminance normalization, contrast sensitivity functions, subband decomposition and modelling of masking effects. From this perceptual model the contrast orientation, length and width as well as the subband amplitude and average luminance, red-green chroma and yellow-blue chroma channels are extracted from characteristic points in the model. The local scores are generated as averaging of the extracted features and the overall score is generated by averaging the local scores.

For the visual information fidelity criterion VIF [16] a more refined model is used which starts with the modeling of the reference image using natural scene statistics (NSS). Furthermore, the possible distortion is modeled as signal gain and additive noise in the wavelet domain and parts of the HVS which have not been covered by the NSS are modeled, i.e. internal neural noise is modeled by using a additive white Gaussian noise model. Using this model the VIF score reflects the fraction of the reference image information which can be extracted from the impaired image.

The Weighted Signal to Noise Ration (WSNR) [17] is defined as the ratio of the average weighted signal power to the average weighted noise power. The weight function used is a contrast sensitivity function (CSF) which is gained by using the frequency response of HVS. A measure of the non-linear HVS response to a single frequency, the contrast threshold function (CTF), is used which is measured over the visible radial spatial frequencies. The CTF is the minimum amplitude necessary to detect a sine wave of a given angular spatial frequency. The CSF is the frequency response obtained by inverting the CTF.

With respect to implementations we used our own code[1] for PSNR, SSIM, LSS, ESS. For C4 the implementation from Carnec et al. was used and for all other VQIs the "MeTriX MuX Visual Quality Assessment Package[2] was used in version 1.1. Also note that UQI, IFC and NQM from Metrix Mux were not included in the evaluation since they are predecessors of other VQIs which were evaluated.

3. EVALUATION OF VISUAL QUALITY INDICES

In the following evaluations the Spearman rank order correlation (SROC) is used to compensate for non linearity. The VQIs were evaluated on two different databases for two reasons. First, we want to show that the shortcomings of the VQIs are not based on the distortion and image types of a single database. Secondly, we want to show that the problems of VQI with lower quality images are not a result of the evaluation method employed in the database assembly. The DMOS value of these two databases was derived differently, the LIVE database uses a linear scale of perceived impairment where observers judge each image individually while the TID database uses a direct comparison of two impaired images where the observer selects the higher quality image and thus creates a ranking in order of perceived impairment.

3.1. Evaluation on the LIVE Database

The first comparison will be on base of the LIVE database[3]. The comparison is between the full range of the database as would be used for regular VQI evaluation, table 1, the low quality part of the database with a DMOS of greater than 80 (70 for gblur in order to keep the number of distortions high enough), table 3, and the high quality range with a DMOS lower than 40, table 2. In these tables value with SROC lower than 0.5 are underlined to show low correlation and the best score per testset is printed in a bold font.

To get a better overview fig. 1 illustrates the relation of the SROC from tables 1 through 6. Figure 2 shows the same

[1] http://www.wavelab.at

[2] http://foulard.ece.cornell.edu/gaubatz/metrix_mux/

[3] http://live.ece.utexas.edu/research/quality/

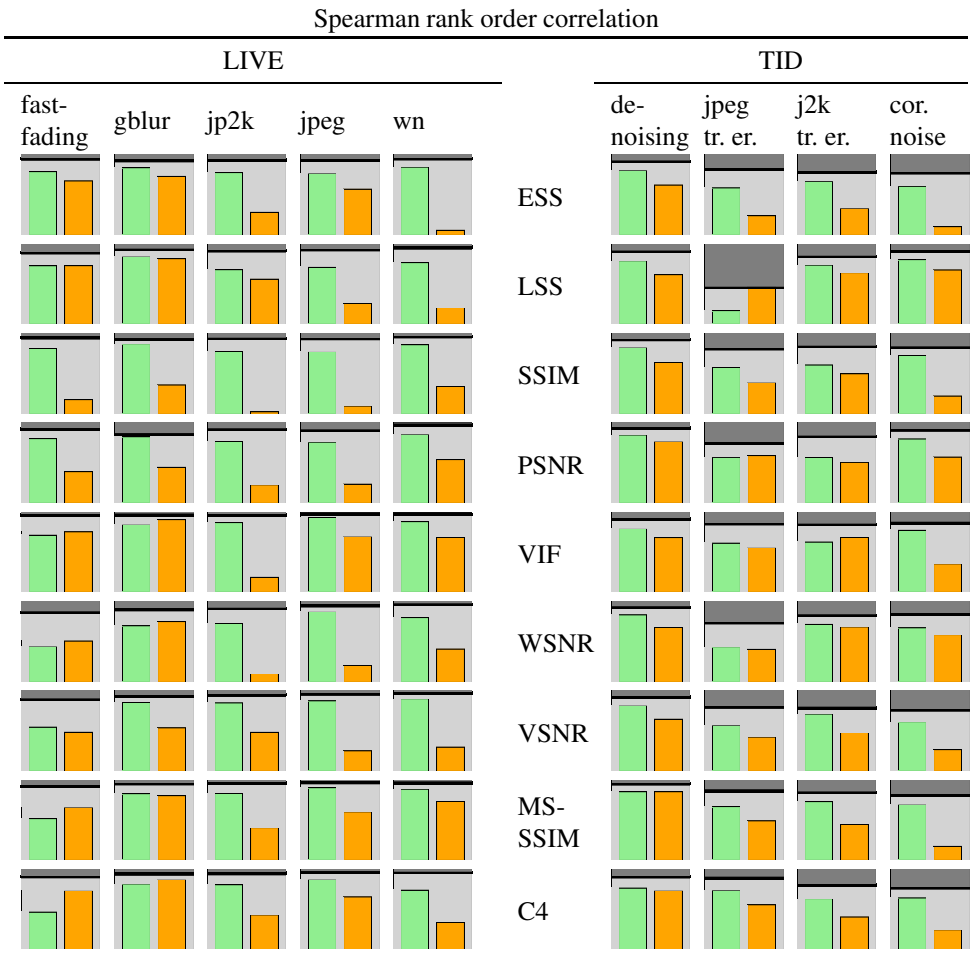

Fig. 1. For each entry the SROC for the full range of qualities is indicated by the line separating the light and dark background ranging from SROC $= 0.0$ at the bottom to SROC $= 1.0$ at the top. Superimposed are two bar charts showing the SROC for the high quality range on the left side and the low quality range on the right side.

Table 1. LIVE Image Quality Assessment Database

	fastfading	gblur	jp2k	jpeg	wn
ESS	0.956	0.939	0.944	0.945	0.958
LSS	0.898	0.941	0.928	0.939	0.967
SSIM	0.957	0.935	0.940	0.940	0.968
PSNR	0.927	0.865	0.923	0.913	0.982
VIF	**0.965**	**0.972**	**0.968**	**0.984**	**0.985**
WSNR	0.873	0.909	0.920	0.958	0.973
VSNR	0.903	0.941	0.955	0.966	0.978
MS-SSIM	0.932	0.958	0.965	0.979	0.973
C4	0.919	0.956	0.959	0.975	0.970

Table 2. LIVE Image Quality Assessment Database, high quality (DMOS ≤ 40)

	fastfading	gblur	jp2k	jpeg	wn
ESS	0.799	0.848	0.790	0.775	0.854
LSS	0.738	0.853	0.688	0.720	0.779
SSIM	**0.821**	**0.879**	0.789	0.780	0.868
PSNR	0.809	0.838	0.780	0.764	0.865
VIF	0.722	0.853	**0.880**	**0.942**	0.889
WSNR	0.444	0.707	0.732	0.881	0.810
VSNR	0.553	0.861	0.856	0.887	**0.908**
MS-SSIM	0.529	0.839	0.841	0.918	0.894
C4	0.473	0.826	0.818	0.883	0.753

information for the Kendall τ_b rank order correlation, for reasons of brevity we did not give tables of τ_b values since overall the behavior is the same as for SROC which is nicely illustrated by the figures given.

It can be directly read from the comparison of low and high quality ranges that the VQIs, with a few exceptions, perform worse for the lower quality range than the higher quality range. Furthermore, the reduced performance for the lower quality range can not be reduced to a lower number of

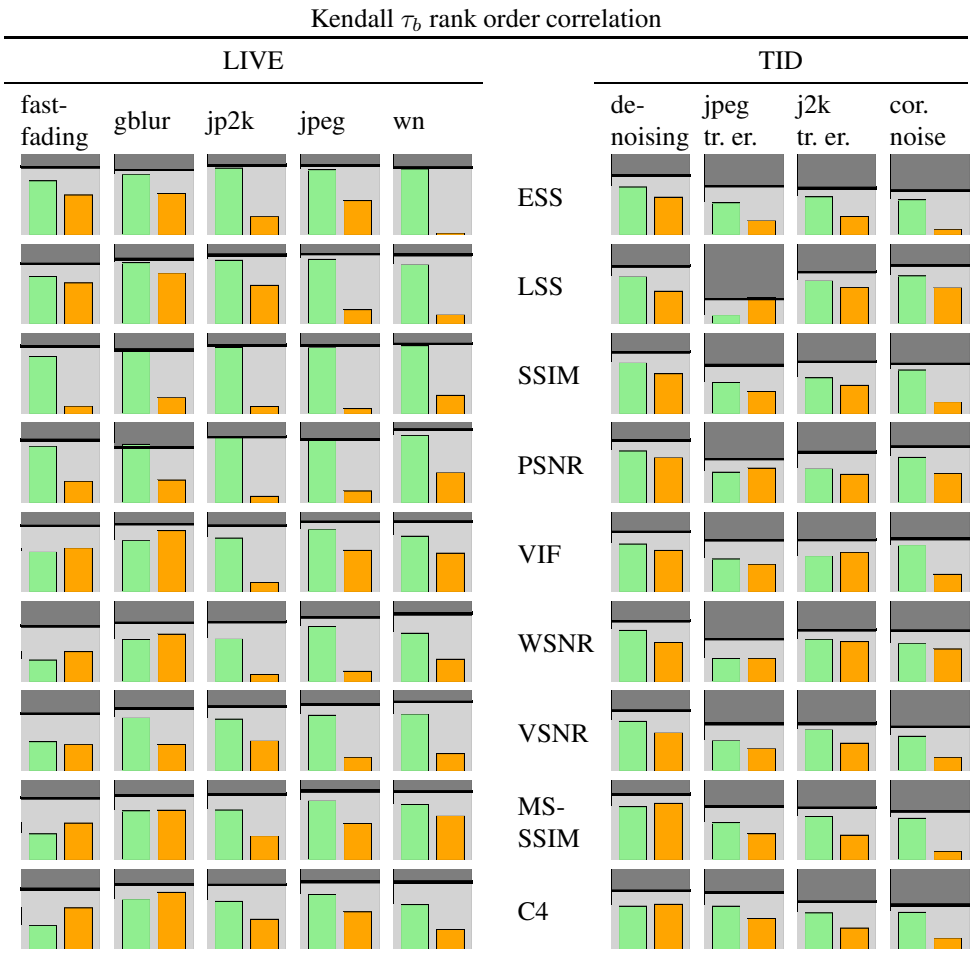

Fig. 2. For each entry the τ_b for the full range of qualities is indicated by the line separating the light and dark background ranging from $\tau_b = 0.0$ at the bottom to $\tau_b = 1.0$ at the top. Superimposed are two bar charts showing the τ_b for the high quality range on the left side and the low quality range on the right side.

Table 3. LIVE Image Quality Assessment Database, low quality (DMOS $>$ 80, *70)

	fastfading	gblur*	jp2k	jpeg	wn
ESS	0.686	0.74	0.291	0.583	0.062
LSS	0.738	0.828	**0.573**	0.268	0.210
SSIM	0.179	0.365	0.027	0.095	0.348
PSNR	0.398	0.451	0.227	0.243	0.549
VIF	**0.767**	**0.919**	0.191	**0.703**	0.692
WSNR	0.514	0.76	0.100	0.209	0.414
VSNR	0.492	0.547	0.491	0.257	0.303
MS-SSIM	0.665	0.819	0.409	0.607	**0.741**
C4	0.741	0.882	0.436	0.666	0.342

comparison images since the higher quality range used is the same distance from the mean of the DMOS values and thus, roughly, the same number of comparison images are used. A notable VQI is the VIF which displays good performance for all cases except high compression rates for JPEG2000 compression where it is among the worst. All VQIs however show certain deficiencies regarding low quality images, even though the actual deficiency is dependant on the distortion introduced. Furthermore, even if the distortion is known beforehand, the evaluation over the full database can be misleading. As an example the best VQI to evaluate highly compressed JPEG2000 images would be the LSS, but the overall performance of LSS regarding JPEG2000 compression is among the worst.

Furthermore, while the reduction in performance is usually in the lower end of the quality spectrum this is not always so. Compare for example the performance of C4 and WSNR for the high and low quality range of the fastfading and gblur testsets. For both VQIs and both testsets the performance on highly impaired images is better than for high quality version.

Table 4. TID2008 Tampere Image Database 2008 (v1.0)

	denoising	jpeg tr. er.	j2k tr. er.	cor. noise
ESS	0.922	0.827	0.789	0.777
LSS	0.912	0.460	0.850	**0.917**
SSIM	0.930	0.818	0.840	0.833
PSNR	0.942	0.752	0.831	0.916
VIF	0.919	0.858	0.851	0.870
WSNR	0.934	0.738	0.834	0.848
VSNR	0.929	0.806	0.791	0.766
MS-SSIM	**0.957**	0.874	**0.853**	0.819
C4	0.918	**0.901**	0.808	0.777

Table 5. TID2008 Tampere Image Database 2008 (v1.0), high quality (DMOS > 3.5)

	denoising	jpeg tr. er.	j2k tr. er.	cor. noise
ESS	0.815	0.596	0.676	0.613
LSS	0.798	0.179	**0.744**	**0.819**
SSIM	0.834	0.582	0.615	0.738
PSNR	0.856	0.577	0.578	0.805
VIF	0.801	0.623	0.634	0.781
WSNR	0.846	0.438	0.720	0.677
VSNR	0.823	0.575	0.715	0.619
MS-SSIM	**0.868**	0.680	0.739	0.703
C4	0.774	**0.745**	0.640	0.655

3.2. Evaluation on the TID Database

The second comparison will be based on the TID Tampere Image Database 2008[4]. Due to reasons of space we only present a subset of the 17 testsets contained in the database, these are "Image denoising", "JPEG transmission errors", "JPEG2000 transmission errors" and "Spatially correlated noise" abbreviated as "denoising", "jpeg tr. er.", "j2k tr. er." and "cor. noise" respectively in the tables. The comparison is again between the full quality range, Table 4, the low quality part DMOS of lower than 3.5, table 6, and the high quality range greater than 3.5, table 5. Again, in these tables values with SROC lower than 0.5 are underlined to show low correlation and the best score per testset is printed in a bold font.

Overall the comparison again shows that the VQIs perform worse for a low quality subset than a high quality subset, even if the performance over the full range is high. However, there are some VQIs which perform better for lower qualities. The LSS for example performs better on the low quality version of the JPEG transmission error testset than on the full quality range.

Furthermore, like with the LIVE database, a high performance of an VQI over the whole quality range can not be taken as indicator that the VQI will perform optimally in either a low or high quality range.

[4] http://www.ponomarenko.info/tid2008.htm

Table 6. TID2008 Tampere Image Database 2008 (v1.0), low quality (DMOS < 3.5)

	denoising	jpeg tr. er.	j2k tr. er.	cor. noise
ESS	0.634	0.248	0.335	0.109
LSS	0.628	0.471	0.649	**0.685**
SSIM	0.650	0.393	0.507	0.224
PSNR	0.773	**0.600**	0.513	0.584
VIF	0.690	0.562	**0.691**	0.357
WSNR	0.682	0.411	0.686	0.590
VSNR	0.652	0.425	0.480	0.269
MS-SSIM	**0.868**	0.497	0.451	0.180
C4	0.743	0.567	0.415	0.247

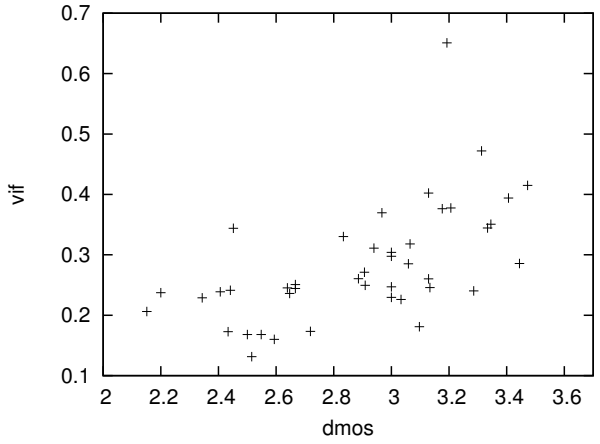

Fig. 3. Scatter plot of VIF over DMOS for the JPEG 2000 transmission errors testset from the TID 2008 database for low quality images (DMOS < 3.5)

To examine in more detail which effects lead to this problem fig. 3 gives a scatter plot of VIF ratings over DMOS for the low quality range of the J2K transmission errors testset, containing 44 of the 100 image in the testset. It can clearly be seen that the overall tendency of higher VIF for higher DMOS ratings holds. It is also clear that locally a high variance in VIF ratings can be observed leading to large mismatches. To illustrate the problem observe that the lowest quality according to VIF would be at about DMOS 2.5, resulting in 8 images being rated higher quality than the corresponding observers would, this is more than 18% of the image in the low quality testset. The same holds regarding the highest quality image according to the VIF.

4. CONCLUSION

It was shown that even seemingly well performing VQIs actually have flaws which can be seen under close scrutiny. When performance is measured only using the full quality range

provided by image evaluation databases these flaws tend to be concealed since the overall correlation between DMOS and VQI overrides the large variance when taking into account a subset of qualities. Two statements can be made regarding the availability of VQIs and the corresponding testing of VQIs. One, there is a lack of VQIs which target the low quality images and performs well over a wide range of distortion types. There are VQIs which are well suited for evaluation of such images when the distortion type is known in advance and a proper VQI can be chosen, however, this becomes less clear when a mix of two or more distortion types can be expected. Second, the evaluation process of VQIs should be done in a more elaborate way, specifically it should differentiate between overall fitness and fitness on low and high quality ranges to better identify shortcomings of certain VQIs. While at least a split into low and high qualities should be done, it might be expedient to differentiate between low, medium and high quality ranges where the database allows, i.e. when enough levels of distortion in the database exist to keep the significance high.

5. REFERENCES

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

[2] N. Ponomarenko, F. Battisti, K. Egizarian, J. Astola, and V. Lukin, "Metrics performance comparison for color image database," in *Fourth international workshop on video processing and quality metrics for consumer electronics*, Arizona, USA, Jan. 2009, p. 6 p.

[3] Mark A. Masry and Sheila S. Hemami, "CVQE: A metric for continuous video quality evaluation at low bit rates," in *Human Vision and Electronic Imaging*, Bernice E. Rogowitz and Thrasyvoulos N. Pappas, Eds., Santa Clara CA, Jan. 2003, vol. 5007 of *SPIE Proceedings*, pp. 116–127.

[4] E. P. Ong, W. Lin, Zhongkang Lu, S. Yao, and M. H. Loke, "Perceptual quality metric for h.264 low bit rate videos," in *IEEE International Conference on Multimedia and Expo*, Toronto, Ont., July 2006, pp. 677–680.

[5] L. Superiori, O. Nemethove, W. Karner, and M. Rupp, "Cross-layer detection of visual impairments in h.264/avc video sequences streamed over umts networks," in *Proceedings of IEEE 1st International Workshop on Cross Layer Design*, Jinan, Shandong, China, Sept. 2007, pp. 96–99.

[6] M. Ries, O. Nemethova, and M. Rupp, "Video quality estimation for mobile h.264/AVC video streaming," *Journal of Communications*, vol. 3, no. 1, pp. 41–50, 2008.

[7] Michael Gschwandtner, Andreas Uhl, and Peter Wild, "Transmission error and compression robustness of 2D Chaotic Map image encryption schemes," *EURASIP Journal on Information Security*, vol. 2007, no. Article ID 48179, pp. doi:10.1155/2007/48179, 16 pages, 2007.

[8] Thomas Stütz and Andreas Uhl, "On efficient transparent JPEG2000 encryption," in *Proceedings of ACM Multimedia and Security Workshop, MM-SEC '07*, New York, NY, USA, Sept. 2007, pp. 97–108, ACM Press.

[9] E. P. Ong, M. H. Loke, W. Lin, Zhongkang Lu, and S. Yao, "Perceptual quality metric for h.264 low bit rate videos," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, Honolulu, Hl, 2007, vol. 1, pp. 889–892.

[10] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, June 2008.

[11] Y. Mao and M. Wu, "Security evaluation for communication-friendly encryption of multimedia," in *Proceedings of the IEEE International Conference on Image Processing (ICIP'04)*, Singapore, Oct. 2004, IEEE Signal Processing Society.

[12] Damon Chandler and Sheile Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284–2298, Sept. 2007.

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

[14] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th IEEE Asilomar Conference on Signals, Systems and Computers*, Nov. 2003, pp. 1398–1402.

[15] Mathieu Carnec, Patrick Le Callet, and Dominique Barba, "Objective quality assessment of color images based on a generic perceptual reduced reference," *Signal Processing: Image Communication*, vol. 23, no. 4, pp. 239–256, Apr. 2008.

[16] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, May 2006.

[17] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 636–650, Apr. 2000.