Sensitivity of Tone Mapped Image Quality Metrics to Perceptually Hardly Noticeable Differences

Nikola Banić and Sven Lončarić
Image Processing Group
Faculty of Electrical Engineering and Computing
University of Zagreb, 10000 Zagreb, Croatia
E-mail: {nikola.banic, sven.loncaric}@fer.hr

***Abstract*—As high dynamic range images are being used more widely, the need for good tone mapping operators (TMOs) i.e. methods for their conversion to low dynamic range images rises as well. In evaluation of results of TMOs objective image quality metrics are often used for practical reasons. Since these metrics only approximate perceptual evaluation, they are sometimes too sensitive to perceptually unimportant details. In this paper such sensitivities of three recent tone mapped image quality metrics are compared: TMQI, TMQI-II, and FSITM. These metrics have been chosen because they are the most appropriate objective quality metrics for the problem of tone mapping. The comparison is performed by using specifically designed tone mapped images to check the measures' susceptibility to perceptually unnoticeable changes in brightness of the resulting image. It is shown that while values of TMQI and FSITM are only slightly affected by such changes, the recent TMQI-II can obtain significantly different values, which brings into question its ability perform a fair TMO comparison. The results are presented and discussed.**

***Index Terms*—High dynamic range, objective image quality assessment, FSITM, low dynamic range, TMQI, TMQI-II, tone mapping.**

I. Introduction

Images with high dynamic range (HDR) i.e. with a high ratio between the largest and smallest intensity are being more widely used with the advance of imaging technology [1]. Since most display devices still support only low dynamic range (LDR) images, there is a need for tone mapping operators (TMOs) i.e. for dynamic range compression methods that convert HDR images to their LDR versions. Tone mapping is a challenging problem and therefore many TMOs have been proposed so far. TMOs are global [2]–[10] if they handle same intensities in the same way across the whole image. On the other hand, if they handle intensities based on the content of their close neighborhood, then they are local [11]–[18]. The main characteristic of global TMOs is their speed and simplicity, while local TMOs are usually more complex and they produce better LDR images of higher quality [19]–[21].

An important part in development of TMOs is the quality evaluation of their results and an accurate way to do that is to perform subjective quality assessment. However, due to a large number of existing TMOs, subjectively comparing a new TMO even with only state-of-the-art TMOs on a larger testing dataset becomes in most cases too slow and impractical. For this reason the objective image quality metrics have been introduced and currently they are often used to simplify and speed up the quality assessment of LDR images produced by a TMO. Since these metrics only approximate subjective evaluation, it can happen that they assess two very similar images very differently i.e. they are sometimes too sensitive to differences that the human visual system does not even notice. If in such cases only the values of these metrics are taken into account, then these differences can erroneously lead to wrong conclusions about the actual performance of different TMOs.

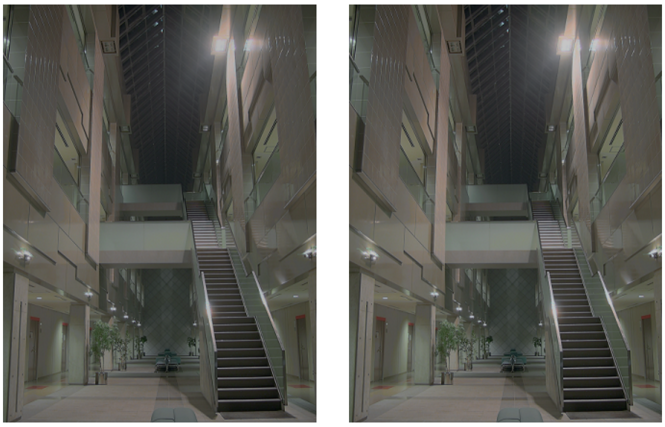
(a) (b)

Fig. 1: Tone mapped images of the same scene. The values of TMQI, TMQI-II, and FSITM^G_TMQI quality measures are for (a) 0.8455, 0.5723, and 0.8352, respectively, and for (b) 0.8635, 0.8044, and 0.8394, respectively.

For this reason it is important to check to what degree are some of the widely used metrics sensitive to such differences. In this paper three recent tone mapped image quality metrics are tested for such sensitivity: TMQI [22], TMQI-II [23], and FSITM [24]. They are chosen because they are the most appropriate objective quality metrics designed specifically for quality assessment of tone mapped images. The testing is performed by using specifically designed tone mapped images to check the measures' sensitivity to alterations of mean brightness of the resulting LDR images. It is demonstrated that for images with slight, but perceptually unnoticeable mean brightness differences the results of TMQI-II can be significantly different, while the values of TMQI and FSITM are affected on a much smaller scale, even in the worst case. This brings into question TMQI-II's practical usability and credibility in fair evaluation and comparison of quality of LDR resulting images produced by using different TMOs.

The paper is structured as follows: Section II describes three recent objective tone mapped image quality metrics, Section III gives the motivation for the comparison of their sensitivity to perceptually unnoticeable differences, in Section IV the comparison is performed and its results are presented and discussed, and Section V concludes the paper.

II. Objective tone mapped image quality metrics

Subjectively assessing the quality of LDR images obtained after carrying out tone mapping of the initial HDR images often results in good and relatively accurate comparison of performance of various TMOs. However, a large drawback of such subjective assessment is that it is slow and it usually takes a lot of time. It also makes TMO development based on measuring improvement over some other TMOs impractical due to the lack of automation. For this reason various objective quality metrics for tone mapped images have been introduced.

One of the widely used objective quality metrics that was also one of the first ones designed specifically for the purpose of objective quality assessment of tone mapped images is the Tone Mapped image Quality Index (TMQI) [22]. It evaluates the structural fidelity and statistical naturalness of a tone mapped image by comparing it to the original HDR image. The final result is a real number in range $[0, 1]$ with a higher value meaning higher quality and vice versa. In [23] TMQI has been upgraded to TMQI-II, which is supposed to be its improved version and additionally there is an iterative procedure for improving an initially tone mapped image in terms of its TMQI-II value. Another recent metric is the Feature Similarity Index For Tone-Mapped Images (FSITM), which is based on local phase information of images and like TMQI-II it was also shown to outperform TMQI. If FSITM is combined with TMQI, it gives better results and for this combination the notation FSITM^C_TMQI [24] is used where C is a color channel. In the rest of the paper the green (G) channel is used because the authors have shown that its usage gives good results. The combination FSITM^G_TMQI was shown [25] to outperform both TMQI and TMQI-II as well. It should be mentioned that these three metrics are currently state-of-the-art in the area of objective quality assessment of tone mapped images. The main advantage of objective quality assessment over subjective quality assessment is that it can be used to automate the evaluation of the performance of a TMO.

However, since objective metrics only approximate perceptual subjective evaluation, there are possible cases where the comparison results obtained by objective and subjective quality assessment significantly differ. A particularly interesting case is when an objective metric is too sensitive to perceptually unnoticeable differences that should be disregarded. This is clearly a drawback because such and similar cases can erroneously lead to unfair comparison of even very similar TMOs.

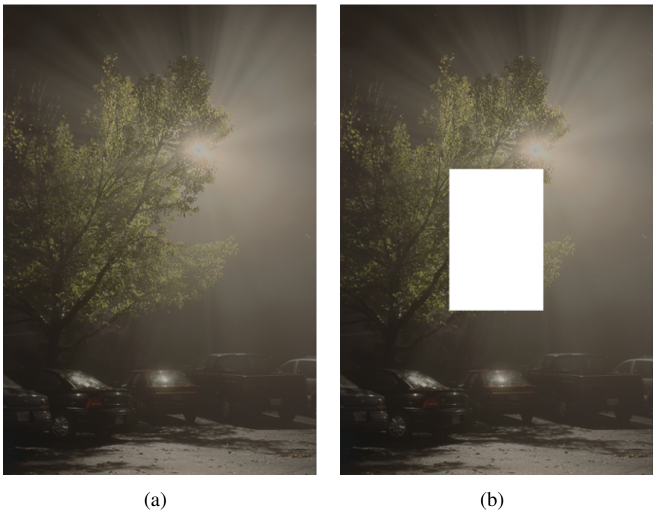
(a) (b)

Fig. 2: Crudely increasing mean image brightness by placing a white rectangle in it with everything else remaining the same. The values of TMQI, TMQI-II, and FSITM^G_TMQI indices are for (a) 0.7993, 0.3861, and 0.8448, respectively, and for (b) 0.7754, 0.8195, and 0.7980, respectively.

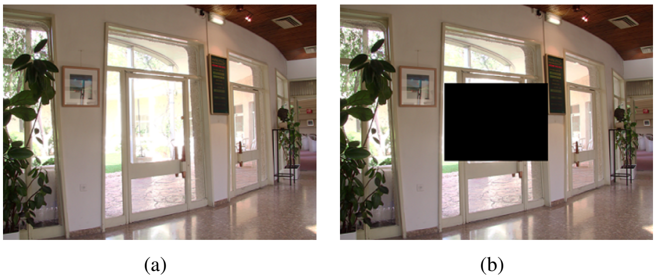
(a) (b)

Fig. 3: Crudely decreasing mean image brightness by placing a black rectangle in it with everything else remaining the same. The values of TMQI, TMQI-II, and FSITM^G_TMQI indices are for (a) 0.9043, 0.4701, and 0.8646, respectively, and for (b) 0.9019, 0.8410, and 0.8395, respectively.

III. Motivation for comparison

The direct motivation for a comparison between sensitivities of different objective tone mapped image quality metrics were the observed significant fluctuations of TMQI-II values for the same images before and after slightly changing their mean brightnesses by multiplying them by a constant. An example is shown in Fig. 1 where the difference of image grayscale means is less than 5. By performing some additional similar experiments with manipulation of image grayscale mean, it becomes evident that TMQI-II is so susceptible to image brightness that it sometimes puts it before the content. If e.g. the mean image brightness is adjusted by introducing artificial content as in Fig. 2 and Fig. 3, the values of TMQI and FSITM^G_TMQI decrease as intuitively expected, but the values of TMQI-II increase significantly despite a clear loss of information. The shown examples are not some rare, specially designed cases and similar results can be obtained for practically any other tone mapped images as well.

Since the shown examples with large content manipulations are highly unlikely to be encountered during development of new TMOs, the mentioned metrics' sensibility should be tested in more realistic conditions. Nevertheless, these examples can point in the direction of more suitable sensitivity tests.

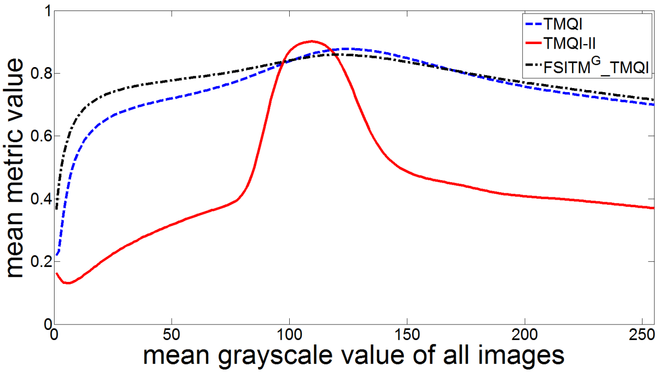

Fig. 4: The impact of forcing all tone mapped images to a given mean grayscale value on the mean metric values.

IV. Experimental results

A. Experimental setup

For further experiments the HDR images available at [26] were used. They were used mainly because they originate from different sources and some of them were even artificially generated, which means that altogether they cover a larger variety of HDR image types. The initial step before carrying out any other experiments was to tone map all of these images by applying Reinhard's TMO [13] implemented in the open-source Luminance HDR software with the same default parameters being used for all images. Reinhard's TMO was chosen mainly because it is a widely known and used TMO and it was shown to give high quality results. Results very similar to the ones described later in the paper could also be obtained by using some other TMO as well. If $\mathbf{I}^{(i)}$ is the LDR resulting image obtained by applying Reinhard's TMO to i-th of the initial HDR images, then for each $\mathbf{I}^{(i)}$ the next step was to create two additional images $\mathbf{I}_A^{(i)}$ and $\mathbf{I}_B^{(i)}$. Their respective j-th pixels were calculated as $\mathbf{I}_A^{(i)}(j) = k_A^{(i)}\mathbf{I}^{(i)}(j)$ and $\mathbf{I}_B^{(i)}(j) = k_B^{(i)}\mathbf{I}^{(i)}(j)$ with $k_A^{(i)} < k_B^{(i)}$. The values of constants $k_A^{(i)}$ and $k_B^{(i)}$ were deliberately chosen so that the difference between a chosen objective quality metric for $\mathbf{I}_A^{(i)}$ and $\mathbf{I}_B^{(i)}$ was maximized under two constraint. The first constraint was that the mean CIELab E^*_{ab} approximated perceptual difference between corresponding pixels of $\mathbf{I}_A^{(i)}$ and $\mathbf{I}_B^{(i)}$ must stay below the just-noticeable difference (JND) threshold of 2.3 [27]. The second constraint was that the values of $k_A^{(i)}$ and $k_B^{(i)}$ must be from set $\{50, 51, ..., 200\}$ to exclude the possibility of unnaturally looking images that could be obtained for too high or too low values of constants $k_A^{(i)}$ or $k_B^{(i)}$.

When this was done for all images $\mathbf{I}^{(i)}$, the result were two new sets A and B with corresponding images $\mathbf{I}_A^{(i)}$ and $\mathbf{I}_B^{(i)}$, which were designed to have slight differences that are perceptually hardly noticeable or not noticeable at all like the ones in Fig. 1. This effectively means that the values of a good objective quality metric for an image from set A and for its corresponding image in set B should differ only slightly. To check whether that holds for TMQI, TMQI-II, and FSITM^G_TMQI, the two mentioned sets were created for each of these metrics and the quality of the obtained images in them was evaluated by calculating these same metrics for them.

B. Numerical results

Table I shows mean values of all objective quality metrics for sets A and B created to maximize the difference for specified metrics. It can be seen that in the individual metrics' worst case scenario of sensitivity to perceptually hardly noticeable differences only TMQI-II is significantly affected. To describe this better, Mann-Whitney U test [28] of the null hypothesis that the distribution of metric values for images in set A is identical to distribution of metric values for images in set B was performed for each metric's worst case. The p-values obtained during the tests for TMQI, TMQI-II, and FSITM^G_TMQI were 0.0811, $3.0260 \cdot 10^{-12}$, and 0.0343, respectively, which clearly shows that TMQI-II is too sensitive.

Another experiment was performed to illustrate the problem more clearly. The initial HDR images were tone mapped by using Reinhard's TMO as was done earlier, but then each image was multiplied by a constant in order to set its mean pixel grayscale value to 1 and then the mean value of all metrics on these images was calculated. This was then repeated by setting the mean pixel grayscale value to every integer in interval $[1, 255]$. The obtained results were as shown in Fig. 4.

C. Discussion

Although in some cases TMQI-II can fail drastically as demonstrated by Table I and Figures 1, 2, and 3, this happens only when the mean grayscale value of a tone mapped image is near the steep parts of the TMQI-II curve shown in Fig. 4. A possible abuse of this situation would be to include this knowledge into a TMO only in order to get a better TMQI-II score and thus seemingly outperform other TMOs. Another less malign case that does not involve including this knowledge into a TMO is when a new TMO accidentally happens to give more images with mean grayscale values favorably valued by TMQI-II than other TMOs do. Since TMQI and FSITM^G_TMQI do not suffer so seriously from this problem, they are probably a significantly better metric choice for evaluating different TMOs in order to determine which of them is supposed to produce results of higher quality. However, it should be mentioned that the results obtained by the iteratively improving an initially given LDR image to gradually improve its TMQI-II metric value [23] still gives high quality results.

V. Conclusions

The sensitivities of several tone mapped image quality metrics to hardly noticeable and unnoticeable differences

TABLE I: Mean values of all objective quality metrics for sets A and B created to maximize the difference for specified metrics.

	Created to maximize TMQI difference			**Created to maximize TMQI-II difference**			**Created to maximize FSITM^G_TMQI difference**		
Created dataset	**TMQI**	**TMQI-II**	**FSITM^G_TMQI**	**TMQI**	**TMQI-II**	**FSITM^G_TMQI**	**TMQI**	**TMQI-II**	**FSITM^G_TMQI**
A	0.7920	0.4847	0.7956	0.8040	0.5481	0.8218	0.7805	0.4492	0.7856
B	0.8147	0.5110	0.8130	0.8178	0.7506	0.8290	0.8012	0.4824	0.8049

in images were compared. For two of them, TMQI and FSITM^G_TMQI, it was shown that this sensitivity is not so high. On the other hand, however, in the case of TMQI-II it was shown to be very high, which can represent a significant problem in practical applications of this metric. A conclusion that can be drawn from the presented experimental results is that for comparison of results of different TMOs it is better to use TMQI and FSITM^G_TMQI instead of TMQI-II.

References

[1] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski, *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010.

[2] J. Tumblin and H. Rushmeier, "Tone reproduction for realistic images," *Computer Graphics and Applications, IEEE*, vol. 13, no. 6, pp. 42–48, 1993.

[3] K. Chiu, M. Herf, P. Shirley, S. Swamy, C. Wang, K. Zimmerman *et al.*, "Spatially nonuniform scaling functions for high contrast images," in *Graphics Interface*. CANADIAN INFORMATION PROCESSING SOCIETY, 1993, pp. 245–245.

[4] G. Ward, "A contrast-based scalefactor for luminance display," *Graphics gems IV*, pp. 415–421, 1994.

[5] C. Schlick, "Quantization techniques for visualization of high dynamic range pictures," in *Photorealistic Rendering Techniques*. Springer, 1995, pp. 7–20.

[6] S. N. Pattanaik, J. Tumblin, H. Yee, and D. P. Greenberg, "Time-dependent visual adaptation for fast realistic image display," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 2000, pp. 47–54.

[7] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive logarithmic mapping for displaying high contrast scenes," in *Computer Graphics Forum*, vol. 22, no. 3. Wiley Online Library, 2003, pp. 419–426.

[8] E. Reinhard and K. Devlin, "Dynamic range reduction inspired by photoreceptor physiology," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 11, no. 1, pp. 13–24, 2005.

[9] G. W. Larson, H. Rushmeier, and C. Piatko, "A visibility matching tone reproduction operator for high dynamic range scenes," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 3, no. 4, pp. 291–306, 1997.

[10] G. J. Braun and M. D. Fairchild, "Image lightness rescaling using sigmoidal contrast enhancement functions," *Journal of Electronic Imaging*, vol. 8, no. 4, pp. 380–393, 1999.

[11] J. Tumblin and G. Turk, "LCIS: A boundary hierarchy for detail-preserving contrast reduction," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 1999, pp. 83–90.

[12] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," *ACM transactions on graphics (TOG)*, vol. 21, no. 3, pp. 257–266, 2002.

[13] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," in *ACM Transactions on Graphics (TOG)*, vol. 21, no. 3. ACM, 2002, pp. 267–276.

[14] R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," in *ACM Transactions on Graphics (TOG)*, vol. 21, no. 3. ACM, 2002, pp. 249–256.

[15] R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "A perceptual framework for contrast processing of high dynamic range images," *ACM Transactions on Applied Perception (TAP)*, vol. 3, no. 3, pp. 286–308, 2006.

[16] L. Meylan and S. Susstrunk, "High dynamic range image rendering with a retinex-based adaptive filter," *Image Processing, IEEE Transactions on*, vol. 15, no. 9, pp. 2820–2830, 2006.

[17] N. Banić and S. Lončarić, "Color Badger: A Novel Retinex-Based Local Tone Mapping Operator," in *Image and Signal Processing*. Springer, 2014, pp. 400–408.

[18] ——, "Puma: A High-Quality Retinex-Based Tone Mapping Operator," in *Signal Processing Conference (EUSIPCO), 2016 24rd European*. IEEE, 2016, pp. 943–947.

[19] J. Kuang, H. Yamaguchi, G. M. Johnson, and M. D. Fairchild, "Testing HDR image rendering algorithms," in *Color and Imaging Conference*, vol. 2004, no. 1. Society for Imaging Science and Technology, 2004, pp. 315–320.

[20] J. Kuang, H. Yamaguchi, C. Liu, G. M. Johnson, and M. D. Fairchild, "Evaluating HDR rendering algorithms," *ACM Transactions on Applied Perception (TAP)*, vol. 4, no. 2, p. 9, 2007.

[21] C. Urbano, L. Magalhães, J. Moura, M. Bessa, A. Marcos, and A. Chalmers, "Tone mapping operators on small screen devices: an evaluation study," in *Computer Graphics Forum*, vol. 29, no. 8. Wiley Online Library, 2010, pp. 2469–2478.

[22] H. Yeganeh and W. Zhou, "Objective Quality Assessment of Tone Mapped Images," *Image Processing, IEEE Transactions on*, vol. 22, no. 2, pp. 657–667, 2013.

[23] K. Ma, H. Yeganeh, K. Zeng, and Z. Wang, "High dynamic range image compression by optimizing tone mapped image quality index," *Image Processing, IEEE Transactions on*, vol. 24, no. 10, pp. 3086–3097, 2015.

[24] H. Ziaei Nafchi, A. Shahkolaei, R. Farrahi Moghaddam, and M. Cheriet, "FSITM: A Feature Similarity Index For Tone-Mapped Images," *Signal Processing Letters, IEEE*, vol. 22, no. 8, pp. 1026–1029, 2015.

[25] ——. (2015, Nov.) FSITM: A Feature Similarity Index For Tone-Mapped Images (Supplementary material). [Online]. Available: http://www.synchromedia.ca/system/files/FSITM_Sup.pdf

[26] High Dynamic Range Image Examples, month=sep, year=2015, url=http://www.anyhere.com/gward/hdrenc/pages/originals.html.

[27] M. Mahy, L. Van Eycken, and A. Oosterlinck, "Evaluation of uniform color spaces developed after the adoption of CIELAB and CIELUV," *Color research and application*, vol. 19, no. 2, pp. 105–121, 1994.

[28] R. Kirk, *Statistics: an introduction*. Cengage Learning, 2007.