Evaluating the Multi-Scale iCID metric

Steven Le Moan[a], Jens Preiss[b] and Philipp Urban[c]

[a]Interactive Graphics Systems Group, Technische Universität Darmstadt,
[b]Institute of Printing Science and Technology, Technische Universität Darmstadt,
[c]Fraunhofer Institute for Computer Graphics Research IGD,
Darmstadt, Germany

ABSTRACT

In this study, we investigate the extent to which an image-difference metric based on structural similarity can correlate with human judgment. We introduce a modified version of the recently published iCID metric and present new results over two large image quality databases. It is particularly noteworthy that the proposed metric yields a correlation of 0.861 with mean opinion scores on the 2013 version of the renowned Tampere Image Database, without dedicated parameter optimization.

1. INTRODUCTION

Despite the fact that there is still much to understand about how the Human Visual System (HVS) interprets an image, recent studies on Image Quality Assessment (IQA) have allowed to reach very high levels of correlation with human judgment[1,2] with low-level features only. In this paper, we focus on Full-Reference IQA metrics*, which take as an input two digital images (reference and reproduction) and provide a score corresponding to the degree of distortion perceived by the HVS.

Recently, Lissner *et al.*[3] introduced the Color-Image-Difference (CID) metric, inspired by the well-known SSIM index,[4] and designed especially to improve the prediction of the latter on chromatic distortions such as created by gamut-mapping algorithms. Later, Preiss *et al.*[5] proposed an improved version – referred to as iCID – in order to account for specific artifacts such as chromatic ringing or chromatic edges. Though the purpose of this most recent publication leans towards optimizing gamut mapping rather than IQA, the iCID metric showed great potential on the 2013 version of the renowned Tampere Image Database (TID2013).[2] Compared to the original CID, this improved version uses less parameters, does not require any training for these parameters and takes into account a larger variety of artifacts by considering more image-difference features.

A more recent study by Zhan *et al.*[6] reported a Spearman correlation larger than 0.89 with MOS on TID2013, which are the best results achieved to this date. This metric (referred to as Visual Saliency Index) is tuned by a certain number of parameters, which are trained over a subset of TID2008. Although the metric proposed here does not yield as high correlations as VSI, it does however perform better than the rest of the state-of-the-art on TID2013 without dedicated training on any database. The scope of this study pertains to evaluating the extent to which a metric based on structural similarity[4] can correlate with human judgment. We introduce a multi-scale version of iCID and we demonstrate its efficiency on two renowned databases.

2. MULTI-SCALE ICID

Let **O** (original) and **R** (reproduction) be two tri-chromatic images (defined e.g. in sRGB). In the iCID framework **O** and **R** are first normalized with an image-appearance model with respect to the viewing conditions (e.g. visual resolution, illuminant, luminance). Particularly, normalizing the images with respect to the visual resolution is crucial in order to consider the differences between the chromatic and achromatic contrast sensitivities of the human visual system. This is done by filtering the input images with contrast sensitivity functions adapted from iCAM (see e.g.[7]). The images are then converted to the nearly perceptually uniform LAB2000HL color space,[8] which is optimized for CIED65. Seven so-called Image-Difference Features (IDF) are then computed, by means of

*Note that the term *metric* is here not used according to its proper mathematical definition. It is however quite established in the image quality community.

terms adapted from the SSIM index[4]: Lightness-Difference ($\mathbf{L}_L$), Lightness-Contrast ($\mathbf{C}_L$), Lightness-Structure ($\mathbf{S}_L$), Chroma-Difference ($\mathbf{L}_C$), Chroma-Contrast ($\mathbf{C}_C$), Chroma-Structure ($\mathbf{S}_C$) and Hue-Difference ($\mathbf{L}_H$). They are extracted in the form of IDF maps which depict their spatial organization, and which are then averaged so as to produce a single score for each IDF. We refer to the original paper by Preiss *et al.*[5] for further explanations. Although all maps are originally computed on a single scale, we propose to compute the contrast and structure terms on 5 different scales, as suggested in.[9] Each map is then averaged and the features are combined into the *multi-scale* iCID (MS-iCID) score, as follows:

$$\text{MS-iCID} = 1 - \mathbf{L}_L^5 \mathbf{L}_C^1 \mathbf{L}_H^1 \prod_{i=1}^{n} \left(\mathbf{C}_L^i (\mathbf{S}_L^i)^3 \mathbf{C}_C^i \mathbf{S}_C^i \right)^{\alpha_i} \quad (1)$$

where n is the number of scales used for the contrast and structure terms. In this study, we chose $n = 5$, as in.[3] The scale coefficients α_i are adapted from the *multi-scale* SSIM (MS-SSIM).[9] Each IDF term $\mathbf{L}_L^i$ represents the average value of the corresponding feature map at scale i. Note that the Lightness-Structure term ($\mathbf{S}_L$) is raised to the power of three, in order to account for the importance of structure in human difference perception.[5] Furthermore, the Lightness-Difference term ($\mathbf{L}_L$) is computed on the smallest scale (5), as suggested in.[9]

In addition, MS-iCID uses three parameters to balance the contribution of the seven IDFs. Note that unlike state-of-the-art metrics whose parameters are trained over a particular image quality database (e.g.[10]), these ones were selected by means of a visual inspection by three expert observers, so as to minimize the artifacts when the metric is used as an objective function to optimize gamut mapping (see[5] Section III.F).

3. EXPERIMENTS

3.1 Benchmark

We compared the proposed MS-iCID metric with seven state-of-the-art metrics: the original single-scale iCID,[5] the Visual Saliency Index (VSI),[6] the feature similarity index (FSIMc),[10] the PSNR-HA,[11] the structural similarity index (SSIM)[4] and its multiscale version (MS-SSIM),[9] as well as the visual information fidelity index (VIF).[12] We compared them over the following databases, which are widely used in the image quality community: TID2013[2] (25 scenes, 3000 reproductions), CSIQ[13] (30 scenes, 866 reproductions), LIVE[1] (29 scenes, 779 reproductions). Note that we assumed a visual resolution of 20 cycles/degree for each database. We used the Spearman Rank Order Correlation Coefficient (SROCC) with Mean Opinion Scores.

Table 1. Performance comparison (SROCC) of IQA metrics - overall

	MS-iCID	iCID	VSI	FSIMc	PSNR-HA	MS-SSIM	SSIM	VIF	Rank MS-iCID
TID2013	0.861	0.813	**0.896**	0.851	0.779	0.786	0.627	0.677	2
CSIQ	0.927	0.922	**0.942**	0.931	0.915	0.913	0.837	0.919	3
LIVE	0.879	0.891	0.902	0.920	0.876	0.903	0.851	**0.953**	6

3.2 Results

Table 1 shows the overall results obtained on each database, whereas Table 2 gives the results obtained for each kind of distortion in TID2013. Finally, Table 3 gives the results obtained for selected scenes from TID2013.

First of all, we note that the multi-scale version of the iCID performs better than the single-scale one on both TID2013 and CSIQ databases, but not on LIVE. In fact, it performs rather poorly – w.r.t. the benchmark – on this latter database. Given that MS-iCID can be seen as an elaborated extension of MS-SSIM for color, and given that the latter performs better than the former by an SROCC of +0.024 on LIVE, we can assume that MS-iCID tends to overestimate the chromatic distortions on this particular database, which comprises mostly achromatic distortions. In other words, the contributions of achromatic and chromatic IDFs might not be well balanced in MS-iCID for LIVE.

Table 2. Performance comparison (SROCC) for each distortion in TID2013.

	MS-iCID	iCID	VSI	FSIMc	PSNRHA	MS-SSIM	SSIM	VIF	Rank MS-iCID
Additive Gaussian noise	0.914	0.908	**0.946**	0.910	0.929	0.865	0.853	0.900	3
Additive Gaussian noise - color	0.820	0.817	0.871	0.854	**0.897**	0.773	0.774	0.830	5
Spatially correlated noise	0.900	0.898	**0.937**	0.890	0.932	0.854	0.862	0.884	3
Masked noise	0.836	0.824	0.770	0.809	0.804	0.807	0.809	**0.845**	2
High frequency noise	0.904	0.900	0.920	0.904	**0.953**	0.860	0.846	0.897	4
Impulse noise	0.867	0.852	0.874	0.825	**0.898**	0.763	0.799	0.854	3
Quantization noise	0.843	0.870	0.875	0.881	**0.909**	0.871	0.806	0.786	6
Gaussian blur	0.968	**0.975**	0.961	0.955	0.944	0.967	0.963	0.965	2
Image denoising	0.944	0.948	0.948	0.933	**0.954**	0.927	0.910	0.892	4
JPEG compression	0.953	0.948	0.954	0.934	**0.961**	0.927	0.910	0.919	3
JPEG2000 compression	0.970	0.958	**0.971**	0.959	0.968	0.950	0.905	0.952	2
JPEG transmission errors	0.894	0.896	**0.922**	0.861	0.802	0.848	0.818	0.841	3
JPEG2000 transmission errors	0.894	0.901	0.923	0.892	**0.959**	0.889	0.870	0.876	4
Non eccentricity pattern noise	0.786	0.773	**0.806**	0.794	0.734	0.797	0.759	0.772	4
Local block-wise distortions	0.479	0.589	0.171	0.553	0.210	0.480	**0.617**	0.531	6
Intensity shift	0.806	**0.814**	0.770	0.749	0.746	0.791	0.777	0.627	2
Contrast change	0.455	0.480	0.475	0.468	0.677	0.463	0.348	**0.838**	7
Change of color saturation	0.825	**0.871**	0.810	0.836	-0.022	-0.410	-0.406	-0.321	3
Multiplicative Gaussian noise	0.859	0.849	**0.912**	0.857	0.902	0.779	0.775	0.848	3
Comfort noise	0.909	0.892	0.924	0.914	**0.938**	0.853	0.819	0.895	4
Lossy compression of noisy images	0.933	0.926	**0.956**	0.949	0.938	0.907	0.911	0.920	4
Image color quantization with dither	0.900	0.896	0.884	0.882	**0.928**	0.855	0.789	0.842	2
Chromatic aberrations	**0.893**	0.884	0.891	0.893	0.882	0.878	0.888	0.885	1
Sparse sampling and reconstruction	0.962	0.951	0.963	0.958	**0.965**	0.948	0.903	0.936	3

Table 3. Performance comparison (SROCC) for the scenes on which MS-iCID performs the best (first row), performs the worst (second row) and ranks the worst (third row), in TID2013. Note that the latter two scenes can be considered as more complex and therefore more challenging for objective IQA than the picture of the woman's face.

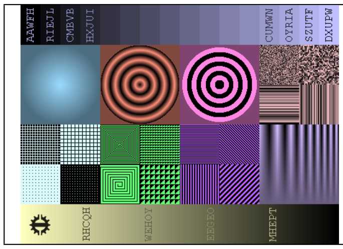
	MS-iCID	iCID	VSI	FSIMc	PSNRHA	MS-SSIM	SSIM	VIF	Rank MS-iCID
	0.920	0.850	**0.936**	0.909	0.880	0.862	0.613	0.757	2
	0.769	0.743	**0.785**	0.729	0.604	0.625	0.500	0.517	2
 AAWFH RIEJL CMEVE EKJUI CUMWN OYRIA SZUTF DXUPW RHCQH MHEPT	0.803	0.811	0.814	**0.851**	0.620	0.824	0.685	0.663	5

Table 4. Performance comparison (SROCC) of MS-iCID with and without saliency pooling

	MS-iCID	MS-iCID with saliency pooling (Signature)	MS-iCID with saliency pooling (SDSP)
TID2013	0.861	0.853	**0.863**
CSIQ	0.927	**0.938**	0.929
LIVE	0.879	**0.899**	0.886

MS-iCID performs especially well however on TID2013, as it yields the second best correlation. This is particularly noteworthy given that, unlike the current best-performing metrics (VSI and FSIMc), MS-iCID was not trained on any database in particular. Looking more specifically into the performances of MS-iCID for different kinds of distortions on TID2013 (Table 2), we observe that it ranks in the first three best metrics for 14 out of 24 distortions. It also ranks among the first two best metrics for 5 distortions: *masked noise*, *Gaussian blur*, *JPEG2000 compression*, *intensity shift* and *chromatic aberrations*. On *intensity shifts*, it is only outperformed by the regular iCID. This is due to the fact that the contribution of structural and contrast artifacts is considered higher in MS-iCID (see Equation 1), although in this particular case, intensity shifts are mostly conveyed through the $\mathbf{L}_L$ IDF. Furthermore, in comparison with the current best-performing metric (VSI), MS-iCID shows a clear amelioration on *masked noise* distortions (+0.066 SROCC) and *image color quantization with dither* distortions (+0.016 SROCC). We deduce from these observations that MS-iCID offers a good compromise in terms of performances for various distortions.

We would also like to draw the reader's attention on the fact that the significance of the difference between these SROCC is difficult to assess, partly due to inter- and intra-observer variabilities. Additionally, it is known that the process of subjective IQA involves both bottom-up and top-down/task-driven mechanisms, whose interaction is still not well understood.[14] Distortions occurring in highly salient places will compel observers to rate the image quality lower than if an equally strong artifact (in terms of e.g. RMSE) would occur in some less salient location (in the background).[15] Motivated by this, we also looked into how saliency pooling could ameliorate MS-iCID, using the Signature-based model by Hou *et al.*,[16] as well as the SDSP model,[17] used in the VSI metric.[6] Results are reported in Table 4 and demonstrate that forcing the metric to weight some regions of the image more than others can indeed yield improvement. Although one might argue on the significance of these improvements, we believe that further exploring the use of visual attention models in MS-iCID is an interesting research perspective, particularly with high-level features.

Finally, we also tested the MS-iCID on the gamut-mapping database used in.[5] We did not observe any noteworthy improvements over the iCID results reported in.[5] This is in agreement to the statement of Lissner et al.[3] that " [...] for the gamut-mapping database, the correlations between IDFs across scales are high". I.e., adding further scales does not give new information and therefore does not lead to an improvement on gamut-mapping databases.

4. CONCLUSIONS

We presented a multi-scale version of the iCID image-difference metric and tested it on three renowned image quality databases, including the 2013 version of the Tampere Image Database. We concluded that it is competitive with the state-of-the-art and best performing metrics such as VSI and FSIMc on all three databases. In particular, it outperforms all except VSI on TID2013, without dedicated training. It also outperforms its single scale version on the TID2013 and CSIQ databases.

ACKNOWLEDGMENTS

This work was supported by the Marie Curie Initial Training Networks (ITN) CP7.0 N-290154 funding.

REFERENCES

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

[2] Ponomarenko, N., Ieremeiev, O., Lukin, V., Egiazarian, K., Jin, L., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al., “Color image database TID2013: Peculiarities and preliminary results,” in [*4th European Workshop on Visual Information Processing EUVIP2013*], (2013).

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

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

[5] Preiss, J., Fernandes, F., and Urban, P., “Color-image quality assessment: From prediction to optimization,” *IEEE Transactions on Image Processing* **23**(3), 1366–1378 (2013).

[6] Zhang, L., Shen, Y., and Li, H., “VSI: A visual saliency induced index for perceptual image quality assessment,” *IEEE Transactions on Image Processing* **23**(10), 4270–4281 (2014).

[7] Reinhard, E., Khan, E. A., Akyz, A. O., and Johnson, G. M., [*Color imaging: fundamentals and applications*], AK Peters, Ltd. (2008).

[8] Lissner, I. and Urban, P., “Toward a unified color space for perception-based image processing,” *IEEE Transactions on Image Processing* **21**(3), 1153–1168 (2012).

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

[10] Zhang, L., Zhang, L., Mou, X., and Zhang, D., “FSIM: a feature similarity index for image quality assessment,” *IEEE Transactions on Image Processing* **20**(8), 2378–2386 (2011).

[11] Ponomarenko, N., Ieremeiev, O., Lukin, V., Egiazarian, K., and Carli, M., “Modified image visual quality metrics for contrast change and mean shift accounting,” in [*CAD Systems in Microelectronics (CADSM), 2011 11th International Conference The Experience of Designing and Application of*], 305–311, IEEE (2011).

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

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

[14] Borji, A. and Itti, L., “State-of-the-art in visual attention modeling,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **35**(1), 185–207 (2013).

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

[16] Hou, X., Harel, J., and Koch, C., “Image signature: Highlighting sparse salient regions,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **34**(1), 194–201 (2012).

[17] Zhang, L., Gu, Z., and Li, H., “SDSP: A novel saliency detection method by combining simple priors.,” in [*ICIP*], 171–175, Citeseer (2013).