https://doi.org/10.2352/issn.2169-2629.2019.27.47

Benchmark of 2D quality metrics for the assessment of 360-deg images

Mohamed-Chaker Larabi, Audrey Girard, Sami Jaballah, and Fan Yu
CNRS, Univ. Poitiers, XLIM, UMR 7252, France

Abstract

Omnidirectional or 360-degree images are becoming very popular in many applications and several challenges are raised because of both the nature and the representation of the data. Quality assessment is one of them from two different points of view: objectively or subjectively. In this paper, we propose to study the performance of different metrics belonging to various categories including simple mathematical metrics, humand perception based metrics and spherically optimized metrics. The performance of these metrics is measured using different tools such as PLCC, SROCC, KROCC and RMSE based on the only publically available database from Nanjing university. The results show that the metric that are considered as optimized for 360 degrees images are not providing the best correlation with the human judgement of the quality.

Introduction

For the past decade, visual media has made great efforts to provide innovative technologies that make the viewer experience more and more realistic. A classic example is 3D image/video technologies providing an additional dimension *i.e.* depth. More recently, emerging technologies such as Virtual Reality (VR) and Augmented Reality (AR) are seen as a further step towards an immersive experience, currently known as immersive Media.

Recent advances in acquisition devices, graphics processing performance and interactive display systems such as HMDs (head mounted devices) have contributed to the rapid deployment of virtual reality applications. One of the most important VR applications to mention is based on omnidirectional or 360-degree images. The latter gives the user the feeling of being part of the visualized scene.

Various technologies can be used to capture a 360-degree image, such as multi-camera array requiring an additional imaging step to obtain a spherical signal, or 360 multi-lenses imaging systems. The omnidirectional content viewed with HMD devices positions the viewer in the center of the sphere, allowing him to freely change his point of view by simply moving his head in the desired direction. Depending on the viewing direction of the user, a part of the spherical surface is made visible. The latter represents a small part of the whole picture/video and is called field of view (FoV) or viewport.

360 imaging is relatively a new field and will require an important effort in order to reach stability in terms of comfort, immersiveness and quality. The latter aspect is vital in order to guarantee a good experience for the user. However, the field for 360 images/videos is still in its infancy and may benefit from the large experience gained by the community in 2D and 3D quality assessment (objective and subjective). Indeed the last decade witnessed hundreds of works related to this topic. This activity is due to the interest of the scientific community in having tools allowing to measure accurately the fidelity of a given algorithm. Moreover, the diversity of content types (2D, S-3D, LDR, HDR, ...), applications (multimedia, medical, security, ...), impairment natures (blocking, blurring, ringing, color shift, ...), created new needs to be addressed. Hence, 360 and stereoscopic 360 are new challenges for the community.

This paper aims at benchmarking a set of quality metrics from the literature, initially developed for 2D images, on this new type of content. The expectation from such a study is to conclude about the usability of the metrics for 360 degrees content. Therefore, several metrics are selected either for their efficiency or for their widespread use in the community. To date several reviews, surveys and chapters have been published regarding image quality metrics for different application fields [1, 2, 3, 4, 5, 6, 7, 8]. With the increase of the activity around the 360 content, some attempts have been made to adapt well known metrics such as the PSNR to the specific characteristics of 360. A selection from these proposals is added to the set of metrics used in this benchmark. Finally, a benchmark cannot be performed without the availability of databases providing opinion scores of observers. The latter aspect is one of the weaknesses of this field because only a few databases exist. Furthermore, there no commonly accepted paradigm to run subjective testing using such a new content and such new devices.

The remainder of this paper is organized as follows: section 2 provide a description of the used quality metrics together with the database allowing to conduct performance evaluation. Section 3 provides the benchmark's results and discusses the performance of each metric for the different types of content. Finally, this paper ends with some conclusions and open questions regarding 360 quality assessment.

Image quality metrics

By browsing the literature one can notice that some metrics, namely MSE, PSNR, SSIM [15], IFC [9], VIF [23], VSNR [10], are often used as anchors to demonstrate the efficiency of a proposed metric. For this study, we intended to cover a large panel of categories of metrics, including the best performing ones. We thus opted for signal-based, structural similarity, feature similarity, information fidelity metrics in addition to a perceptually weighted metric and a visual difference predictor. In addition, we included the first attempts to fit the metrics with this new type of content and whose are mostly used in the compression field.

Signal fidelity metrics: MAE, PSNR, MSE and PAMSE

The mean absolute error (MAE) is one of the simplest metrics. It calculates the residual between the reference and the impaired images for every pixel, taking only the absolute value of each so that negative and positive residuals do not cancel out. MAE describes the typical magnitude of the residuals and is formulated as:

$$MAE(I,I') = \frac{1}{N}|I - I'|, \tag{1}$$

with I and I' are two images being compared and N is the number of pixels.

Peak Signal-to-Noise Ratio (PSNR) is undoubtedly the most widely used metric to date even with the advent of an impressive number of metrics. It is very popular especially because of the simplicity of its description, its good understandability and low complexity. PSNR comes originally from the signal processing community and aims at assessing fidelity thanks to the ratio between the maximum possible power of a signal and the power of corrupting noise affecting it. PSNR has been widely used for the evaluation of image compression algorithms where the original signal is the unaltered image and the impairments introduced by codecs represent the noise.

PSNR is expressed in terms of the logarithmic decibel (dB) scale based on the mean square error (MSE) as described below.

$$PSNR = 10\ log_{10} \frac{I_{max}^2}{MSE} \tag{2}$$

where I_{max} is the maximum possible value of a pixel, and MSE represents:

$$MSE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} [I(x,y) - I'(x,y)]^2 \tag{3}$$

with I and I' are two images being compared, $I(x,y)$ is the value of the pixel (its luminance) and W (resp. H) is the width (resp. height) of the images. For 8-bit images, I_{max} is 255. Hence, PSNR equation can be written:

$$PSNR = 20\ log_{10} \frac{255}{\sqrt{MSE}} \tag{4}$$

There is no standard formulation of PSNR for multi-channel (*e.g.* RGB) or multi-view (*e.g.* stereoscopic) images. Several authors use either an average of MSEs over all channels / views or an average of individual PSNRs.

Typical values of PSNR for lossy compression range between 30 and 50 dB where higher values indicate better fidelity. However, one must be very careful when interpreting results because their validity highly depends on the content, the compared algorithms and the spatial or spectral distribution of the noise [11]. Moreover, two impaired images may obtain the same PSNR value while their visual inspection indicates an important difference of quality. For instance, if the same noise is applied with a uniform spatial distribution on an image or concentrated on one object of interest, this will provide the same PSNR value while the perceptual impact is incomparable. Therefore, the perceptual validity of PSNR is very disputable and several studies demonstrated its inefficiency on specific impairments.

By having a close look at the PSNR formula, on can notice that MSE is playing an important role in this inefficiency. Wang and Bovik conducted a very important study entitled "*Mean Squared Error: Love It or Leave It?*" where they demonstrated the lacking efficiency of this metric while being confident that it has many years ahead of it [12].

For the selection of patches, we apply a recently developed metric, called the perceptual-fidelity aware mean squared error (PAMSE) [13] to find the best matched patches. PAMSE is a Gaussian-smoothed MSE that is computationally efficient in comparison to other metrics. The PAMSE approximates a weighted sum of the gradient of MSE, the Laplacian of MSE, and MSE itself. It is formulated as follows:

$$PAMSE(I,I') = \frac{1}{N} \|h \otimes (I - I')\|_2{}^2, \tag{5}$$

where h is a Gaussian smooth filter and N is the number of pixels.

MS-SSIM: Multi-Scale Structural Similarity Index Metric

Structural similarity has been proposed as a computational way to overcome the drawbacks of traditional perceptual image quality metrics. The latter are often based on several stages mimicking the HVS behavior like: 1) the pre-processing stage taking into account the low-pass filtering performed by the eye, 2) the channel decomposition stage reproducing the operations performed at the visual cortex, 3) the error normalization process providing the ability to weight the error in each channel and finally 4) the error pooling stage combining the whole information into a single score per pixel or per image. Despite the complex structure of these metrics, their performance may be debatable in comparison with their computational cost.

The Structural Similarity Index Metric (SSIM) proposed by Wang et al. [15] has taken an important place in the quality evaluation community and beyond, thanks to the very interesting tradeoff between complexity and correlation with the human judgement. However, it is quite difficult to achieve the unanimity of users when dealing with this metric. Indeed, it is somehow difficult to interpret results when two impaired images are close in terms of quality leading to similar conclusions as for PSNR about the usability. Moreover, as reported by the authors themselves, SSIM is a single-scale metric while the viewing conditions are varying. To cope with this limitation, an extension of this metric called Multi-Scale SSIM (MS-SSIM) has been proposed [14]. The extension inherits all the features introduced in the single-scale version. To avoid redundancy, we focused here only on the extended version of the structural similarity.

The MS-SSIM metric takes as input the reference and impaired images and compares two features called contrast c and structure s defined by:

$$c(I,I') = \frac{2\sigma_I \sigma_{I'} + c_1}{\sigma_I^2 + \sigma_{I'}^2 + c_1}, \tag{6}$$

$$s(I,I') = \frac{\sigma_{II'} + c_2}{\sigma_I + \sigma_{I'} + c_2}, \tag{7}$$

where σ_* and σ_{**} represent the variance and the covariance of the luminance, respectively. c_* are constants used for computation stability.

This processing of scale 1 is iterated at every scale and moving from scale to scale is performed by applying a low-pass filter and downsampling the filtered image by a factor of 2 until scale M. While contrast and structure are computed at each scale, another feature called luminance l is computed only on the smallest scale (M) as described below:

$$l_M(I,I') = \frac{2\mu_I \mu_{I'} + c_3}{\mu_I^2 + \mu_{I'}^2 + c_3}. \tag{8}$$

where μ_* is the luminance mean. The MS-SSIM score is obtained by computing and combining the aforementioned features on local image patches i at different scales j, as described by equation 9.

$$MS{-}SSIM_i = [l_M(I_i, I_i')]^{\alpha_M} \prod_{j=1}^{M} [c(I_{j,i}, I_{j,i}')]^{\beta_j} [s(I_{j,i}, I_{j,i}')]^{\gamma_j} \tag{9}$$

Exponents α_M, β_j and γ_j are used to adjust the relative importance of the different features. The weighting values are obtained by means of a psycho-physical study conducted with a panel of ten observers. These local scores are then averaged into a single score per image.

FSIM: Feature Similarity Index Metric

It is known that the HVS is attracted by low-level features, such as edges or zero-crossings, on a given image. This is due to the important information conveyed by the most salient feature helping in the scene interpretation task. Therefore, any noticeable change on an image will be naturally expressed by the discussed features. Based on these observations, Zhang et al. proposed to exploit two important features namely phase congruency (PC) and gradient magnitude (GM) [16]. The former, PC, raised from the idea that features are considered as noticeable at points where the phase is maximal for the Fourier components. This is confirmed by physiological and psychophysical studies on how the mammalian visual system detects and identifies salient features in an image. The latter feature, *i.e. GM*, is used to cope with the fact that PC is contrast invariant while the HVS is sensitive to image local contrast and this important feature has to be taken into account.

Based on the above description, Zhang et al. formulated their proposed metric, based on phase congruency and gradient magnitude, as given by equation 10.

$$FSIM = \frac{\sum_i S_i \cdot PC_i^{max}}{\sum_i PC_i^{max}} \tag{10}$$

where $\sum_i$ represents a summation over local image patches i; $PC_i^{max} = max(PC_{I_i}, PC_{I'_i})$ and $S_i = [S_{PC_i}]^{\alpha}[S_{GM_i}]^{\beta}$ is the weighted combination of PC and GM similarities between the original image patch I_i and the impaired one I'_i. While α and β may be used to give more importance to one feature or the other, the author made the choice to give them an equal importance. Similarity between PCs (S_{PC}) and GMs (S_{GM}) are expressed as given below:

$$S_{PC} = \frac{2PC_I \cdot PC_{I'} + c_1}{PC_I^2 + PC_{I'}^2 + c_1}, \tag{11}$$

$$S_G = \frac{2GM_I \cdot GM_{I'} + c_2}{GM_I^2 + GM_{I'}^2 + c_2}. \tag{12}$$

Finally, the feature $GM = \sqrt{G_x^2 + G_y^2}$ is computed using horizontal and vertical gradients by an operator such as Sobel, Prewitt or Scharr; the latter behaves slightly better as reported by the authors. *PC* meanwhile is expressed as the ratio between the local energy E_{θ_j} along orientation θ_j and the local amplitude A_{n,θ_j} on scale n and orientation θ_j as shown below:

$$PC = \frac{\sum_j E_{\theta_j}}{c_3 + \sum_n \sum_j A_{n,\theta_j}} \tag{13}$$

Constants c_* are used for computation stability.

Visual information fidelity: VIF

Based on the information theory, Sheikh et al. proposed a visual fidelity metric for image quality assessment [23]. It is seen as a measure allowing to quantify the level of information that can be extracted by the brain from a given scene. Hence, the metric rely on natural scene statistics (NSS), HVS properties and a distortion model. This metric comes from the extension of a previous work by the same authors [9] in which they proposed an information theoretic criterion for image fidelity based on NSS.

The assumption behind the VIF metric is that the random field (RF) from a wavelet decomposition subband of an image, $RF_{I'}$, can be defined as:

$$RF_{I'} = G.RF_I + V \tag{14}$$

where RF_I is the random field of the subband from the reference image, G is a deterministic scale gain field, and V is a stationary additive zero-mean Gaussian noise random field.

We have chosen to shorten the description of this metric because its mathematic demonstration is long and can hardly be summarized. The reader can refer to papers [9, 23] for a complete description and demonstration. Finally, this metric is rather efficient but its main drawback is that the provided information is a single score and not a map gathering the spatial distribution of the perceived distortion.

Gradient Magnitude Similarity Deviation (GMSD)

The Gradient Magnitude Similarity Deviation metric (GMSD) has emerged as a highly efficient model for post-compression quality analysis [19]. Like many other full reference quality metrics, it computes image quality in two steps. First, a Local Quality Map (LQM) is computed by locally comparing original and impaired images. The overall quality is then determined from the LQM by using a pooling strategy. Several choices exist for the pooling strategy where the simplest is the average one. However, since different parts of the image contribute differently to image quality, a weighted pooling strategy yields better results. GMSD employs standard deviation for obtaining the overall quality score from the similarity map.

Visual difference predictor (VDP): HDR-VDP-2

The VDP-like metrics follow a bottom-up approach when comparing the input images, *i.e.* original and impaired, in general. The framework of these predictors is highly based on the simulation of several processes happening in the HVS. The VDP metric introduced by Scott Daly [18] was interesting because of the proposed HVS simulation rather than its efficiency. It is only applicable to low-dynamic range (LDR) images and its complexity was a real issue at that time. An extension of VDP to higher dynamic range, called HDR-VDP, has been proposed by Mantiuk et al. [20, 21]. It operates on the full range of luminance, but cannot be applied to strongly distorted images, since it is considered as a near-threshold metric. Unfortunately, both discussed metrics have not been rigorously experimented and calibrated using psychophysical experiments.

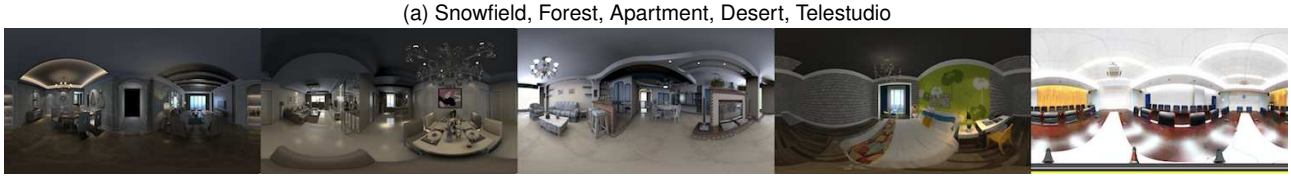
Despite its name, HDR-VDP-2 is considered as a breakthrough solution in comparison to the aforementioned metrics [24]. It provides three types of maps and relies on both a comprehensive model of the HVS characteristics and a sound extension to a broad range of viewing conditions. Hence, the visual difference predictor models the optical and retinal pathway taking into account 1) the light scattering happening at various levels, especially with HDR scenes, and 2) the spectral sensitivity of rods and LMS cones in addition to the luminance masking effect due to their regulation of the incoming light. At a higher level of the HVS, HDR-VDP-2 considers the overall noise affecting each subband of the multi-scale decomposition as an accumulation of: 1) a signal independent noise obtained from the contrast sensitivity function measurements, and 2) a signal dependent noise related to contrast masking.

In this paper, we used the output Q described by equation 15 which comes from the pooling strategy proposed by the authors.

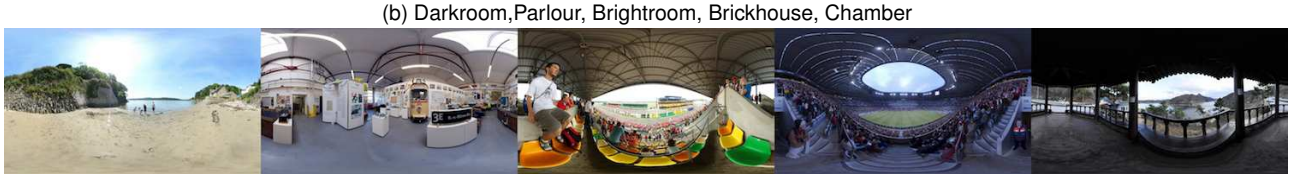
$$Q = \frac{1}{F.O} \sum_{f=1}^{F} \sum_{o=1}^{O} w_f \, log\left(\frac{1}{I} \sum_{i=1}^{I} D^2[f,o](i) + \varepsilon\right), \tag{15}$$

where i is the pixel index and I is their total number. $D[f,o]$ is the noise-normalized signal difference for the f^{th} spatial fre-

(a) Snowfield, Forest, Apartment, Desert, Telestudio



(b) Darkroom,Parlour, Brightroom, Brickhouse, Chamber



(c) Seaside, Fair, Race, Stadium, Pavilion

Figure 1: Thumbnails of the 15 original 4K images in the database.

quency band and o^{th} orientation, w_f is the per-band weighting and $\varepsilon = 10^{-5}$ is used to avoid computation instabilities.

Recently, an extension named HDR-VDP-2.2 has been proposed where the main changes lie in the improvement of the frequency-based pooling considered as a constrained optimization problem [22]. We used this extension for the purpose of this study since it showed the best performance.

Spherical extensions of PSNR

Because of the progress of 360 image coding approaches, the compression community defined or exploited several extensions of the PSNR with the aim to account for the specificity of omnidirectional images [26].

Weighted spherical PSNR (WS-PSNR)

For computing WS-PSNR, squared error of each pixel (i, j) of the tested frames is weighted by the scaling factor $w(i, j)$ corresponding to the ratio between the tested projection and the unit spherical surface. The distribution of $w(i, j)$ is illustrated in Fig. 2 for ERP projection. For a reference frame R and tested frame T, WS-PSNR is calculated as:

$$WS-PSNR = \sum w(i,j)(R(i,j)-T(i,j))^2 \quad (16)$$

Figure 2: Spatial distribution of weight for WS-PSNR for ERP projections.

Spherical PSNR w/o interpolation (S-PSNR-NN)

Calculate PSNR based on a set of points uniformly sampled on the sphere. To find the sample value at the corresponding position on the projection plane, nearest neighbor rounding is applied. The two inputs to the metric calculation can have different resolutions and/or projection formats

Spherical PSNR with interpolation (S-PSNR-I)

Calculate PSNR based on a set of points uniformly sampled on the sphere. To find the sample value at the corresponding position on the projection plane, bicubic interpolation is applied. The two inputs to the metric calculation can have different resolutions and/or projection formats

Experiments

This section describes the testing conditions including the database and associated subjective experiments. In order to compare the performance of the different metrics on the used database, the definition of the used performance metrics (SROCC, PLCC, KROCC and RMSE) are given. Finally, this section gives the results and discusses the usability of the metrics depending on the resolution and quality factors.

Testing conditions

For these experiments, we used the 360 image database from Nanjing university [25]. It is composed of 15 images of 4k resolution in bmp format. In order to increase the reference images and the testing conditions, images were used with four different resolutions using the bi-cubic interpolation: 4096×2160 (4K), 2560×1440 (2K), 1920×1080 (1K) et 1280×720 (720p). Images are then distorted using JPEG compression with three quality factor 100 (high), 60 (medium), and 25 (low).

The test protocol used the HTC Vive HMD, a personal computer with a CPU i7-6700K, GPU nvidia gtx 1070. The ACR Single Stimulus methodology was adopted. Subjects were asked to freely navigate in the 360 environment. Images (resolution 4k to 720p and a quality factor from 100 to 25)are shown in a random way. The exposition time is fixed to 20s with gray image shown during 5s between the tests. Scores are collected orally on a continuous scale [0,100]. These scores are then normalized using Z-scores and averaged per image.

Performance metrics

Pearson's correlation coefficient (PLCC)

Pearson's correlation coefficient r [27] is used for data on the interval or ratio scales, and is based on the concept of covariance. When an X, Y sample are correlated they can be said to covary; or they vary in similar patterns.

The product-moment r statistic is given by :

$$PLCC = \frac{n\sum_{i=0}^{n} X_i Y_i - (\sum_{i=0}^{n} X_i)(\sum_{i=0}^{n} Y_i)}{\sqrt{\left(\left[n\sum_{i=0}^{n} X_i^2 - (\sum_{i=0}^{n} X_i)^2\right]\left[n\sum_{i=0}^{n} Y_i^2 - (\sum_{i=0}^{n} Y_i)^2\right]\right)}} \quad (17)$$

where n is the number of pairs of scores. The degree of freedom is $df = n2$.

Spearman rank order correlation (SROCC)

The Spearman rank correlation coefficient [29], r_s (or Spearman's rho), is used with ordinal data and is based on ranked scores. Spearman's rho is the nonparametric analog to Pearson's r.

The process for Spearman's correlation first requires ranking the X and Y scores: the analysis is then performed on the ranks of the scores, and not the scores themselves. The paired ranks are then subtracted to get the values of d, which are then squared to eliminate the minus sign. If there is a strong relationship between X and Y then paired values should have similar ranks. The test statistic is given by :

$$SROCC = 1 - \frac{6\sum_{i=0}^{n} d_i^2}{n(n^2-1)} \tag{18}$$

Kendall rank order correlation (KROCC)

The Kendall correlation coefficient can be computed as the difference between two probabilities related to the same and inverse order of both lists [28].

$$KROCC = 2.\frac{P-Q}{n(n-1)} \tag{19}$$

where P corresponds to the score pairs corresponding to the same image located in the same order and Q in the reverse one.

Root Mean Square Error (RMSE)

The root mean square error (RMSE) indicates the accuracy and precision of the model and is expressed in the original units of measure. Accurate prediction capability is indicated by a small RMSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{N}(X-Y)^2} \tag{20}$$

results and discussion

The results obtained on the whole database can be found on Table 1. It gives the correlation coefficients and RMSE on all images. One can notice that the best metric in this case is by far S-PSNR-NN, followed by $FSIM_c$ which is not optimized for spherical content. Besides this couple of metrics, most of the other selected metrics are behaving similarly except HDR-VDP-2 which is giving very bad results. These results are relatively surprising because one would have expected that spherical-optimized metrics will appear among the top ones. The fact of having $FSIM_c$ as the second best is also surprising. Besides, this result shows that color is having an interesting impact on the performance when comparing the results with $FSIM$.

Table 2 gives more detailed results depending on the used resolution. For 4K, $MS-SSIM$ provides the best performance whatever the used coefficient. It is followed by *GMSD* and $FSIM_c$. The performance are relatively good for almost all the metrics. The least performing metric is $HDR-VDP-2$. Nevertheless even though the performance of the spherically-based metrics are average, they do not appear among the best metric. This is questioning because it should perform better because it is taking into account the geometrical aspect. The same conclusion/ranking does not hold for 2K. The results are mixed. For instance, *PLCC* and *RMSE* considers $FSIM_c$ as the best while SROCC/KROCC consider PAMSE. Again the spherical-based metrics are not ranking well. The results for 1K is also different because it almost considers GMLD as the best. At 720p, FSIM-c is again the best while *MAE* is incredibly low.

Table 1: Performance metric on the whole database.

Metrics	PLCC	SROCC	KROCC	RMSE
MAE	0.485	0.434	0.297	11.819
VIFp	0.449	0.43	0.294	12.08
PAMSE	0.486	0.467	0.32	11.813
GMSD	0.493	0.465	0.318	11.763
SSIM	0.513	0.46	0.315	11.605
MS-SSIM	0.457	0.412	0.278	12.022
FSIM	0.692	0.667	0.487	9.753
FSIMc	0.716	0.685	0.504	9.437
PSNR	0.478	0.486	0.335	11.872
WS-PSNR	0.47	0.494	0.344	11.971
S-PSNR-NN	**0.751**	**0.723**	**0.548**	**8.917**
S-PSNR-I	0.585	0.526	0.386	10.962
HDR-VDP-2	0.235	0.227	0.148	13.136

Table 3 gives the performance results per quality factor. While the compression quality is high, surprisingly S-PSNR-NN is ranking the best. However for the medium and low quality factors, $FSIM_c$ is considered as the best whatever is the performance metric. Nevertheless, the results of S-PSNR-N are still showing very good correlation. The other metrics are performing relatively low except FSIM and S-PSNR-I. WS-PSNR provide very bad results compared to the other spherically-based metrics.

Conclusion

In this paper, the aim was to benchmark 13 quality metrics from the literature, most of them initially developed for 2D images, on a 360-deg images database. The selected metrics cover a large spectrum in terms of quality assessment categories. The results obtained on a publicly available database have lead to rather surprising conclusions. For instance, the metrics optimized for such a content are not providing the good results. The results vary depending on the resolution and the quality factor. However, if one is requested to give a recommendation, $FSIM-C$ would represent the best tradeoff for the used database. This initial work highlighted the difficulty of finding appropriate metrics for 360. Hence, as a future work, a comprehensive database should be built by taking into account the different characteristics of such media. Besides, novel approaches that are fully dedicated to 360 are needed instead of weighting existing metrics.

References

[1] K. Seshadrinathan and A. Bovik, Automatic prediction of perceptual quality of multimedia signals - a survey, Int. Journal of Multimedia Tools and Applications, vol. 51, no. 1, pp. 163- 186, 2011.

[2] A. Moorthy and A. Bovik, Visual quality assessment algorithms: What does the future hold? International Journal of Multimedia Tools and Applications, vol. 51, no. 2, pp. 675-696, 2011.

[3] L. Zhang, A comprehensive evaluation of full reference image quality assessment algorithms, International Conference on Image Processing (ICIP), pp. 1477-1480, 2012.

[4] D. M. Chandler, Seven challenges in image quality assessment: Past, present, and future research, Signal Proc., vol. 2013, p. 53, 2013.

[5] M.-C. Larabi, A. Saadane, and C. Charrier, Quality assessment approaches, in Digital Color. Wiley, 2013, pp. 265-306.

[6] M.-C. Larabi, A. Saadane, and C. Charrier Quality assessment of still images,in Advanced Color Image Processing and Analysis. Springer New York, 2013, pp. 423-447.

[7] A. Beghdadi, M.-C. Larabi, A. Bouzerdoum, and K. Iftekharuddin, A survey of perceptual image processing methods, Signal Processing: Image Communication, vol. 28, no. 8, pp. 811 - 831, 2013.

Table 2: Performance evaluation of the different metrics depending on the used resolution.

	4k				2k				1k				720p			
Metrics	PLCC	SROCC	KROCC	RMSE	PLCC	SROCC	KROCC	RMSE	PLCC	SROCC	KROCC	RMSE	PLCC	SROCC	KROCC	RMSE
MAE	0.69	0.695	0.499	3.53	0.756	0.789	0.582	4.976	0.793	0.794	0.578	4.719	0.129	0.151	0.086	2.87E+11
VIFp	0.784	0.692	0.509	3.029	0.792	0.828	0.634	4.643	0.849	0.845	0.638	4.093	0.75	0.763	0.553	5.674
PAMSE	0.791	0.704	0.513	2.985	0.839	**0.842**	**0.651**	4.133	0.885	0.856	0.642	3.607	0.851	0.768	0.551	4.478
GMSD	0.834	0.692	0.505	2.692	0.83	0.808	0.602	4.242	**0.908**	**0.867**	0.659	**3.244**	0.861	0.744	0.522	4.34
SSIM	0.782	0.682	0.493	3.038	0.778	0.784	0.592	4.771	0.827	0.84	0.634	4.358	0.756	0.726	0.527	5.575
MS-SSIM	**0.845**	**0.748**	**0.578**	**2.607**	0.849	0.83	0.616	4.019	0.89	0.866	**0.665**	3.523	0.848	0.761	0.553	4.516
FSIM	0.798	0.717	0.511	2.939	0.853	0.812	0.572	3.967	0.876	0.832	0.616	3.729	0.854	0.769	0.559	4.43
FSIMc	0.812	0.697	0.493	2.848	**0.856**	0.821	0.598	**3.931**	0.892	0.823	0.596	3.498	**0.863**	**0.776**	**0.561**	**4.302**
PSNR	0.64	0.677	0.489	3.75	0.522	0.649	0.475	4.25E+11	0.781	0.804	0.604	4.835	0.682	0.679	0.484	6.234
WS-PSNR	0.632	0.652	0.461	3.781	0.718	0.753	0.552	5.29	0.784	0.803	0.608	4.81	0.719	0.689	0.494	5.921
S-PSNR-NN	0.632	0.652	0.461	3.78	0.718	0.753	0.549	5.291	0.784	0.804	0.61	4.807	0.719	0.688	0.492	5.921
S-PSNR-I	0.632	0.652	0.461	3.78	0.479	0.572	0.444	6.672	0.545	0.539	0.418	6.49	0.5	0.455	0.332	7.382
HDR-VDP-2	0.475	0.515	0.349	4.292	0.549	0.612	0.42	6.355	0.538	0.609	0.404	6.525	0.508	0.539	0.367	7.343

Table 3: Performance evaluation of the different metrics depending on the quality factor

	Qf 100				Qf 60				Qf 25			
Metrics	PLCC	SROCC	KROCC	RMSE	PLCC	SROCC	KROCC	RMSE	PLCC	SROCC	KROCC	RMSE
MAE	0.189	0.212	0.155	10.701	0.292	0.299	0.206	10.826	0.017	0.097	0.066	3.14E+11
VIFp	0.134	0.127	0.088	9.4E+11	0.235	0.171	0.127	11.004	0.211	0.182	0.122	13.325
PAMSE	0.085	0.131	0.094	10.859	0.326	0.245	0.168	10.707	0.3	0.235	0.163	13.003
GMSD	0.066	0.071	0.042	10.875	0.319	0.254	0.168	10.728	0.369	0.278	0.197	12.67
RECO	0.463	0.421	0.302	9.662	0.217	0.215	0.155	11.051	0.167	0.18	0.129	13.474
SSIM	0.12	0.063	0.041	10.823	0.321	0.25	0.171	10.723	0.264	0.222	0.155	13.149
MS-SSIM	0.12	0.038	0.027	10.825	0.007	0.055	0.047	11.32	0.261	0.047	0.019	13.258
FSIM	0.767	0.741	0.551	6.987	0.823	0.818	0.615	6.438	0.773	0.784	0.573	8.645
FSIMc	0.798	0.776	0.567	6.567	**0.891**	**0.882**	**0.686**	**5.131**	**0.83**	**0.832**	**0.629**	**7.613**
PSNR	0.162	0.176	0.127	10.758	0.354	0.347	0.241	10.587	0.361	0.323	0.223	12.715
WS-PSNR	0.159	0.165	0.127	10.76	0.342	0.336	0.235	10.644	0.387	0.282	0.188	12.571
S-PSNR-NN	**0.901**	**0.88**	**0.69**	**4.726**	0.818	0.818	0.62	6.519	0.806	0.802	0.621	8.07
S-PSNR-I	0.301	0.258	0.202	10.392	0.633	0.594	0.452	8.764	0.5	0.557	0.398	12.483
HDR-VDP	0.401	0.398	0.279	9.986	0.291	0.27	0.186	10.832	0.306	0.276	0.203	12.981

[8] L. He, F. Gao, W. Hou, and L. Hao, Objective image quality assessment: a survey, Int. Jour. of Computer Mathematics, pp. 1-15, 2013.

[9] H. Sheikh, A. Bovik, and G. de Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Trans. Image Process., vol. 14, no. 12, pp. 2117-2128, 2005.

[10] D. Chandler and S. Hemami, VSNR: a wavelet-based visual signal-to-noise ratio for natural images, IEEE Trans. Image Process., vol. 16, no. 9, p. 2284?2298, 2007.

[11] Q. Huynh-Thu and M. Ghanbari, The accuracy of psnr in predicting video quality for different video scenes and frame rates, Telecommunication Systems, vol. 49, no. 1, pp. 35-48, 2012.

[12] Z.Wang and A.Bovik,Mean squared error:Love it or leave it? a new look at signal fidelity measures, Signal Processing Magazine, IEEE, vol. 26, no. 1, pp. 98-117, 2009.

[13] W. Xue, X. Mou, L. Zhang, X. Feng, Perceptual fidelity aware mean squared error, IEEE Intl. Conf. Comp. Vis., 2013, pp. 705-712.

[14] Z.Wang,E.P.Simoncelli,and A.C.Bovik,Multi-scalestructural similarity for image quality assessment, in Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers, 2003, pp. 1398- 1402.

[15] 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 Proc., vol. 13, no. 4, pp. 600-612, avril 2004.

[16] L. Zhang, X. Mou, and D. Zhang, FSIM: A Feature Similarity Index for Image Quality Assessment. IEEE Transactions on Image Processing, no. 99, pp. 1-1, Jan. 2011.

[17] Z.Wang and Q.Li,Information content weighting for perceptual image quality assessment, IEEE Transactions on Image Processing, vol. 20, no. 5, pp. 1185 - 1198, 2011.

[18] S. Daly, The visible differences predictor: an algorithm for the assessment of image fidelity, in Digital images and human vision, Andrew B. Watson, Ed. Cambridge: MIT Press, Oct. 1993, pp.179-206.

[19] W. Xue, L. Zhang, X. Mou, A.C. Bovik, Gradient magnitude similarity deviation: a highly efficient perceptual image quality index. IEEE Trans. Image Process. 23(2), 684-695 (2014)

[20] R. Mantiuk, K. Myszkowski, and H.-P. Seidel, Visible difference predicator for high dynamic range images, IEEE Int. Conference on Systems, Man and Cybernetics, vol. 3, pp. 2763-2769, 2004.

[21] R. Mantiuk, S. J. Daly, K. Myszkowski, and H.-P. Seidel, Predicting visible differences in high dynamic range images: model and its calibration, in Proc. SPIE, vol. 5666, 2005, pp. 204-214.

[22] M. Narwaria, R. K. Mantiuk, M. P. Da Silva, and P. Le Callet, HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images, JEI, 24(1) 2015.

[23] H. Sheikh and A. Bovik, Image information and visual quality, IEEE Trans. Image Process., vol. 15, no. 2, pp. 430-444, 2006.

[24] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, Hdr-vdp-2 : A calibrated visual metric for visibility and quality predictions in all luminance conditions, ACM Siggraph, 2011.

[25] M. Huang, Q. Shen, Z. Ma, X. Cao, A. C. Bovik and P. Gupta, "Modeling the Perceptual Quality of Immersive Images Rendered on Head Mounted Displays: Resolution and Compression".

[26] Zakharchenko, V., Choi, K.P., Park, J.H.: Quality metric for spherical panoramic video. In: Optics and Photonics, SPIE Optical Engineering + Applications, San Diego, 2016. pp. 1-9.

[27] Pearson, E.S., Hartley, H.O.: Biometrika Tables for Statisticians,vol.1.Cambridge University Press (1966)

[28] Kendall, M.G.: Rank Correlation Methods. Charles Griffin Company, Ltd., London (1975).

[29] Huck, S., Cormier, W.H.: Reading statistics and research. Harper Collins (1996).