This copy represents the peer reviewed and accepted version of paper:
Farid, M.S., Lucenteforte, M., Grangetto, M.
"**Objective quality metric for 3D virtual views**," in proc. IEEE International Conference on Image Processing 2015.
doi: 10.1109/ICIP.2015.7351499

The published version is available at
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7351499

OBJECTIVE QUALITY METRIC FOR 3D VIRTUAL VIEWS

Muhammad Shahid Farid, Maurizio Lucenteforte, Marco Grangetto

Dipartimento di Informatica, Università degli Studi di Torino
Corso Svizzera 185, 10149 Torino, Italy

ABSTRACT

In free-viewpoint television (FTV) framework, due to hardware and bandwidth constraints, only a limited number of viewpoints are generally captured, coded and transmitted; therefore, a large number of views needs to be synthesized at the receiver to grant a really immersive 3D experience. It is thus evident that the estimation of the quality of the synthesized views is of paramount importance. Moreover, quality assessment of the synthesized view is very challenging since the corresponding original views are generally not available either on the encoder (not captured) or the decoder side (not transmitted). To tackle the mentioned issues, this paper presents an algorithm to estimate the quality of the synthesized images in the absence of the corresponding reference images. The algorithm is based upon the cyclopean eye theory. The statistical characteristics of an estimated cyclopean image are compared with the synthesized image to measure its quality. The prediction accuracy and reliability of the proposed technique are tested on standard video dataset compressed with HEVC showing excellent correlation results with respect to state-of-the-art full reference image and video quality metrics.

Index Terms— Quality assessment, depth image based rendering, view synthesis, FTV, HEVC

1. INTRODUCTION

Depth perception in 3D television (3DTV) is achieved by rendering two views of the scene captured at sightly different viewpoints. The latest 3D display technologies - free-viewpoint television (FTV) [1] and future Super Multiview (SMV) displays [2], are capable to provide hundreds of high resolution views with base line distance smaller than interocular distance [3]. Both FTV and SMV require huge number of views to provide the viewer the freedom to roam around a scene. Due to various hardware, economic and bandwidth constraints acquisition and transmission of such huge number of views is not possible. Therefore, only few views are captured and transmitted; the rest are synthesized on the receiver side.

This work was partially supported by Sisvel Technology research grant.

High quality immersion requires efficient 3D content representation and compression as well as computationally viable view synthesis techniques to generate good quality novel views. Multiview videos plus depth (MVD) format [4] has gained widespread acceptability due to its provision for intermediate virtual views generation and efficient compression. MVD has been adopted for future FTV and SMV technologies for both compression and display [5]. In MVD format in addition to texture images, gray scale depth maps representing the per pixel depth value are also available which permit the generation of novel views through *Depth Image Based Rendering* (DIBR) techniques [6].

Efficient compression of MVD data is central to 3D television processing chain and a number of compression friendly MVD data representations e.g., [7–9] have been proposed resulting in the developments of novel codecs. With the introduction of the novel state-of-the-art High Efficiency Video Coding (HEVC) [10] the Joint Collaborative Team on 3D Video Coding Extension Development (JCT-3V) has recently developed extensions of HEVC: Multiview-HEVC (MV-HEVC) and 3D-HEVC [11]. 3D-HEVC exploits additional coding tools and achieves the best compression ratio for MVD data [11] and it also guarantees promising quality virtual views with View Synthesis Optimization (VSO) coding tool. To enable autostereoscopy additional views on the receiver side are generated through DIBR. Since there are only few original coded views, the overall user experience largely depends on the synthesized views.

Quality assessment of synthesized images is a challenging task as the corresponding reference images are not available. Furthermore, the generated images suffer from various types of artifacts e.g., compression artifacts [12], synthesis distortion [13], textural and structural distortions due to poor quality depth maps [14–16]. Recent studies [17, 18] tested various existing 2D image quality metrics to assess the quality of stereoscopic and synthesized images and concluded that none of them is suitable in this context. In the recent years a number of quality metrics for stereoscopic images have been proposed [19] which mostly extends the existing 2D quality metrics. However, little attention has been paid to the quality evaluation of synthesized images. In [20] quality of the virtual view is assessed by comparing the structures e.g. edges of the original and the warped images. It is limited to struc-

tural distortion estimation and cannot be used to represent the overall quality of the virtual image as it does not compute the color related artifacts. CSED (Color and Sharpness of Edge Distortion) [21] is another full reference quality metric to assess the quality of the virtual image. It targets the hole regions to assess the color distortion and uses the edge sharpness of the reference and virtual images to assess structural distortion. Battisti et al. [22] proposed to weight more the distortion created around the human body that largely affects the quality of the whole synthesized view. They proposed 3DSwIM metric which assesses the virtual image quality by detecting the skin regions and aligning them with the reference images to determine the distortion. View Synthesis Quality Assessment (VSQA) metric [23] combines SSIM with weighting functions derived from contrast, orientation and texture maps of the reference and synthesized views to assess the quality of the virtual pictures.

The previous discussion shows that existing techniques for quality evaluation of synthesized views are either full reference or reduced reference. However, in the FTV scenario a large number of views are synthesized whose references are not available. Therefore, the existing techniques become ineffective. This paper proposes a solution to the problem thanks to the following contributions:

- definition of a novel Synthesized Image Quality Evaluator (SIQE) to assess the quality of depth based rendered images in absence of corresponding reference images;
- exploitation of the cyclopean eye theory and divisive normalization transform to infer the quality of a picture rendered from a stereo pair (video plus depth);
- statistical analysis of the quality prediction accuracy obtained on a set of video plus depth sequences compressed with HEVC: the SIQE performance is compared versus other state-of-the-art full reference image and video quality metrics showing competitive results.

The rest of the paper is organized as follows: the proposed quality metric is described in Sect. 2, followed by experimental evaluation in Sect. 3. The conclusions are drawn in Sect. 4.

2. PROPOSED SYNTHESIZED IMAGE QUALITY EVALUATOR

The proposed quality metric is build around the Béla Julesz's Cyclopean Perception theory [24] and Divisive Normalization transform [25]. The Cyclopean Perception refers to the formation of a virtual image in our mind from the stimuli received from the left and the right eye. The cyclopean image (or the *mental image*), is a view obtained by fusing the left and the right views as if it was captured by a virtual eye (usually referred to as the cyclopean eye) placed in between the two eyes. This process is graphically shown in Fig. 1.

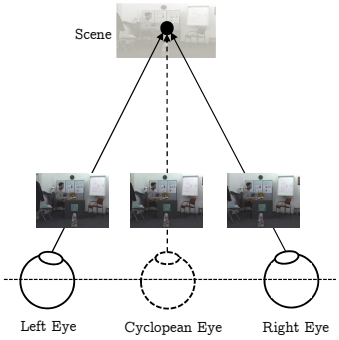

Fig. 1. The cyclopean image as perceived by mind.

Divisive normalization (DN) is based upon the standard psychophysical and physiological model and it has been used to study the nonlinear behaviors of cortical neuron in biological vision. The use of DN model in image quality assessment was pioneered by Teo and Heeger in [25]. Moreover, it is shown that DN model achieves statistical independence [26] and can be represented by Gaussian scale mixture (GSM) model. Using GSM model a number of quality assessment techniques have been proposed using various spatial and frequency based DN transforms e.g., [27, 28]. In the proposed quality model we exploit the divisive normalization to estimate the statistical characteristics of the uncompressed stereo images and fuse them together to obtain the statistical model of the cyclopean image. Depth image based rendering is used to generate the intermediate virtual image from the coded stereopair and the corresponding depth maps. The statistical model of this synthesized image is also computed by using the divisive normalization. This model is compared to the reference cyclopean image model to estimate the quality degradation in the synthesized image due to compression and 3D warping. Fig. 2 shows the block diagram of the whole algorithm.

Let V_l, V_r be the uncompressed left and right texture images and let V_s be a synthesized image obtained from the compressed texture and depth images (as shown in Fig. 2). The size of all the views is $M \times N$. Let T_l, T_r and T_s are the divisive normalized images of V_l, V_r and V_s respectively which are created in the spatial domain similar to [28, 29].

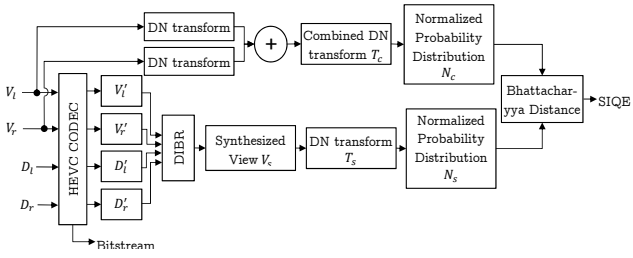

Fig. 2. Block diagram of proposed SIQE technique.

Any T_k is computed from V_k, $k \in \{l, r, s\}$ as follows:

$$T_k(u,v) = \frac{V_k(u,v) - \mu_k(u,v)}{\sigma_k(u,v) + \epsilon} \quad (1)$$

where ϵ is a small constant used to avoid division by zero. $\mu_k(u,v)$ and $\sigma_k(u,v)$ are local average and standard deviation computed over a block of size $m \times n$ centered at (u,v). These are computed as:

$$\mu_k(u,v) = \sum_{i=-m}^{m} \sum_{j=-n}^{n} w(i,j) V_k(u+i, v+j)$$

$$\sigma_k(u,v) = \sqrt{\sum_{i=-m}^{m} \sum_{j=-n}^{n} w(i,j)[V_k(u+i,v+j) - \mu_k(u,v)]^2}$$

where w is 2D symmetric Gaussian weight function computed as:

$$w(i,j) = \frac{1}{2\pi\sigma^2} e^{-\frac{i^2+j^2}{2\sigma^2}} \quad (2)$$

The statistical characteristics of the cyclopean image are estimated from the DN representations T_l, T_r of the left and the right views respectively. To this end, we propose to exploit the histogram of the DN image $T_c = [T_l|T_r]$, obtained by simple concatenation of the left and right images. In particular, from the histogram of T_c we compute the normalized probability distribution $\mathcal{N}_c$ using κ equally spaced bins; it follows that: $\sum_{i=1}^{\kappa} \mathcal{N}_c(i) = 1$. Then, the normalized distribution $\mathcal{N}_s$ of the DN synthesized image T_s is computed using the same number of bins. Fig. 3 shows the $\mathcal{N}_c$ and $\mathcal{N}_s$ curves for a sample image from Poznan_Street video sequence (see Tab. 1). The $\mathcal{N}_c$ models the distribution of the cyclopean image estimated from the left and the right views, whereas the $\mathcal{N}_s$ curves show the distribution of the intermediate synthesized image obtained through DIBR from the respective sequence coded at 6 different quality levels. The non-overlapped areas between the $\mathcal{N}_c$ and $\mathcal{N}_s$ curves represents the distortion in the respective synthesized image which can be estimated by computing the difference between the two distributions. To this end we propose to exploit the *Bhattacharyya coefficient* (ρ) since it has been already shown to be more reliable than other metrics, e.g. the Mahalanobis distance. The Bhattacharyya coefficient is used to estimate the similarity between the two distributions as follows:

$$\rho(\mathcal{N}_c, \mathcal{N}_s) = \sum_{x \in \kappa} \sqrt{\mathcal{N}_c(x)\mathcal{N}_s(x)} \quad (3)$$

Finally, the proposed SIQE metric is computed as the difference between the two models computed through the *Hellinger distance* [30], defined as:

$$SIQE = \sqrt{1 - \rho(\mathcal{N}_c, \mathcal{N}_s)} \quad (4)$$

The SIQE measures the distortion in the synthesized image i.e., smaller the value, better the quality of the image.

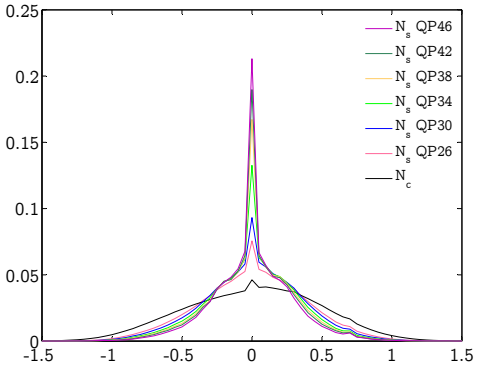

Fig. 3. Reference normalized distribution $\mathcal{N}_c$ and synthesized image normalized distributions $\mathcal{N}_s$ of the 1st frame from Poznan_Street test video sequence coded at QP={26,30,34,38,42,46}.

3. EXPERIMENTAL EVALUATION AND RESULTS

The performance of the proposed SIQE metric is evaluated on synthesized videos generated from compressed MVD video sequences, by comparing its grading with existing widely used full reference image and video quality metrics. All the results and the software to reproduce them are available at http://www.di.unito.it/~farid/3DQA/SIQE.html. We have used 4 standard video sequences listed in Tab. 1. Each video (texture plus depth), is independently encoded at 6 quality levels using the novel *High Efficiency Video Coding* (HEVC) [10] HM 11.0 reference software (main profile) with QP={26,30,34,38,42,46}. Using DIBR algorithm [31], 36 different intermediate view sequences are generated using a pair of videos plus depths coded with different QPs. This means a total of 144 synthesized videos are used in evaluation. In all experiments the parameters ϵ=1, m=7, n=7 and κ=300 are used.

The performance of the proposed quality metric is evaluated by comparing it with state of the art full reference video quality metrics as well as with popular 2D image quality metrics. According to Video Quality Expert Group (VQEG) FRTV Phase-I and Phase-II tests the NTIA/ITS video quality model (VQM) [32] performed the best with the highest correlation (0.91) with subjective testing. Due to its best

Table 1. Test Dataset Details. S# is sequence label, #F is the number of frames, V_l, V_r and V_s represent the left, right and the synthesized views respectively, and FR is the frame rate.

S#	Sequence	#F	V_l	V_r	V_s	Size	FR
S1	Poznan_Hall2	200	7	6	5	1920×1088	25
S2	Poznan_Street	250	5	3	4	1920×1088	25
S3	Book_Arrival	100	10	8	9	1024×768	16
S4	Balloons	300	1	5	3	1024×768	30

Table 2. Pearson Linear Correlation Coefficient (PLCC).

S#	VQM	FSIM	SSIM	MSSIM	iwSSIM	PSNR	UQI
S1	0.9211	0.8894	0.9098	0.8884	0.8848	0.6489	0.9725
S2	0.9457	0.8621	0.9165	0.9201	0.9038	0.8623	0.9937
S3	0.8941	0.8531	0.8707	0.8547	0.8463	0.8694	0.9679
S4	0.8485	0.8923	0.8974	0.9002	0.8884	0.7781	0.8799
Avg:	0.9024	0.8742	0.8986	0.8909	0.8808	0.7897	0.9535

Table 3. Root Mean Square Error (RMSE).

S#	VQM	FSIM	SSIM	MSSIM	iwSSIM	PSNR	UQI
S1	0.0412	0.0059	0.0050	0.0045	0.0135	0.5834	0.0099
S2	0.0506	0.0071	0.0094	0.0086	0.0158	0.6546	0.0122
S3	0.0597	0.0089	0.0108	0.0093	0.0177	0.7260	0.0164
S4	0.0653	0.0078	0.0094	0.0071	0.0153	0.6819	0.0275
Avg:	0.0542	0.0074	0.0087	0.0074	0.0156	0.6615	0.0165

performance we selected VQM for performance evaluation of proposed metric. In experiments NTIA General Model of VQM is used with spatial scaling and temporal registration (1 second) features. Moreover, we compare SIQE with 6 widely used full reference 2D image quality metrics - SSIM [33], FSIM [34], MSSIM [35], iwSSIM [36], Peak Signal to Noise Ratio (PSNR), and UQI [37]. For performance evaluation we use Pearson linear correlation coefficient (PLCC) for *prediction accuracy* test and root mean square error (RMSE) for *prediction error*. Before computing these parameters, according to VQEG recommendations [38] the scores are mapped with monotonic nonlinear regression. The following logistic function outlined in [39] is used for regression mapping:

$$Q_p = \beta_1 \left(\frac{1}{2} - \frac{1}{\exp \beta_2 (Q - \beta_3)} \right) + \beta_4 Q + \beta_5 \qquad (5)$$

where Q_p is the mapped score and $\beta_1, \cdots, \beta_5$ are the regression model parameters.

Tab. 2 lists the Pearson linear correlation coefficient values achieved by SIQE. The proposed metric achieves high correlation with all the reference image and video quality metrics. It has average PLCC of 0.9024 with VQM. It exhibits PLCC larger than 0.87 with all quality metrics except PSNR (that in turn is known to correlate poorly with the actual image quality). The prediction error in terms of root mean square error (RMSE) reported in Tab. 3 also reflects the accurate and reliable of performance of the proposed quality metric.

4. CONCLUSIONS

In this paper a quality assessment algorithm is presented to estimate the quality of DIBR synthesized images in the absence of the corresponding references. The algorithm uses the original uncompressed input views to estimate the statistical characteristics of the cyclopean image by using the divisive normalization transform, which are compared to those of the synthesized image to estimate the compression and DIBR warping artifacts in the novel view. The metric is tested in on standard MVD sequences and compared to 7 widely used full reference image an video quality metrics to evaluate its performance. The evaluation results show the effectiveness of the proposed quality metric.

5. REFERENCES

[1] M. Tanimoto, "FTV: Free-viewpoint Television," *Signal Process.-Image Commun.*, vol. 27, no. 6, pp. 555 – 570, 2012.

[2] M.P. Tehrani et al., "Proposal to consider a new work item and its use case - rei : An ultra-multiview 3D display," *ISO/IEC JTC1/SC29/WG11*, July-Aug 2013.

[3] Y. Takaki, "Development of super multi-view displays," document JTC1/SC29/WG11/M32164, Jan 2014.

[4] A. Smolic et al., "Multi-view video plus depth (MVD) format for advanced 3D video systems," doc. ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q6, 2007.

[5] "Call for proposals on 3D video coding technology," Mar 2011, documnet ISO/IEC JTC1/SC29/WG11.

[6] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," in *Proc. SPIE*, 2004, vol. 5291, pp. 93–104.

[7] M. Domanski et al., "High efficiency 3D video coding using new tools based on view synthesis," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3517–3527, 2013.

[8] M.S. Farid, M. Lucenteforte, and M. Grangetto, "Panorama view with spatiotemporal occlusion compensation for 3D video coding," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 205–219, Jan 2015.

[9] T. Maugey, A. Ortega, and P. Frossard, "Graph-based representation for multiview image geometry," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1573–1586, May 2015.

[10] G.J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, 2012.

[11] K. Muller et al., "3D High-Efficiency Video Coding for Multi-View Video and Depth Data," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3366–3378, Sept 2013.

[12] A. Boev et al., "Classification and simulation of stereoscopic artifacts in mobile 3dtv content," in *Proc. SPIE*, 2009, vol. 7237, pp. 72371F–72371F–12.

[13] L. Fang et al., "An analytical model for synthesis distortion estimation in 3D video," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 185–199, Jan 2014.

[14] P. Merkle et al., "The effects of multiview depth video compression on multiview rendering," *Signal Process.-Image Commun.*, vol. 24, no. 12, pp. 73 – 88, 2009.

[15] M.S. Farid, M. Lucenteforte, and M. Grangetto, "Edge enhancement of depth based rendered images," in *IEEE Int. Conf. Image Process. (ICIP)*, Oct 2014, pp. 5452–5456.

[16] M.S. Farid, M. Lucenteforte, and M. Grangetto, "Edges shape enforcement for visual enhancement of depth image based rendering," in *IEEE 15th Int. Workshop Multimedia Signal Process. (MMSP)*, 2013, pp. 406–411.

[17] E. Bosc et al., "Towards a new quality metric for 3-D synthesized view assessment," *IEEE J. Sel. Top. Signal Process.*, vol. 5, no. 7, pp. 1332–1343, Nov 2011.

[18] J. You et al., "Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis," in *Proc. VPQM*, 2010.

[19] A.K. Moorthy and A.C. Bovik, "A survey on 3D quality of experience and 3D quality assessment," in *Proc. SPIE*, 2013, vol. 8651, pp. 86510M–86510M–11.

[20] E. Bosc, P. Le Callet, L. Morin, and M. Pressigout, "An edge-based structural distortion indicator for the quality assessment of 3D synthesized views," in *Proc. Picture Coding Symp. (PCS)*, May 2012, pp. 249–252.

[21] H. Shao, X. Cao, and G. Er, "Objective quality assessment of depth image based rendering in 3dtv system," in *Proc. IEEE 3DTV-CON*, May 2009, pp. 1–4.

[22] F. Battisti et al., "Objective image quality assessment of 3d synthesized views," *Signal Process.-Image Commun.*, vol. 30, pp. 78 – 88, 2015.

[23] P.-H.i Conze, P. Robert, and L. Morin, "Objective view synthesis quality assessment," in *Proc. SPIE*, 2012, vol. 8288, pp. 82881M–82881M–14.

[24] B. Julesz, "Cyclopean perception and neurophysiology," *Invest. Ophthalmol. Vis. Sci.*, vol. 11, no. 6, pp. 540–548, 1972.

[25] P.C. Teo and D.J. Heeger, "Perceptual image distortion," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Nov 1994, vol. 2, pp. 982–986.

[26] R.P.N. Rao, B.A Olshausen, and M.S. Lewicki, *Probabilistic models of the brain: Perception and neural function*, Mit Press, 2002.

[27] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE J. Sel. Top. Signal Process.*, vol. 3, no. 2, pp. 202–211, April 2009.

[28] A. Mittal, A.K. Moorthy, and A.C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec 2012.

[29] D.L. Ruderman, "The statistics of natural images," *Netw.-Comput. Neural Syst.*, vol. 5, no. 4, pp. 517–548, 1994.

[30] E. Hellinger, "Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen.," *J. Reine Angew. Math.*, vol. 136, pp. 210–271, 1909.

[31] M.S. Farid, M. Lucenteforte, and M. Grangetto, "Depth image based rendering with inverse mapping," in *Proc. IEEE 15th Int. Workshop Multimedia Signal Process. (MMSP)*, Sept 2013, pp. 135–140.

[32] M.H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sept 2004.

[33] Z. Wang et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, April 2004.

[34] L. Zhang, D. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug 2011.

[35] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*, 2003, vol. 2, pp. 1398–1402.

[36] Zhou Wang and Qiang Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[37] Z. Wang and A.C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, March 2002.

[38] Video Quality Expert Group, "Final report from the video quality experts group on the validation of objective models of video quality assessment," 2003.

[39] H.R. Sheikh, M.F. Sabir, and A.C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov 2006.