No-reference Stereoscopic Image Quality Assessment with Saliency-guided Binocular Feature Consolidation

Xiaogang Xu, Yang Zhao, and Yong Ding, *Member, IEEE*

***Abstract*—Stereoscopic image quality assessment, especially the method that does not require the presence of a pristine image for reference, has become a fundamental yet challenging issue. In this paper, we propose a no-reference framework based on multi-scale feature extraction and saliency-guided feature consolidation. The underlying multi-scale features are extracted from both left and right views including two aspects: a) global features based on natural scene statistics derived from the distribution of local mean subtracted contrast normalized coefficients and b) local features including spatial and spectral entropy features. The feature consolidation is implemented by a weighted process for each of features from right and left views where the weights are obtained by visual saliency detection. In addition, the kurtosis and skew of disparity distribution are measured to capture the statistical changes of disparity information. Finally, we apply support vector regression for feature pooling and image quality mapping. Experimental results on public databases confirm the promising performance of the proposed method that correlates highly with human judgments on image quality.**

***Index Terms*—Stereoscopic image quality assessment, saliency detection, multi-scale features, feature consolidation**

I. Introduction

PERCEPTUAL quality assessment of stereoscopic images (3D-IQA) plays an essential and fundamental role in the design, performance monitoring and optimization of 3D image and video processing systems, which falls into two categories: subjective judgment and objective evaluation. While the subjective methods are usually cumbersome, time-consuming and expensive for real-time applications, it is more valuable to evaluate the quality of 3D content objectively that is consistent with the human subjective perception [1]. Typically, objective 3D-IQA can be classified into with-reference (WR) or no-reference (NR) methods according to whether original information is demanded for reference. In most practical cases, the original information regarding reference is unavailable. NR methods are thus highly desirable and have potentially much broader applicability.

Unfortunately, investigation dedicated to NR 3D-IQA is still immature. Earliest algorithms are simply applying off-the-shelf 2D methods on the left and right stereo views respectively [2]. Owing to neglecting the complex binocular effects which have an important impact upon image quality sensation, these approaches give out unsatisfactory results. In general, the success of 3D-IQA depends greatly on quality-aware feature description that reflects the perceived quality intrinsically. Current methods attempt to make use of the binocular effects as well as other important properties of HVS and natural scene statistics (NSS) for feature extraction. Zhou et al. constructed a no-reference method that utilized a binocular vision mechanism by training a pristine multivariate Gaussian model to extract the quality-predictive features [1]. Shao et al. proposed a metric by learning the characteristics of binocular receptive fields' features [3]. More recently, with the development of machine learning tools, learning a regression model for mapping image features into quality scores becomes popular [4]. For example, in [5], with NSS-based feature extraction from the cyclopean view and disparity map, support vector regression (SVR) was used to learn a regression function for quality prediction. Certainly, with the stereo image pair and disparity map to construct an intermediate image for further feature extraction and quality assessment is a promising solution to effectively utilize the depth information [6], but the construction process including binocular fusion and disparity computation is rather computationally complex. Furthermore, there are considerably fewer approaches that correlate highly with human judgments of quality and work well across multiple classes of distortions.

This study puts forward an effective method for NR 3D-IQA, in which multi-scale features are extracted globally and locally to ensure a comprehensive representation for the stereoscopic image, and these features are further weighted for fusion using a saliency model. The main contributions are as follows: First, a new framework based on feature consolidation is proposed which improves the computational efficiency greatly in contrast to the time-consuming process of intermediate image construction in traditional methods [6]. Second, the global and local features not only reflect image quality complementarily, but also take the multi-scale property of HVS into account. Third, the saliency-guided feature consolidation simulates the binocular rivalry effect effectively by defining weights of features based on saliency map detected from each view. Besides, for each stereo pair, the statistical changes of disparity information are captured as additional features. Experiments validate the attractive performance of the proposed method that it is highly consistent with the subjective evaluation and robust across databases and distortions.

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456".

X. Xu, Y. Zhao, Y. Ding are with the College of Information Science and Electronic Engineering, Zhejiang University, Zhejiang 310027, China (e-mail: xiaogangxu@zju.edu.cn; zhaoyang@vlsi.zju.edu.cn, dingy@vlsi.zju.edu.cn)

II. THE PROPOSED METHOD

Inspired by the success of NSS-based feature extraction in 2D-IQA which is developed on the hypothesis that the presence of distortions inevitably alters the natural statistical properties of images, thereby rendering them (and consequently their statistics) unnatural, we present a powerful model based on global and local feature descriptions to capture the statistical behavior of stereoscopic images complementarily.

A. Global feature extraction

Global feature extraction is executed in the spatial domain, in which we compute locally normalized luminance coefficients through local mean subtraction and subsequent divisive normalization. It is revealed that applying the local non-linear operation into log-contrast luminance can remove local mean displacements from zero log-contrast and normalizing the local variance of the log-contrast will exert a decorrelating effect [8]. Such an operation be applied to a given image is expressed as

$$\begin{aligned} I_T(i,j) &= \frac{I(i,j)-\mu(i,j)}{\sigma(i,j)+C} \\ \mu(i,j) &= \sum_{k=-K}^{K}\sum_{l=-L}^{L} w_{k,l} I_{k,l}(i,j) \\ \sigma(i,j) &= \sqrt{\sum_{k=-K}^{K}\sum_{l=-L}^{L} w_{k,l}\left(I_{k,l}(i,j)-\mu(i,j)\right)^2} \end{aligned} \tag{1}$$

where $i \in 1,2,\ldots,M$, $j \in 1,2,\ldots,N$ are spatial indices respectively, K and L are the image's height and width. $C = 1$ is a constant which prevents the instability in case of denominator tending to zero. $w_{k,l}$ is a 2D Gaussian weighting function which is sampled out to 3 standard deviations and then rescaled to an unit volume. The transformed luminance $I_T(i,j)$ is the mean subtracted contrast normalized (MSCN) coefficients. Experiments in [7] demonstrated that the statistical properties of these coefficients would be changed by the presence of distortions. So, the perceptual quality can be predicted by quantifying such changes. Moreover, it has been observed that the histogram of MSCN coefficients of a natural image exhibits a Gaussian like appearance [8]. In this way, we adopt a generalized Gaussian distribution (GGD) to obtain the spectrum of the image statistical properties. By zero mean distribution being set for GGD, it is given as following

$$\begin{aligned} f(x;\alpha,\sigma^2) &= \frac{\alpha}{2\beta\,\Gamma(1/\alpha)}\exp\left(-\left(\frac{|x|}{\beta}\right)^{\alpha}\right) \\ \beta &= \sigma\sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}} \end{aligned} \tag{2}$$

where $f(x;\alpha,\sigma^2)$ is the normalized number of coefficients, x is the MSCN coefficient, $\Gamma(\cdot)$ is the Gamma function [9]. In the expression, there are two key parameters including the variance σ^2 and α controlling the distribution shape. Consequently, these two parameters are selected to be the global features,

$$f_g = \{\alpha, \sigma^2\} \tag{3}$$

The detail computation for these global features can refer to [9]. Besides, since multi-scale representation can provide more information of an image, in this Letter, the global features are extracted from two scales, yielding 2×2=4 features. Images with different scales are obtained by pyramid decomposition.

B. Local feature extraction

It is claimed that distortions may affect the local entropy of an image which can indicate the amount of information contained within the image. Different types and degrees of distortions exert distinctly different influences on the spatial entropy values as well as spectral entropy values [10]. The spatial entropy is a function of the probability distribution of local pixel values revealing the statistical characteristics in pixel level. It can be expressed as following

$$E_s = -\sum_{x} p(x)\log_2 p(x) \tag{4}$$

where x is the pixel value in a block and $p(x)$ is the probability density correspondingly.

Likewise, in implementation, the spectral entropy feature which is even more indicative of distortion types than spatial entropy, can be regarded as a probability distribution function for local DCT coefficient values,

$$E_f = -\sum_{i}\sum_{j} c(i,j)\log_2 c(i,j) \tag{5}$$

where $c(i,j)$ are the normalized DCT coefficients in a block. Also, the spatial and spectral entropy are extracted from both right and left views of a stereoscopic image pair in different scales, respectively.

Furthermore, to reduce the image content dependency that may pose a challenge in quality assessment, statistical measures such as mean and skew for both spatial entropy and spectral entropy are computed and used as quality-aware features. There we use 60% of the center elements to construct the local features

$$f_l = \{mean(E_s),\ skew(E_s),\ mean(E_f),\ skew(E_f)\} \tag{6}$$

As the spatial and spectral entropy is extracted from three scales, 4×3=12 local features are yielded. Combined with the global features, there are all 16 features extracted for the presentation of left and right views, respectively.

C. Saliency-guided feature consolidation

Since saliency map can reflect the energy distribution in images, it highlights the regions which attract human attention. It has been demonstrated that the current soundness of visual saliency modelling is sufficient for IQA to yield a statistically meaningful gain in their performance [11]. Recent years, many saliency models aiming at detecting areas that human visual systems are concerned about, have been developed. However, the problem of computation complexity is still a challenge. In this Letter, we adopt the simple method proposed in [12] for the visual saliency detection. The residual of an image is extracted firstly in spectral domain, then the corresponding saliency map in spatial domain is captured. The procedure of computing a saliency map for a given image can be described as following,

$$
\begin{aligned}
A(f) &= \Re(F[I(x)]) \\
P(f) &= \wp(F[I(x)]) \\
L(f) &= \log(A(f)) \\
R(f) &= L(f) - h_n(f) \cdot L(f) \\
S(x) &= g(x) \cdot F^{-1}[\exp(R(f) + P(f))]^2
\end{aligned} \tag{7}
$$

where $I(x)$ is the given image, F and F^{-1} denote the Fourier Transform and Inverse Fourier Transform, respectively. $A(f)$ is amplitude information and $\Re$ is an operator to compute the amplitude. $P(f)$ denotes the phase spectrum of the image and $\wp$ is an operator to extract phase information. $L(f)$ is log spectrum of the image. $h_n(f)$ is a local average filter used to approximate the shape of $A(f)$. $R(f)$ is the spectral residual of the image. And $g(x)$ is a Gaussian filter used to smooth the saliency map. $S(x)$ is the captured saliency map.

Based on the fact that binocular strength is a weight sum of monocular stimulus strength [13], we put forward a saliency-guided strategy to consolidate the features extracted from left and right views into a set of features. More specifically, we make use of saliency map to derive the weights for each view's features in order to simulate binocular rivalry. In our framework, feature extraction and saliency detection are both implemented in multiple scales, thus the feature consolidation is multi-scale, too. For example, at the n-th scale, we use f_g^n, f_l^n and f^n to denote the combined global, local and the final features, respectively. Similarly, $f_{g\_left}^n$ and $f_{g\_right}^n$ are the global features of the left and right views extracted by Eq. (3). $f_{l\_left}^n$ and $f_{l\_right}^n$ are local features of left and right views extracted by Eq. (6). Correspondingly, at the n-th scale, the totality of elements in the left and right saliency maps are notated by S_{left}^n and S_{right}^n. The procedure of saliency-guided consolidation is,

$$
\begin{aligned}
f_g^n &= \frac{S_{left}^n}{S_{left}^n + S_{right}^n} f_{g\_left}^n + \frac{S_{right}^n}{S_{left}^n + S_{right}^n} f_{g\_right}^n \\
f_l^n &= \frac{S_{left}^n}{S_{left}^n + S_{right}^n} f_{l\_left}^n + \frac{S_{right}^n}{S_{left}^n + S_{right}^n} f_{l\_right}^n
\end{aligned} \tag{8}
$$

$$
f^n = \{ f_g^n, f_l^n \} \tag{9}
$$

D. Disparity information

It is believed that human stereoscopic perception is strongly dependent on the left and right views of a stereo pair and disparity information between them as well. The existence of distortions will also alter the statistical characteristics of the disparity map. Since there is no thoroughly study on disparity computing for 3D-IQA, a structural-similarity-based scheme utilizing a minimum disparity criterion was proposed by Chen *et al*. [6]. In this Letter, the disparity map is generated mainly based on this method.

For each stereo pair, we attempt to capture the statistical changes of disparity information which are related to the perceived stereoscopic quality using simple statistical measures such as the kurtosis and skew of the disparity distribution.

$$
\begin{aligned}
k &= E[(D - E(D))^4] / E[(D - E(D))^2] \\
s &= E[(D - E(D))^3] / (E[(D - E(D))^2])^{3/2}
\end{aligned} \tag{10}
$$

where k and s are the kurtosis and skew of the disparity map, respectively. Thus, for each pair of stereo pair, we extract two features from the corresponding disparity map.

E. Image quality mapping by SVR

For 3D-IQA, the goal of objective quality mapping is to construct a function to give out an objective score of a stereo image pair. That is, the objective score is calculated by using the extracted features and a mapping function. In order to build the mapping relationship between the consolidated features of a stereo image pair and the corresponding subjective scores, in this Letter, we employ the ε-SVR with a radial basis function kernel for regression model learning which has been proven to be an effective machine learning method for high-dimensional data pooling [14]. For the quality evaluation of a stereoscopic image pair, by feeding the consolidated features into the trained SVR model, the overall quality score will be given out. To be more specific, the overall quality for a stereo image pair is predicted by utilizing the regression model $M_{\varepsilon\text{-SVR}}$ constructed in advance and the extracted features in Eq. (9) and Eq. (10). That is, the final quality is predicted as $Q = M_{\varepsilon\text{-SVR}}(k, s, f^n | n = 0, \ldots, N)$.

In the implementation of the training-testing procedure, we adopt five-fold cross-validation where databases are randomly partitioned into two non-overlapping groups: 80% for training and 20% for testing. To eliminate the bias, such a training-testing procedure is run 1000 times, and the median results of the trials are selected for the final validation.

III. RESULTS AND DISCUSSION

For performance validation, two publicly available databases, LIVE Phase I and Phase II, where distorted stereoscopic pairs cover five distortion types including JPEG and JPEG2K compression, additive white Gaussian noise (WN), Gaussian blur (Blur) and Rayleigh fast fading channel distortion (FF). To measure the performance of 3D-IQA methods quantitively, two commonly-used indicators are used: Spearman rank order correlation coefficient (SROCC) which measures the prediction convergence and monotonicity and Person linear correlation coefficient (PLCC) which provides an estimation of the prediction accuracy. The larger values of PLCC and SROCC mean the better performance.

Table I shows the comparison on LIVE Phase I in terms of PLCC and SROCC where the best results are highlighted in boldface. Although Phase I is a valuable database, it only has symmetrically distorted images. In real applications, both views of a stereoscopic image pair may be distorted symmetrically or asymmetrically. Therefore, the performance comparison on Phase II which is listed in Table II, is more constructive.

Obviously, from Table I and Table II, we can find that: (1) Early methods in [14] and [16] deliver poor performance because they are both the extension from 2D-IQA, indicating that directly applying 2D methods into stereoscopic image

quality assessment is unreasonable. (2) NSS-based approaches, such as the methods in [2], [3], [4] and the proposed method, achieve comparative better performance. But it should be noted that some NSS-based metrics do not work well with some special types of distortion. For instance, the method in [1] is not very prominent for JP2K and JPEG distortions. (3) It is instructive to observe that the approach in [15] and our method provide preferable performance resulting from utilizing saliency map detection. (4) Some methods perform well on Phase I database while deliver poor performance on Phase II. (5) Taking binocular features into account is a good ideal to improve the overall performance on the whole database, for example, the metrics in [1] and [17] and ours. In contrast, by taking the advantages of NSS and saliency based binocular feature combination, the proposed method outperforms other metrics in terms of not only the high consistency with subjective evaluation, but also the robustness across different symmetrical or asymmetrical distortions.

TABLE I
The PLCC and SROCC Performance Comparison On LIVE PHASE I

Method	PLCC						
	JP2K	JPEG	WN	Gblur	FF	**σ**	ALL
Zhou[1]	0.848	0.626	0.925	0.899	0.707	0.017	0.887
Shao[2]	0.901	0.456	0.919	0.950	—	0.055	0.878
Chen[4]	0.907	0.695	0.917	0.917	0.735	0.012	0.895
Akhter[14]	0.905	0.729	0.904	0.617	0.503	0.031	0.626
Ryu[15]	0.860	0.630	0.940	0.960	0.780	0.018	0.800
Zhou[17]	—	—	—	—	—	—	0.928
Proposed	**0.951**	**0.738**	**0.948**	**0.966**	**0.848**	**0.009**	**0.949**

Method	SROCC						
	JP2K	JPEG	WN	Gblur	FF	**σ**	ALL
Zhou[1]	0.837	0.638	**0.931**	0.833	0.649	0.016	0.892
Shao[2]	0.870	0.429	0.930	**0.914**	—	0.057	0.866
Chen[4]	0.863	0.617	0.919	0.878	0.652	0.020	0.891
Akhter[14]	0.866	0.675	0.914	0.555	0.640	0.023	0.383
Ryu[15]	—	—	—	—	—	—	0.860
Zhou[17]	0.824	0.614	0.915	0.916	—	0.020	0.887
Proposed	**0.903**	**0.678**	0.905	0.907	**0.800**	**0.010**	**0.934**

TABLE II
THE PLCC AND SROCC PERFORMANCE COMPARISON ON LIVE PHASE II

Method	PLCC						
	JP2K	JPEG	WN	Gblur	FF	**σ**	ALL
Zhou[1]	0.634	0.647	0.904	0.967	0.851	0.023	0.818
Shao[2]	0.803	0.710	0.887	**0.983**	—	0.014	0.820
Zhang[3]	0.782	0.583	0.796	0.924	—	0.020	0.763
Akhter[14]	0.766	0.786	0.722	0.795	0.674	0.003	0.568
You[16]	0.732	0.674	0.546	0.976	0.839	0.027	0.721
Zhou[17]	—	—	—	—	—	—	0.861
Proposed	**0.900**	**0.815**	**0.972**	**0.982**	**0.891**	**0.005**	**0.926**

Method	SROCC						
	JP2K	JPEG	WN	Gblur	FF	**σ**	ALL
Zhou[1]	0.553	0.593	0.593	0.869	0.828	0.022	0.825
Shao[2]	0.790	0.730	0.882	**0.942**	—	0.009	0.789
Zhang[3]	0.793	0.581	0.780	0.865	—	0.015	0.708
Akhter[14]	0.724	0.649	0.714	0.682	0.559	0.004	0.543
You[16]	0.731	0.523	0.482	0.923	0.839	0.004	0.721
Zhou[17]	0.717	0.593	0.891	0.903	—	0.022	0.823
Proposed	**0.861**	**0.771**	**0.936**	0.922	**0.858**	**0.004**	**0.910**

IV. CONCLUSIONS AND FUTURE WORK

Based on multi-scale quality-aware features extracted locally and globally, binocular feature consolidation by a saliency map guided weighting strategy, and SVR-based quality mapping, we propose a novel no-reference framework for stereoscopic image quality assessment. Experimental results show that the proposed method is a promising solution that is consistent with the subjective judgement and stable across different distortions.

REFERENCES

[1] W. Zhou, L. Yu, W. Qiu, Z. Wang and M. Wu. "Utilizing binocular vision to facilitate completely blind 3D image quality measurement," Signal Processing, vol.129, pp.130-136, Dec.2016

[2] J. You, L. Xing, A. Perkis, and X. Wang, "Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis," in Proc. Int. Workshop Video Process. Qual. Metrics Consum. Electron., Scottsdale, AZ, USA, pp. 4033-4036, Sep. 2010

[3] F. Shao, "Learning receptive fields and quality lookups for blind quality assessment of stereoscopic images," IEEE Trans. Cybernetics, vol.46, no.3, pp.730-743, Mar. 2016

[4] W Zhang, "Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network," Pattern Recognition, vol.59, no.C, pp.176-187, Jan.2016

[5] M.-J. Chen, L. K. Cormack, and A. C. Bovik, "No-reference quality assessment of natural stereopairs," IEEE Trans. Image Process., vol.22, no. 9, pp. 3379–3391, 2013

[6] M.-J. Chen, C.-C. Su, D.-K. Kwon, L. K. Cormack, and A. C.Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," Signal Process.: Image Commun., vol. 28, no. 9, pp. 1143–1155, Oct. 2013.

[7] 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.

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

[9] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video," IEEE Trans. Circuits Syst. Video Technol., vol. 5, no. 1, pp. 52–56, Feb. 1995.

[10] L. Liu, B. and H. Huang, "No-reference image quality assessment based on spatial and spectral entropies," Signal Process.: Image Commun., vol.29, no.8, pp.856–863, 2014,

[11] L. Zhang, Y. Shen, H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," IEEE Trans. Image Process., vol.23, no.10, pp.4270-4281, Oct.2014

[12] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, Minneapolis, Minnesota, June.2007, pp.1-8

[13] W. J. Levelt, On Binocular Rivalry. The Hague, The Netherlands: Mouton, 1968

[14] R. Akhter, J. Baltes, Z. M. Parvez Sazzad, and Y. Horita, "No-reference stereoscopic image quality assessment," in Proc. SPIE, vol.7524, no.2, pp.75240T-75240T-12, Feb. 2010

[15] S. Ryu and K. Sohn, "No-reference quality assessment for stereoscopic images based on binocular quality perception," IEEE Trans. Circuits Syst. Video Technol., vol. 24, no. 4, pp. 591–602, Apr. 2014.

[16] J. You, L. Xing, A. Perkis, and X. Wang, "Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis," in Proc. Int. Workshop Video Process. Qual. Metrics Consum. Electron., Scottsdale, AZ, USA, pp. 4033-4036, Sep. 2010

[17] W. Zhou and L. Yu, "Binocular Responses for No-Reference 3D Image Quality Assessment," IEEE Trans. Multimedia, vol. 18, no. 6, pp. 1077-1084, June.2016