

# Stereoscopic Image Quality Assessment Based on Binocular Adding and Subtracting

Jiachen Yang<sup>1</sup>, Bin Jiang<sup>1(✉)</sup>, Chunqi Ji<sup>1</sup>, Yinghao Zhu<sup>1</sup>, and Wen Lu<sup>2</sup>

<sup>1</sup> School of Electrical and Information Engineering,  
Tianjin University, Tianjin 300072, China

{yangjiachen, jiangbin, jcq., zhuyinghao}@tju.edu.cn

<sup>2</sup> School of Electronic Engineering, Xidian University, Xi'an 710071, China  
luwen@xidian.edu.cn

**Abstract.** There has been a great concern on blind image quality assessment in the field of 2D images, however, stereoscopic image quality assessment (SIQA) is still a challenging task. In this paper, we propose an efficient blind image quality assessment model for stereoscopic images according to binocular adding and subtracting channels. Different from other SIQA methods which focus on complex binocular visual properties, we simply use the visual information from adding and subtracting to describe binocularity (also known as ocular dominance) which is closely related to distortion types. To better evaluate the contribution of each channel in SIQA, a dynamic weighting is introduced according to local energy. Meanwhile, distortion-aware features based on wavelet transform are utilized to describe visual degradation. Experimental results on 3D image databases demonstrate the potential of the proposed framework in predicting stereoscopic image quality.

**Keywords:** Ocular dominance · Binocularity  
Blind image quality assessment · Adding and subtracting

## 1 Introduction

In recent years, there has been great progress in Image Quality Assessment (IQA). The emergence of various IQA databases and the proposal of IQA theory greatly enrich the way to evaluate image quality [1]. At the beginning, 2D-IQA metrics mainly focus on the difference between reference and distorted images (known as Full-Reference (FR) methods). For example, the well-known Structural Similarity Index Measurement (SSIM) measures image quality from the perspective of image formation, and Visual Information Fidelity (VIF) explores the consistence between image information and distortion. The rising of blind (No-Reference, NR) IQA methods has a huge influence on IQA, such as Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE) [2], Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [3], Blind Image Integrity Notator using DCT Statistics-II (BLIINDS-II) [4]. They utilize

distortion-aware/sensitive features to evaluate image quality, and therefore the availability of visual features is of great importance. Meanwhile, there are also deep learning based frameworks [5]. These newly proposed IQA methods achieve great success in predicting image quality.

However, Bosc *et al.* have proved that 2D-IQA metrics are not applicable to SIQA because of the weak correlation with binocular visual properties [6]. To solve this problem, scholars and experts have been long focusing on binocular visual characteristics. Chen *et al.* designed an intermediate image closely resembles the cyclopean image [8]. Shao *et al.* classified the stereoscopic images into non-corresponding, binocular fusion and binocular suppression regions [9]. Ryu *et al.* models the binocular quality perception in context of blurriness and blockiness [10]. Yang *et al.* proposed a quality index by evaluating binocular subtracting and adding.

Recently, it has been found that the Human Visual System has separately adaptable channels for adding and subtracting the visual signals from two views. Compared with previous research on binocular interactions, we simply focus on the adding and subtracting channels to demonstrate binocular visual properties. In addition, ocular dominance which produces binocularity is considered to characterize the receptive field properties from monocular response to binocular response, and it is closely related to different distortions. Therefore, we try to use the information from subtracting and adding to measure the ocular dominance. With this method, we also greatly reduce the amount of computation in modeling complicated binocular visual properties. However, to what extent each channel contributes to binocular visual perception is still undiscovered. To solve this problem, we take the local energy response as a weighting index to balance their performance in stereo perception, since energy maps provide useful local binocular rivalry information, which may be combined with the qualities of single-view images to predict 3D image quality [12].

Another key point would be how to extract distortion-aware features from the adding and subtracting channels to describe visual degradation. There have been a great number of effective ways to extract Natural Scene Statistics (NSS) features. For example, General Gaussian Distribution (GGD) model has been used to fit the non-Gaussian distribution from frequency domain, and statistical properties of MSCN coefficients are also explored in spatial domain. However, sometimes these properties would change with different visual content. Taking this shortcoming of these methods into account, He *et al.* studied the exponential attenuation characteristics of the magnitude, variance and entropy in different wavelet subbands. In this paper, those features from wavelet transform are utilized to represent the properties of natural scenes since they are less sensitive to the image content. Meanwhile, the feature extraction procedure is rather computationally efficient.

Our framework is based on the adding and subtracting theory which describes ocular dominance. As a result, NSS features are extracted from the adding and subtracting channels. Compared with the time-consuming methods, the proposed framework achieves a good balance between efficiency and prediction accuracy.

The rest of the paper is organized as follows. Section 2 presents the framework of the proposed metric. The experimental results and analysis are given in Sect. 3, and finally conclusions are drawn in Sect. 4.

## 2 Background and Motivation

Before the proposed method, we refer some basic theory about stereo image quality, which will give us some motivation. As for stereo image quality, it refers to the machine method to keep consistent with subjective feelings of humans. In the processing of acquisition and display, the stereo images will be affected by the existing noise interference, which will make the images have a certain difference with the original images. In this way, it will give a bad visual perception.

For human eyes, it can make use of emitted light to see the objects, and the final images will be sent to the visual center through internal light system. In this way, it can produce the visual perception. Based on the previous researches, there are three important visual neural pathways for people brains. Specially, they are ventral pathway, dorsal pathway and visual pathway. For ventral pathway, it contains the primary visual cortex named  $V1$ , secondary visual cortex named  $V2$  and  $V3$ . And it also contains the ventral extrastriate cortex named  $V4$ .

When people make use of consciousness and perception, all of the above parts will combine to determine the movement and location of the object. Judging from above, human visual system(HVS) is very complex and it contains physiological structure. In addition, some high level visual cortex interaction should be used for the visual pathway in order to finish the circulation mechanism and the feedback types.

Through the production of random point of view, the absence of any cue will happen in the human eyes, which will affect the perception of the depth. In stereo images perception, the human eye's binocular disparity and convergence can be made use of. And then the binocular disparity can be regarded as the most important part for the three-dimensional technology.

In general, the horizontal distance of the two human eyes is about 6 cm, which can help people get the certain difference in scene. In detail, the projected object in left and right eyes are different respectively. The difference can be named binocular disparity. As for disparity, it can be divided into two parts: vertical and horizontal parallax. specially, the horizontal parallax is the most important for the depth perception. And vertical parallax can only determine the perception comfort. In other word, if the vertical parallax is bad, the human will feel uncomfortable perception.

In this connection, a lot of researchers to study the visual characteristics of the human eye binocular fusion, and establish a corresponding stereo vision model, the corresponding mathematical model is as follows:

Model of Eye-Weighting (EW): Engel *et.al* proposed a binocular weighted (EW) model. The model also contains a weighting factor for the binocular fusion process, but unlike the simple model, the model coefficients can be obtained by

integrating the square root of each eye signal autocorrelation function. And it can be described as

$$C = ((W_L \cdot E_L)^2 + (W_R \cdot E_R)^2)^{\frac{1}{2}} \tag{1}$$

Model of Vector Summation (VS): Therefore, Curtis *et al.* Found that the binocular fusion graph is the sum of the two normalized orthogonal vectors, and the vector sum (VS) model is proposed.

$$C = \sqrt{I_L^2 + I_R^2 + I_L \cdot I_R} \tag{2}$$

Among them, IL and IR represent the left and right view information.

Model of Gain Control (GC): Ding et al. on the basis of previous research proposed the corresponding gain control model, they pointed out that people each eye visual gain information not only from in its input signal, but also from in the other eye gain control signal energy.

$$C = \frac{I_L}{1 + I_R} + \frac{I_R}{1 + I_L} + 0.1 \cdot I_L \cdot I_R \tag{3}$$

Model of Neural Network (NN): A neural network (NN) model is proposed by Cogan et al.

$$C = \frac{E_L}{1 + E_L + E_R} \cdot I_L + \frac{E_R}{1 + E_L + E_R} \cdot I_R \tag{4}$$

He proposed that the *log - Gabor* model can well reflect the simple visual cell function of the human eye. Using *log - Gabor* filter to volume response calculation of gain control system model of weighted value, according to the US in the literature, we define  $[\eta_{s,o}, \zeta_{s,o}]$  as a different direction and size of the filter response, this chapter of the *log - Gabor* filter  $G_{s,o}(\omega, \theta)$  is defined as shown.

$$G_{s,o}(\omega, \theta) = \exp\left[-\frac{(\log(\omega/\omega_s))^2}{2\sigma_s^2}\right] \cdot \exp\left[-\frac{\theta - \theta_0}{2\sigma_o^2}\right] \tag{5}$$

Among them, *s* and *o* is respectively spatial scale and orientation information,  $\theta$  said direction angle information,  $\omega_s$  and  $\omega_o$  is used to determine the filter energy,  $\omega$  and  $\omega_s$  represent the normalized radial filter frequency and the center frequency. According to this, we can calculate the signal *X* position in the size of *s*, *o* direction of the local energy information for the:

$$E_o(X) = \sqrt{F_o(X)^2 + H_o(X)^2} \tag{6}$$

Among them,

$$F_o(X) = \sum_s \eta_{s,o}(X) \tag{7}$$

$$H_o(X) = \sum_s \zeta_{s,o}(X) \tag{8}$$

And the energy of the filter is:

$$E(X) = \sum E_o(X) \tag{9}$$

At last ,we can see the energy value is the weight value of the left eye and right eye.

Motivated by the above model, we find that binocular adding and subtracting is another obvious model which can affect the stereo perception. However, there is few research results that pay attention on this issue.

### 3 The Proposed Framework

The proposed framework contains two parts: binocular adding and subtracting and NSS feature extraction. At last, SVM is used to connect the NSS features with image quality. Detailed information about the framework is shown in Fig. 1.

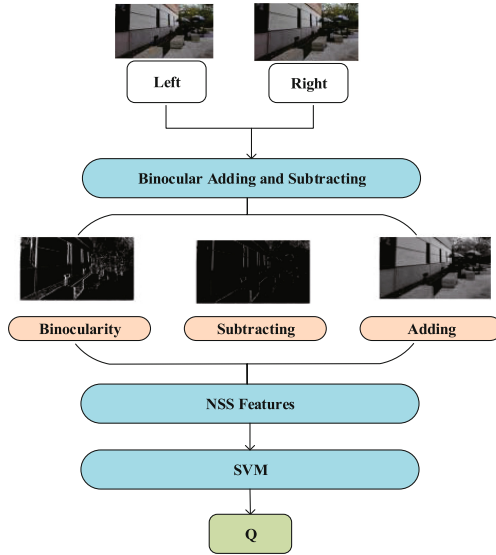
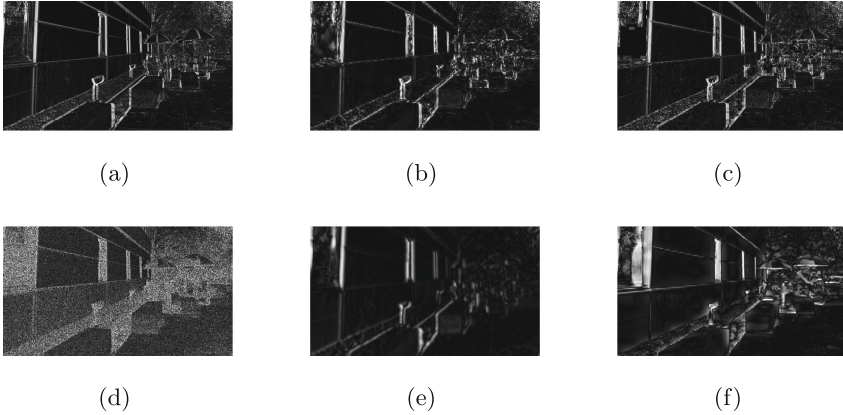


Fig. 1. Framework of the proposed image quality assessment metric

#### 3.1 Binocular Adding and Subtracting

HVS has separately adaptable channels for adding and subtracting the neural signals from the two views. Encoding the adding channel  $A$  and subtracting channel  $S$  between two stereo-halves can be used for stereopsis.

$$\begin{aligned} A &= L + R \\ S &= |L - R| \end{aligned} \tag{10}$$



**Fig. 2.** Examples of binocularity. (a) Original image pair. (b) JP2K compressed image pair. (c) JPEG compressed image pair. (d) White noised image pair. (e) Fast faded image pair. (f) Gaussian blur image pair

Examples of binocular adding and subtracting are shown in Fig. 2. It can be seen that subtracting and adding show totally different visual information from each other, the subtracting channel contains the difference between two views because of the viewing angle, which could be considered as an alternative of disparity. Considering that the computation of disparity is extremely time-consuming, we take subtracting channels as another choice to achieve the balance between better efficiency and prediction results. In comparison, the adding channel is more similar to a information-enriched map.

In order to model neural mechanism for binocular processing, the responses from subtracting and adding channels are used to characterize the ocular dominance. To be specific, the binocularity is defined as :

$$b = \frac{|W_{left}| - |W_{right}|}{|W_{left}| + |W_{right}|} \quad (11)$$

where  $W_{left}$  and  $W_{right}$  are the monocular response from each view, and  $b$  represents the degree of binocularity. A large absolute value of  $b$  represents a weak binocular response, and vice versa. To better visualize ocular dominance in terms of distortion, Fig. 2 shows that ocular dominance changes with regard to different distortions, and therefore it could be an index which characterize binocular visual properties.

However, to the best of our knowledge, which channel contributes more to visual perception is still unsolved. Therefore we proposed a weighting scheme to combine the two channels together. Based on the fact that distortion in either view may affect the consistency between the two views and lead to binocular rivalry, a reasonable way to characterize this property is the local energy map since higher-energy regions are more likely to attract visual attention. Therefore, a local energy based weighting scheme is adopted to balance the adding and subtracting channels. Here, the energy for each channel is obtained by summing the

local variances using an  $11 \times 11$  circular-symmetric Gaussian weighting function  $w = \{\omega_i | i = 1, 2, \dots, N\}$ , with standard deviation 1.5 samples, normalized to unit sum ( $\sum_{i=1}^N \omega_i = 1$ ). And then local energy is calculated by

$$e_{in} = \left( \sum_{i=1}^N \omega_i (x_i - \mu_{in})^2 \right)^{\frac{1}{2}} \quad (12)$$

where  $\mu_{in} = \sum_{i=1}^N \omega_i x_i$  is the local mean value. Finally the energy map is computed from 4scales by:

$$e_l = \frac{1}{n_s \times M} \sum_{i=1}^{n_s} \sum_{j=1}^M e_{in,j-l} \quad (13)$$

### 3.2 NSS Feature Extraction

He *et al.* have demonstrated that the tertiary properties of the wavelet coefficients reflect the self-similar property of scenes. In particular, the exponential attenuation characteristics of the magnitude, variance and entropy in different wavelet subbands are less specific in their representation of an image, and thus are utilized to represent the generalized behaviors of natural scenes.

Specifically, the magnitude  $m_k$  is used to encode the generalized spectral behavior, the variance  $v_k$  to describe the fluctuations of energy, and the entropy  $e_k$  to represent the generalized information, following Eqs. 14–16.

$$m_k = \frac{1}{N_k \times M_k} \sum_{j=1} \sum_{i=1} \log 2 |C_k(i, j)| \quad (14)$$

$$v_k = \frac{1}{N_k \times M_k} \sum_{j=1} \sum_{i=1} \log 2 |C_k(i, j) - m_k| \quad (15)$$

$$e_k = \sum_{j=1}^{N_k} \sum_{i=1}^{M_k} p[C_k(i, j)] \ln p[C_k(i, j)] \quad (16)$$

where  $C_k(i, j)$  represents the  $(i, j)$  coefficient of the  $k$ -th subband,  $M_k$  and  $N_k$  are the length and width of the  $k$ -th subband, respectively;  $p[\cdot]$  is the probability density function of the subband.

In our framework, an image is decomposed into 4 scales and 8 wavelet subbands without distinguishing the the low-high and high-low subbands because of their similarity in statistics. The vertical and horizontal subbands with an identical mark in the same scale are combined through averaging after the above process. Finally, there are 24 features extracted from each channel in total

$$f = [m_1, m_2, \dots, m_8, v_1, v_2, \dots, v_8, e_1, e_2, \dots, e_8]^T \quad (17)$$

## 4 Experimental Results and Analyses

### 4.1 Stereo Database

To verify the performance of the proposed method, the LIVE 3D Image Quality Databases (Phase I and Phase II) of the University of Texas at Austin are used [13], which are shown in Fig. 3. Database Phase I database contains 365 stereopairs with symmetric distortion and Database Phase II contains 360 stereopairs with both asymmetric and symmetric distortion, including JPEG compression, JP2K compression, white noise (WN), Gaussian blur (Blur) and fast fading (FF).



Fig. 3. LIVE 3D image quality databases

### 4.2 Performance Measure

For performance evaluation, four commonly used indicators are adopted: Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC), Kendall Rank-order Correlation Coefficient (KROCC), and Root Mean Squared Error (RMSE) between subjective scores and objective scores after nonlinear regression. For nonlinear regression, we use a 4-parameter logistic mapping function:

$$DMOS_P = \frac{\beta_1 - \beta_2}{e^{(Q - \beta_3)/|\beta_4|} + 1} + \beta_2 \quad (18)$$

where  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$  and  $\beta_4$  are the parameters to be fitted. A better match is expected to have higher PLCC, SROCC, KROCC and lower RMSE.

In the prediction phase, the stereopairs in each database were randomly divided into two parts, with 80% for training and 20% for test. To ensure that the proposed framework is robust, 1000 iterations of training are performed by



varying the splitting of data over the training and test sets, and the median value of all iterations is chosen as the final prediction model. All the parameters of our SVM model are the same for different databases.

In order to demonstrate its efficiency, the proposed method is compared with several existing state-of-art IQA metrics, including five 2D metrics (DIIVINE [2], BLIINDS-II [4] and BRISQUE [3]), and four 3D metrics (Lin’s scheme [7], Chen’s scheme [8], Shao’s scheme-A [9] and Shao’s scheme-B [11]). For those 2D-extended BIQA metrics, feature vectors extracted from the left and right views are averaged, and SVM is used to train a regression function. For Lin’s scheme, the FI-PSNR metric is adopted into the comparison. For Chen’s scheme, the adopted 2D metric is MS-SSIM which performs the best. For Shao’s schemes, experimental results in the reference are directly adopted.

### 4.3 Overall Performance on LIVE 3D Image Database

As shown in Table 1, the values of PLCC, SROCC, KROCC, and RMSE of each metric are reported, where the indicator that gives the best performance is highlighted in bold. It can be observed that the overall performance of the proposed framework on Database Phase-I significantly outperforms other IQA metrics. Shao’s scheme also achieves rather competitive performances. The outstanding performance partially demonstrates the efficiency of the proposed framework.

**Table 1.** Overall performances on LIVE 3D image database

Criteria	LIVE 3D image database phase I				LIVE 3D image database phase II			
	PLCC	SROCC	KROCC	RMSE	PLCC	SROCC	KROCC	RMSE
DIIVINE [2]	0.9252	0.9233	0.7474	6.2245	0.7758	0.7707	0.6023	7.1221
BLIINDS-II [4]	0.9117	0.9087	0.7220	6.7368	0.7865	0.7167	0.5427	6.9708
BRISQUE [3]	0.9119	0.9083	0.7197	6.7314	0.7513	0.7242	0.5481	7.4496
Lin [7]	0.8645	0.8559	0.6559	8.2424	0.6584	0.6375	0.4701	8.4956
Chen [8]	0.9161	0.9153	0.7360	6.5740	0.9067	0.9068	0.7314	4.7587
Shao-A [9]	0.9350	0.9251	-	5.8155	0.8628	0.8494	-	5.7058
Shao-B [11]	0.9565	0.9449	-	4.7552	0.9265	<b>0.9106</b>	-	4.3381
Proposed	<b>0.9594</b>	<b>0.9447</b>	<b>0.8004</b>	<b>4.5086</b>	<b>0.9350</b>	0.9097	<b>0.7347</b>	<b>4.2739</b>

Results on asymmetric distorted stereopairs are also listed. According to the experimental results on Database Phase II which contains both symmetric and asymmetric distortion, the proposed framework overtakes the other metrics by a large extent. The significant difference further confirms the previous conclusion that our framework can effectively predict the quality of stereoscopically viewed images. However, predicting the quality of asymmetric distorted stereopairs is still a challenge, since all the metrics show less consistency with subjective test.

Considering that samples of training and test are selected from the same dataset, experiments on individual databases are not sufficient to explain the

**Table 2.** Cross-database performances on LIVE 3D image database (training/test)

Criteria	LIVE I /LIVE II		LIVE II /LIVE I	
	PLCC	SROCC	PLCC	SROCC
DIIVINE [2]	0.6161	0.5177	0.5375	0.4954
BLIINDS-II [4]	0.6986	0.6716	0.7652	0.7398
BRISQUE [3]	0.5949	0.4582	0.5718	0.5564
Shao-B [11]	0.7791	0.7514	<b>0.8936</b>	<b>0.8917</b>
Proposed	<b>0.8127</b>	<b>0.7809</b>	0.8464	0.8280

generality and stability of the evaluation model, and therefore cross-database experiments are also carried out in this part. Table 2 also gives the detailed information of cross-database test. Note that two FR metrics (Lin, Chen and Shao’s schemes) are not reported. Obviously, the performance of the metrics significantly declines compared with individual dataset, because the source images and distortion types are not consistent. Another probable reason would be that Database Phase-I only contains both symmetrically distorted stereopairs. However, our metric still has a relatively good predictive ability, when metrics are trained on LIVE phase I and tested on LIVE phase II.

**Table 3.** Running time

Criteria	DIIVINE [2]	BLIINDS-II [4]	BRISQUE [3]	Lin [7]	Chen [8]	Ours
Time(s)	6800	21923	<b>37</b>	82	5143	145

In addition to the excellent performance on predicting image quality, the proposed method is also computationally efficient, as shown in Table 3. The total running time on LIVE 3D Image Database-Phase I (365 stereopairs with a resolution of  $640 \times 360$  for each view) is only 145s on a computer with a Core i7 CPU. It means that predicting a stereopair costs less than 0.4s. Although it is not the most efficient one, it achieves better balance between accuracy and time. Compared with those metrics which spend a lot of time in modeling complex binocular visual properties (e.g. Chen’s scheme) and computing NSS features (BLIINDS-II), the proposed method is much less time-consuming. However, designing an quality assessment index applicable to real-time image (video) processing is still challenging.

## 5 Conclusions

In this paper, an ocular dominance based quality index for stereoscopic images is proposed, in which the NSS features are used to describe the visual degradation on stereopairs. Based on the fact that binocular adding and subtracting are

closely related to stereo perception, we use the NSS features from three channels (namely, binocularity, adding, and subtracting) to predict visual quality. Experiments further confirm that the proposed framework is highly consistent with subjective test, and the computation is relatively efficient.

In the future, we will pay much attention to the research on binocular visual properties, and explore more effective ways to describe image degradation.

**Acknowledgments.** The heading should be treated as a This research is partially supported by National Natural Science Foundation of China (No. 61471260), and Natural Science Foundation of Tianjin (No. 16JCYBJC16000).

## References

1. Lin, W., Kuo, C.C.J.: Perceptual visual quality metrics: a survey. *J. Vis. Commun. Image Represent.* **22**(4), 297–312 (2011)
2. Moorthy, A.K., Bovik, A.C.: Blind image quality assessment: from natural scene statistics to perceptual quality. *IEEE Trans. Image Process. Publ. IEEE Sig. Process. Soc.* **20**(12), 3350–3364 (2011)
3. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc.* **21**(12), 4695 (2012)
4. Saad, M.A., Bovik, A.C., Charrier, C.: Blind image quality assessment: a natural scene statistics approach in the DCT domain. *IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc.* **21**(8), 3339 (2012)
5. Fei, G., Tao, D., Gao, X., et al.: Learning to rank for blind image quality assessment. *IEEE Trans. Neural Netw. Learn. Syst.* **26**(10), 2275–2290 (2015)
6. Bosc, E., Pepion, R., Callet, P.L., et al.: Towards a new quality metric for 3-D synthesized view assessment. *IEEE J. Sel. Top. Signal Process.* **5**(7), 1332–1343 (2011)
7. Lin, Y.H., Wu, J.L.: Quality assessment of stereoscopic 3D image compression by binocular integration behaviors. *IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc.* **23**(4), 1527 (2014)
8. Chen, M.J., Su, C.C., Kwon, D.K., et al.: Full-reference quality assessment of stereopairs accounting for rivalry. *Image Commun.* **28**(9), 1143–1155 (2013)
9. Shao, F., Li, K., Lin, W., et al.: Full-reference quality assessment of stereoscopic images by learning binocular receptive field properties. *IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc.* **24**(10), 2971–2983 (2015)
10. Ryu, S., Sohn, K.: No-reference quality assessment for stereoscopic images based on binocular quality perception. *IEEE Trans. Circuits Syst. Video Technol.* **24**(4), 591–602 (2014)
11. Shao, F., Tian, W., Lin, W., et al.: Toward a blind deep quality evaluator for stereoscopic images based on monocular and binocular interactions. *IEEE Trans. Image Process.* **25**(5), 2059–2074 (2016)
12. Wang, J., Rehman, A., Zeng, K., et al.: Quality prediction of asymmetrically distorted stereoscopic 3D images. *IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc.* **24**(11), 3400–3414 (2015)
13. Sheikh, H., Wang, Z., Cormack, L., Bovik, A.: Live image quality assessment database release 2 (2005)