



No-Reference Stereoscopic Video Quality Assessment Based on Spatial-Temporal Statistics

Jiufa Zhang^(✉), Lixiong Liu, Jiachao Gong, and Hua Huang

Beijing Laboratory of Intelligent Information Technology,
Beijing Institute of Technology, Beijing, People's Republic of China
{zhangjyf, lxliu, jc_gong, huahuang}@bit.edu.cn

Abstract. Stereoscopic video quality assessment (SVQA) has become the necessary support for 3D video processing while the research on efficient SVQA method faces enormous challenge. In this paper, we propose a novel blind SVQA method based on monocular and binocular spatial-temporal statistics. We first extract the frames and the frame difference maps from adjacent frames of both left and right view videos as the spatial and spatial-temporal representation of the video content, and then use the local binary pattern (LBP) operator to calculate spatial and temporal domains' statistical features. Besides, we simulate binocular fusion perception by performing weighted integration of generated monocular statistics to obtain binocular scene statistics and motion statistics. Finally, all the computed features are utilized to train the stereoscopic video quality prediction model by a support vector regression (SVR). The experimental results show that our proposed method achieves better performance than state-of-the-art SVQA approaches on three public databases.

Keywords: Stereoscopic video quality assessment · Spatial-temporal · Structural statistics · No-reference

1 Introduction

In recent years, more and more 3D video contents have been demanded and produced. The need for efficient 3D video quality prediction approaches is significantly increasing. The quality representation methods for 3D video mainly include Quality of Service (QoS) and Quality of Experience (QoE). Since human is the ultimate receiver of video contents, the research on QoE becomes particularly important in the field of 3D video perception. Therefore, it is of great significance to study the perceptual quality evaluation methods of 3D video, especially objective methods.

Depending on whether pristine stereo videos can be used, objective stereoscopic video quality assessment (SVQA) methods can be divided into full-reference (FR), no-reference (NR) and reduced-reference (RR) methods. Since it is difficult to guarantee available pristine stereo videos, making it more practical to study the NR SVQA method. In this paper, we consider addressing the problem of NR stereoscopic video quality prediction.

Some SVQA algorithms have been developed, which have good prediction performance in recent years. However, there still exists room for improving on SVQA methods in terms of prediction performance and compute complexity. The initial SVQA methods simply utilize existing IQA methods such as PSNR and SSIM [1], by obtaining per frame's evaluation score from a video, then using average pooling method to obtain the video's quality score. Obviously, these methods only focus on the spatial information in the video but ignoring the temporal features, depth information, or binocular effects in human visual system (HVS). Consequently, their quality scores are unsatisfactorily consistent with the subjective perception scores. To address this issue, some researchers considered the above aspects to design an efficient SVQA model. Han et al. [2] proposed a FR SVQA method 3D-STs, which utilized structural tensor for salient areas from adjacent frames to predict video quality scores. Yu et al. [3] considered salient frames and binocular perception according to the internal generative mechanism of HVS, and designed a RR SVQA framework. However, these two methods rely on the salient regions and take into account spatial changes on video quality, but ignore the impact of temporal changes on video quality.

Very recently, several SVQA methods have been presented by integrating temporal information. Qi *et al.* [4] proposed a stereo just-noticeable difference model which mainly considers the masking effect in both spatial and temporal domains to evaluate the perceptual quality for stereo videos. Galkandage *et al.* [5] used IQA methods based on a HVS model that considers binocular suppression and recurrent excitation to evaluate frame quality, and then introduced an optimized temporal pooling method to associate the frame quality with the video quality. Yang *et al.* [6] utilized spatial and temporal information in curvelet domain and spatial-temporal optical flow features, and proposed a blind SVQA model named BSVQA. In addition, Yang *et al.* [7] jointly focused on spatial-temporal salient model and sparse representation to calculate the de-correlated features and then predicted stereo video's quality scores by using the deep-learning network. Chen *et al.* [8] integrated natural scene statistics and auto-regressive prediction-based disparity entropy measurement to propose a NR SVQA method. These methods fully consider the video spatial information and even the binocular perceptual characteristics, but utilize less temporal information or just use it as the pooling strategy for spatial features. Besides, these methods execute complex spatial-temporal domain transforms or train deep learning networks for predicting stereo video quality scores, which makes them time-consuming and complex. Considering the impact of binocular perception based on weighted fusion [9, 10] and enriched structural information in image, we attempt to simulate spatial-temporal processing by simply computing monocular and binocular spatial-temporal statistics in our proposed SVQA method.

In this paper, we propose a blind SVQA method that uses enriched structural statistics of spatial-temporal video content. The method firstly extracts the monocular structural statistics from both left and right stereo video frames in spatial domain. We then account for the impact of binocular fusion and weight generated monocular features, yielding binocular statistical features of left and right videos. In addition, following recent evidence regarding the measurement of temporal information of stereoscopic videos by motion intensity features [6], we utilize the frame difference maps from adjacent frames of both view videos as motion representation in spatial-

temporal domain. Similar to the process of the spatial information, we perform binocular integration on the monocular features extracted from the frame difference maps. We finally use both monocular and binocular features obtained from the left and right video frames and corresponding frame difference maps to train the stereoscopic video quality prediction model by a support vector regression (SVR) [11].

The rest of this paper is organized as follows. Section 2 details the framework of our proposed method and features extracted from stereoscopic video sequences. In Sect. 3 we test the performance of our method on three public SVQA databases. We conclude the paper and discuss future work in Sect. 4.

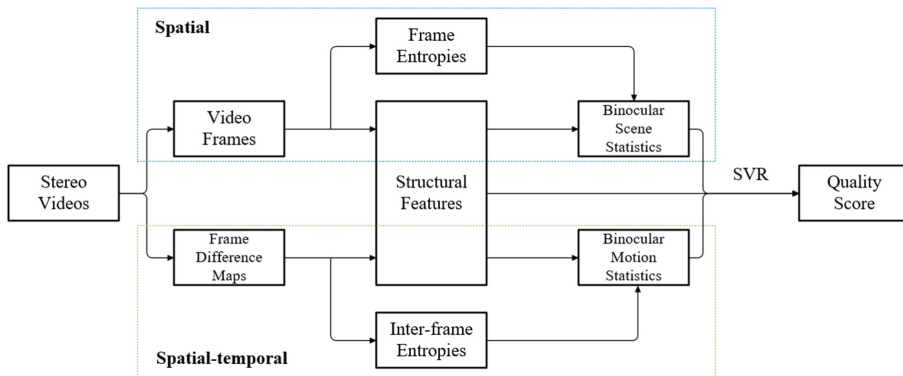


Fig. 1. Framework of our proposed NR SVQA method.

2 Proposed Method

In this section, we present our proposed NR SVQA and show its framework in Fig. 1. As shown in Fig. 1, stereo videos are first processed into video frames associated with monocular spatial information and frame difference maps corresponding to spatial-temporal motion information. We then use the local binary pattern (LBP) operator to calculate the monocular structural statistical features in spatial and spatial-temporal domain. Considering that the binocular perception, especially binocular fusion, plays an important role in the human visual perception of 3D contents, we further integrate the spatial/spatial-temporal statistics by weighting them to obtain binocular scene/motion statistics. Finally, a SVR is used to map all the monocular and binocular statistical features to stereoscopic video quality predictions.

2.1 Overview of LBP

Since HVS is sensitive to structural information in natural scene [1] and local descriptors can effectively represent scene's structural information [6], we utilize the LBP operator, an efficient structural texture descriptor, to process the structural information obtained from the stereo videos. Different from the traditional LBP calculation, here we simply consider the binary relationship of values between the central

pixel and its neighborhood pixels in the local image region to obtain the non-uniform LBP operator. For a local region of image, we define the LBP operator through calculating the binary relationship between central pixel p_c and its neighborhood pixels p_i as

$$LBP_{P,R} = \sum_{p=0}^{P-1} \begin{cases} 1, & p_c - p_i \geq 0 \\ 0, & p_c - p_i < 0 \end{cases} \quad (1)$$

where P means the number of neighborhood pixels, R is the radius of the neighborhood, and i denotes the positional order of neighboring pixels. In our proposed method, we set the parameters $P = 8$ and $R = 1$ to empirically simplify the algorithm complexity. It is obvious that $LBP_{P,R}$ has 9 output patterns. Thus, we can transform a natural image into an LBP statistical map with structure patterns between 0 and 8. Finally, we calculate the LBP statistical distribution features as the features of the stereoscopic videos. The structural statistical features are given by

$$h(k) = \frac{G(k)}{\sum_{k=0}^P G(k)} \quad (2)$$

where k denotes the possible structure patterns, and $G(\cdot)$ denotes the number each pattern occurs, which is given by

$$G(k) = \sum_{m=1}^M \sum_{n=1}^N f(LBP_{P,R}(m, n), k) \quad (3)$$

where M and N mean image size, (m, n) are the coordinates of image pixels, and $f(\cdot)$ is used to calculate the correspondence as

$$f(x, y) = \begin{cases} 1, & x = y \\ 0, & \text{else} \end{cases} \quad (4)$$

2.2 Spatial Statistical Features

Due to the complexity of stereo videos, different types and degrees of distortion may occur symmetrically or asymmetrically in both left and right view videos, causing the change of structural information in spatial domain [2]. Consequently, for a stereo video, we first extract its left and right video frames, which can maximize the retention of spatial information. Then, we utilize the LBP operator described above to extract spatial statistical features of left and right video frames, respectively. Different from viewing 2D videos, human eyes perceive differences of binocular perception, such as binocular fusion, binocular rivalry and binocular suppression when viewing stereoscopic videos [3]. Since the binocular perception which has been fully studied in SIQA can greatly influence the perceptual quality when viewing stereo scenes [9, 10], it is valuable to imitate the binocular perception during processing the stereo videos' spatial and temporal information. Since the hybrid combination of features can model 3D image quality based on Bayesian theory [20], instead of directly merging left and right

view video frames, we simplify the binocular fusion by pooling monocular spatial statistics to represent binocular perception, which is inspired by [19, 20]. Considering that the entropy can effectively represent the spatial changes caused by distortion [12] or scene changes occurring in left and right frames, we further integrate computed spatial statistics of left and right video frames with the weighted global entropy to generate binocular spatial statistics h_s :

$$h_s = \frac{\varepsilon_{sl} + C}{\varepsilon_{sl} + \varepsilon_{sr} + C} \cdot h_{sl} + \frac{\varepsilon_{sr} + C}{\varepsilon_{sl} + \varepsilon_{sr} + C} \cdot h_{sr} \tag{5}$$

where h_{sl} and h_{sr} are the monocular statistical features of the left and right video frames respectively, C is a constant to avoid the instability, and ε_{sl} and ε_{sr} are the entropy of the corresponding left and right video frames respectively, which is defined as

$$\varepsilon = - \sum_v p(v) \log_2 p(v) \tag{6}$$

where v mean the pixel value from 0 to 255 in an image, while $p(\cdot)$ means the empirical probability density. Exemplar left and right video frames from the SVQA database [4] and corresponding calculated spatial statistics h_{sl} , h_{sr} and h_s are shown in Fig. 2. It can be seen that there exists strong correlation between binocular statistics and monocular features.

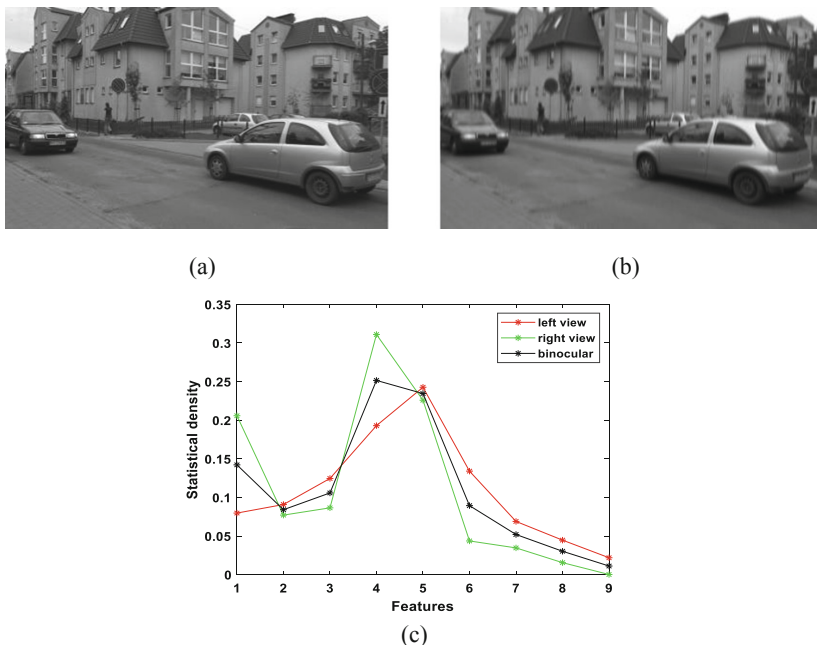


Fig. 2. Video frames and their corresponding statistical feature density. (a) The left video frame (b) the right video frame, and (c) the statistical feature density of monocular and binocular statistics.

2.3 Spatial-Temporal Statistical Features

Considering that videos contain rich temporal information such as motion intensity, which can affect visual perception significantly [6, 13], several SVQA methods have been developed to address the Stereoscopic VQA problem by using temporal information. Since the frame difference map as a widely used spatial-temporal representation, can be used to highlight motion information that exists in the video [6], we also calculate a frame difference map I_d from both adjacent frames of left and right view videos to show temporal changes. The frame difference map is defined as

$$I_d = |I_t - I_{t+1}| \quad (7)$$

where I_t and I_{t+1} denote two adjacent frames at times t and $t + 1$ from left or right view video, respectively.

Extracting structural statistical features from frame difference maps can effectively represent the intensity of motion in the scene [6]. Thus, like the spatial statistical feature extraction, the spatial-temporal statistical features of frame difference maps of both videos, which are denoted as h_{dl} and h_{dr} , are calculated by using LBP structural operator firstly. Then, considering the pooling of monocular features can simulate binocular perception [20] and the motion perception is an important part of binocular perception, we assume that the binocular motion statistics h_d is integrated to reflect binocular motion perception by weighting monocular motion statistics, which is defined as

$$h_d = \frac{\varepsilon_{dl} + C}{\varepsilon_{dl} + \varepsilon_{dr} + C} \cdot h_{dl} + \frac{\varepsilon_{dr} + C}{\varepsilon_{dl} + \varepsilon_{dr} + C} \cdot h_{dr} \quad (8)$$

where ε_{dl} and ε_{dr} are the entropy of the corresponding frame difference maps of left and right view videos, respectively. Figure 3 shows the frame difference maps computed on two different stereo videos from the SVQA database and their corresponding binocular statistical density. As we can see from Fig. 3(a)–(d), frame difference maps can highlight the motion intensity between adjacent frames and possible temporal distortion. The binocular statistical features extracted from different frame difference maps can also show the feature distribution diversity for different motion scenes as shown in Fig. 3(e).

In summary, we extract monocular statistics h_{sl} , h_{sr} , h_{dl} and h_{dr} from the left and right view video frames, and weighted binocular statistics h_s and h_d from their corresponding frame difference maps. These features are calculated from frame-level in video sequence. In this paper, we simply take the average of each frame's features as the video-level features. Furthermore, considering multiscale processing has been proved to improve the efficacy of IQA algorithms [9], we extract 9 features computed on all six types of statistics at two scales, yielding $6 \times 2 \times 9 = 108$ features. All these features are empirically trained a SVR, whose kernel function is RBF (Radial Basis Function), margin of tolerance defaults to 0.001 and penalty factor is set to 5, to predict stereoscopic video quality scores [11].

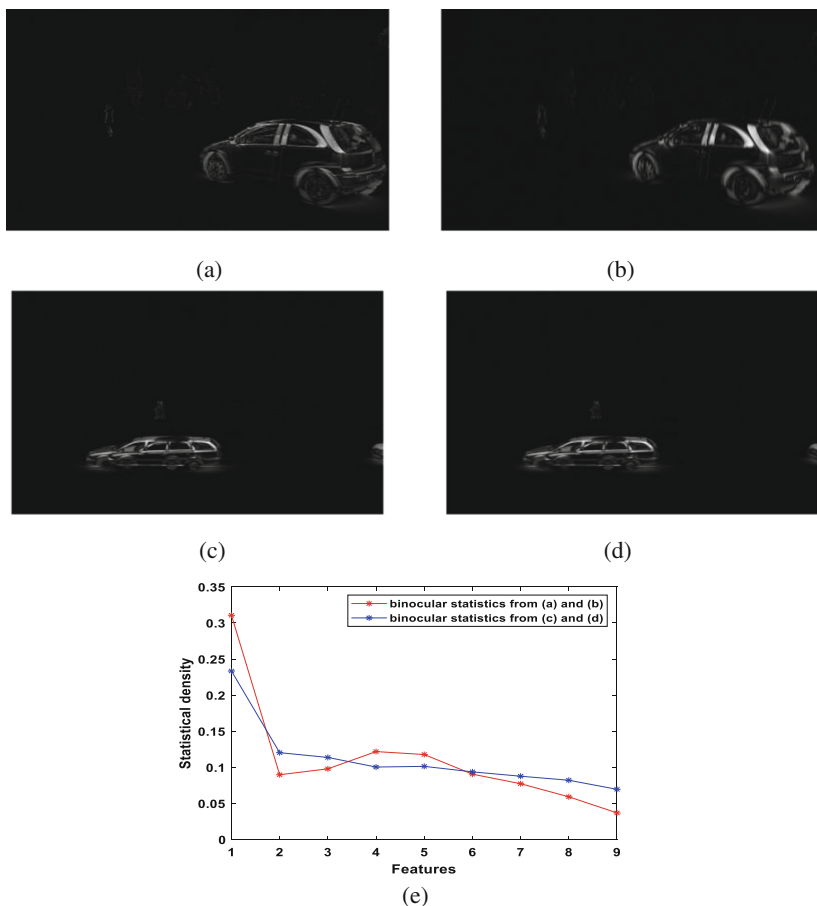


Fig. 3. Frame difference maps from different stereo videos and their corresponding binocular motion statistical feature density. (a) The left frame difference map and (b) the right frame difference map of Poznan Street sequence, (c) the left frame difference map and (d) the right frame difference map of Outdoor sequence, and (e) the statistical feature density of binocular statistics from two groups of frame difference maps.

3 Experimental Results

We verified the performance of our proposed model on three public SVQA databases. These databases are the SVQA database [4], the Waterloo-IVC 3D Video Quality Databases Phase I [14] and Phase II [15]. The SVQA database mainly contains two distortion types: Blur and improper H.264 compression. There exist 450 stereo video sequences of frame rates 25 fps. The Waterloo-IVC 3D Video Quality Database Phase I database contains totally 176 stereo video sequences which are obtained from 4 pristine stereo video sequences by compressing symmetrically or asymmetrically. Besides, the Waterloo-IVC 3D Video Quality Database Phase II contains 528 stereo video sequences

obtained from different types of coding and levels of low-pass filtering symmetrically or asymmetrically. We used the Pearson linear correlation coefficient (PLCC), the Spearman rank correlation coefficient (SRCC), the Kendall rank-order correlation coefficient (KRCC) and the Root mean squared error (RMSE) as the evaluation criteria. All three databases were randomly divided into two independent subsets: 80% for training and the remaining 20% for testing. We repeated the training-testing process 1000 times and used the median values of evaluation criteria across 1000 iterations as the final performance metrics. Note that we used the luminance component of the videos, i.e. sample Y components from the YUV videos.

3.1 Performance on Three SVQA Databases

We selected the popular IQA algorithms PSNR and SSIM [1], and several state-of-the-art VQA models PHVS-3D [16], 3D-STC [2], SJND-SVA [4], VQM [17], Sliva [18], Yang [6], and Yang [7], to make the performance comparison on all three databases. Since the source codes of these SVQA approaches are not publicly available, we obtained their experimental results from the original papers [2, 4, 6, 7, 16–18]. In particular, we also listed the results of weighted VQA model mentioned in [15] for the performance comparison on both Waterloo-IVC 3D Video Quality Databases Phase I and Phase II. The results are shown in Fig. 4 and Tables 1, 2 and 3, respectively.

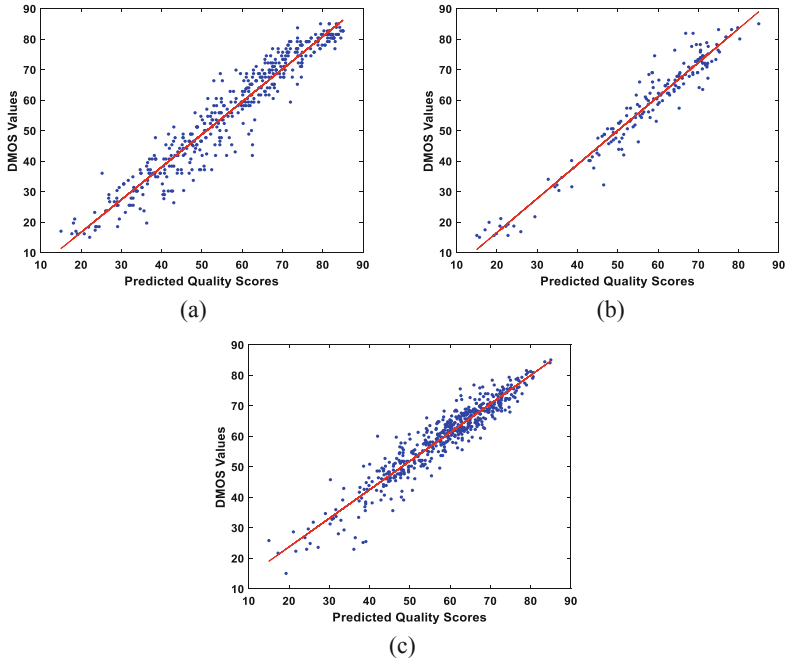


Fig. 4. Scatter plots of predicted quality scores of our proposed method against DMOS on the (a) SVQA, (b) Waterloo-IVC 3D Video Quality Phase I, and (c) Waterloo-IVC 3D Video Quality Phase II.

Table 1. Comparison of seven methods on the SVQA database.

Metric	SRCC	PLCC	KRCC	RMSE
PSNR	0.5319	0.5395	0.3767	0.8374
SSIM	0.4988	0.5025	0.3488	0.8598
PHVS-3D [16]	0.7195	0.7082	0.5353	0.7021
3D-STIS [2]	0.8338	0.8311	0.6553	0.5520
SJND-SVA [4]	0.8379	0.8415	0.6650	0.5372
Yang [6]	0.9175	0.9208	0.7730	0.3709
Yang [7]	0.9111	0.9141	0.7605	0.4018
Proposed	0.9247	0.9300	0.7719	0.3625

Table 2. Comparison of seven methods on the Waterloo-IVC Phase I.

Metric	SRCC	PLCC	KRCC	RMSE
PSNR	0.5335	0.7085	0.3850	15.4516
SSIM	0.4789	0.5045	0.3436	18.9039
VQM [17]	0.6321	0.7912	–	13.3905
Weighted VQM [15]	0.8655	0.9191	–	8.6273
Silva [18]	0.6856	0.7416	–	14.6893
Proposed	0.8949	0.9383	0.7380	7.4890

Table 3. Comparison of seven methods on the Waterloo-IVC Phase II.

Metric	SRCC	PLCC	KRCC	RMSE
PSNR	0.3152	0.3286	0.2190	11.6362
SSIM	0.3374	0.3237	0.2432	11.6001
VQM [17]	0.6287	0.7019	–	8.7759
Weighted VQM [15]	0.8042	0.8496	–	6.4976
Silva [18]	0.5184	0.5566	–	10.1051
Proposed	0.8852	0.8814	0.7085	5.8018

Table 4. Performance of spatial and spatial-temporal features on the SVQA database.

Features	Spatial	Spatial-temporal	All
SRCC	0.9140	0.8875	0.9247

Table 5. Computational complexity of three methods on the SVQA database.

Metric	Time (s)
PSNR	4.4490
SSIM	41.0248
Proposed	105.3997

From Tables 1, 2 and 3, it may be shown that our proposed model performed better than all other metrics on the SVQA database, the Waterloo-IVC 3D Video Quality Phase I and Phase II databases. The scatter plot distribution in Fig. 4 shows the scatter plots of predicted quality scores of our proposed method versus DMOS on the three SVQA database. It may be observed that our proposed method correlated well with human subjective judgements on all three databases.

3.2 Performance of Spatial and Spatial-Temporal Features

We tested the performance of spatial and spatial-temporal separately on the SVQA database to verify the validity of the computed features. The SROCC results are listed in Table 4. Clearly, both spatial and spatial-temporal features can capture space-time distortion well, and a combination of these feature achieves better performance.

3.3 Computational Complexity

We also tested the compute complexity of our proposed method on the SVQA database, using a PC with i5-3.2 GHz CPU and 16 GB RAM. Since the source codes of all of selected VQA algorithms are not publicly available and there are no compute complexity experiments in their literatures, only the compute complexity of PSNR and SSIM are analyzed here. We tested three methods on a same pair of videos, and the running time results are listed in Table 5. Clearly, our method is slower than PSNR and SSIM. Considering that our method has better prediction as compared with all of selected methods performance and is not computed in transform domain, our model may be referred to as a relatively fast SVQA method.

4 Conclusion

In this paper, we propose a blind SVQA method based on monocular and binocular spatial-temporal statistics. We utilize the frames of left and right view videos as the spatial representation and extract monocular statistics and weighted binocular scene statistics by using LBP operator. The generated binocular statistics only computed on low-level content makes it possible to design a fast algorithm. Since temporal changes such as motion intensity have a significant impact on stereoscopic video perception, similar to spatial statistical feature extraction, we further extract monocular statistics and binocular motion statistics from frame difference maps as spatial-temporal features. The proposed model was validated on three public SVQA databases, and shown to achieve significant performance improvement. In near future, we will pay more attention on the binocular perceptual mechanism of HVS and design more effective SVQA models.

Acknowledgments. This work is supported by the National Natural Science Foundation of China under grant 61672095 and grant 61425013.

References

1. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004)
2. Han, J., Jiang, T., Ma, S.: Stereoscopic video quality assessment model based on spatial-temporal structural information. In: *Proceedings IEEE Conference on Visual Communication and Image Processing*, pp. 1–6. IEEE, Sarawak (2013)
3. Yu, M., Zheng, K., Jiang, G., Shao, F., Peng, Z.: Binocular perception based reduced-reference stereo video quality assessment method. *J. Vis. Commun. Image Represent.* **38**, 246–255 (2016)
4. Qi, F., Zhao, D., Fan, X., Jiang, T.: Stereoscopic video quality assessment based on visual attention and just-noticeable difference models. *Signal Image Video Process.* **10**(4), 737–744 (2016)
5. Galkandage, C., Calic, J., Dogan, S., Guillemaut, J.Y.: Stereoscopic video quality assessment using binocular energy. *IEEE J. Sel. Top. Signal Process.* **11**(1), 102–112 (2017)
6. Yang, J., Wang, H., Lu, W., Li, B., Badii, A., Meng, Q.: A no-reference optical flow-based quality evaluator for stereoscopic videos in curvelet domain. *Inform. Sci.* **414**, 133–146 (2017)
7. Yang, J., Ji, C., Jiang, B., Lu, W., Meng, Q.: No reference quality assessment of stereo video based on saliency and sparsity. *IEEE Trans. Broadcast.* **64**(2), 341–353 (2018)
8. Chen, Z., Zhou, W., Li, W.: Blind stereoscopic video quality assessment: from depth perception to overall experience. *IEEE Trans. Image Process.* **27**(2), 721–734 (2018)
9. Liu, L., Liu, B., Su, C., Huang, H., Bovik, A.C.: Binocular spatial activity and reverse saliency driven no-reference stereopair quality assessment. *Signal Process. Image Commun.* **58**, 287–299 (2017)
10. Geng, X., Shen, L., Li, K., An, P.: A stereoscopic image quality assessment model based on independent component analysis and binocular fusion property. *Signal Process. Image Commun.* **52**, 54–63 (2017)
11. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**(3), 27 (2011). <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
12. Liu, L., Liu, B., Huang, H., Bovik, A.C.: No-reference image quality assessment based on spatial and spectral entropies. *Signal Process. Image Commun.* **29**(8), 856–863 (2014)
13. Jiang, G., Liu, S., Yu, M., Shao, F., Peng, Z., Chen, F.: No reference stereo video quality assessment based on motion feature in tensor decomposition domain. *J. Vis. Commun. Image Represent.* **50**, 247–262 (2018)
14. Wang, J., Wang, S., Wang, Z.: Quality prediction of asymmetrically compressed stereoscopic videos. In: *Proceedings IEEE International Conference Image Processing*, pp. 1–5. IEEE, Quebec City (2015)
15. Wang, J., Wang, S., Wang, Z.: Asymmetrically compressed stereoscopic 3D videos: quality assessment and rate-distortion performance evaluation. *IEEE Trans. Image Process.* **26**(3), 1330–1343 (2017)
16. Jin, L., Boev, A., Gotchev, A., Egiazarian, K.: 3d-DCT based perceptual quality assessment of stereo video. In: *Proceedings Eighteenth IEEE International Conference Image Processing*, pp. 2521–2524. IEEE, Brussels (2011)
17. Pinson, M., Wolf, S.: A new standardized method for objectively measuring video quality. *IEEE Trans. Broadcast.* **50**(3), 312–322 (2004)
18. De Silva, V., Arachchi, H.K., Ekmekcioglu, E., Kondoz, A.: Toward an impairment metric for stereoscopic video: a full-reference video quality metric to assess compressed stereoscopic video. *IEEE Trans. Image Process.* **22**(9), 3392–3404 (2013)

19. Fan, Y., Larabi, M.C., Cheikh, F.A., Fernandez-Maloigne, C.: No-reference quality assessment of stereoscopic images based on binocular combination of local features statistics. In: Proceedings IEEE Conference on Image Processing, pp. 3538–3542. IEEE, Athens (2018)
20. Shao, F., Li, K., Lin, W., Jiang, G., Yu, M.: Using binocular feature combination for blind quality assessment of stereoscopic images. *IEEE Signal Process. Lett.* **22**(10), 1548–1551 (2015)