STEREOSCOPIC VIDEO QUALITY ASSESSMENT MODEL BASED ON SPATIAL-TEMPORAL STRUCTURAL INFORMATION

Jingjing Han, Tingting Jiang, Siwei Ma

National Engineering Lab. for Video Technology, Key Lab of Machine Perception (MOE),
School of EECS, Peking University, Beijing, 100871 China
{jjhan, ttjiang, swma}@jdl.ac.cn

ABSTRACT

Most of the existing 3D video quality assessment methods estimate the quality of each view independently and then pool them into unique objective score. Besides, they seldom take the motion information of adjacent frames into consideration. In this paper, we propose an effective stereoscopic video quality assessment method which focuses on the inter-view correlation of spatial-temporal structural information extracted from adjacent frames. The metric jointly represents and evaluates two views. By selecting salient pixels to be processed and discarding the others, the processing speed is significantly improved. Experimental results on our stereoscopic video database show that the proposed algorithm correlates well with subjective scores.

Index Terms—Stereoscopic video quality assessment, spatial-temporal structure, inter-view correlation, human-visual system (HVS), asymmetric coding

1. INTRODUCTION

With the rapid development of 3D display technology, 3D video has entered the entertainment life of the public. Just by wearing a pair of glasses, people can enjoy the immersive feeling of stereoscopic films at the cinema or in front of a computer with three-dimensional display function. Compared with the traditional 2D video, 3D videos contain additional information and two eyes see different images at the same time. After binocular fusion, two views produce a single image and a sense of stereoscopic depth in the Human Visual System (HVS). The process of acquisition, coding, transmission and display may cause distortions with different types and intensities in the left and right views, which may result in different qualities sensed by HVS. Therefore, how to accurately evaluate the quality of 3D videos is an important problem worth exploring.

As the final receptor of a video is human-beings, the most accurate and reliable method of the quality assessment is to use the scores rated by observers, which is called the subjective quality assessment. However, this method takes too much time, labor and money, and cannot be applied in real-time applications. In order to make up the shortcomings, the objective quality assessment method is used in practice, which estimates the quality by applying a quantitative mathematical model. A major issue is whether the results obtained are consistent with the subjective feelings.

At present, stereoscopic image and video quality evaluation models can be mainly divided into two categories. The first category is assessing the quality of experience (QoE) or whether observers feel comfortable when watching 3D images and videos. An objective metric for stereoscopic crosstalk perception is proposed in [1] by considering the crosstalk level, camera baseline and scene content. In [2], the authors extract simple statistical features from both the disparity map and the spatial image and then select those which correlate well with the subjective scores using Principal Component Analysis (PCA) or Forward Feature Selection (FFS) to assess the quality of experience. The second category is to measure the degree of distortion caused by compression or coding. A no-reference objective quality assessment method aimed at JPEG coded stereoscopic images is introduced in [3]. It measures artifacts (including blockiness and zero-crossing) separately for blocks in the left and right views and evaluates the disparity between them. The work in [4] classifies pixels into edge, texture and smooth regions and assigns different weights to the local quality obtained by traditional 2D Image Quality Metrics (IQMs) of these regions and then averages between views. Wang *et al.* [5] proposed a stereo image quality assessment algorithm considering binocular spatial sensitivity. The algorithm takes Just Noticeable Difference (JND) values as the binocular spatial sensitivity of each pixel in the image pair and uses them as the weights to IQMs. In [6], the spatial frequency domain (SFD) information is extracted to estimate the quality of images coded asymmetrically. Perceptual quality metric (PQM) for 2D videos is described in [7] and its application on stereoscopic videos rendered from color plus depth format using Depth-Image-Based-Rendering (DIBR) is addressed meanwhile. All of the above algorithms calculate the quality of both views separately and use different weighting strategies without considering the correlation between views. Jin *et al.* [8] find similar blocks between views to form 3D arrays. After applying 3D-DCT transform to each 3D array,

they compute the modified MSE between the coefficients of the original and distorted blocks. Although the metric correlates well with subjective scores of four 3D video sequences according to [8], it does not take into account of the motion information of adjacent frames and depends on the performance of stereo matching which is time-consuming thus cannot be applied in real-time applications.

In this paper, we propose an effective stereoscopic video quality assessment method based on the spatial-temporal structure (STS) metric. STS is previously used for 2D videos quality evaluation in [9]. The metric constructs 3D structure tensors which consider both edges in the spatial domain and motion in the temporal domain. Descriptors including eigenvalue and eigenvectors extracted from the 3D structure tensor are compared to evaluate video quality. We extend the STS metric to stereoscopic video quality assessment by constructing a joint descriptor which represents two views jointly without disparity estimation and considers inter-view correlation between 3D structure tensors of both views. Based on the joint descriptor, the left view is evaluated as the reference view in HVS and the right view is regarded as the auxiliary view. The distortion of the video is evaluated based on the difference between the joint descriptors of the distorted video and the reference video.

The rest of the paper is organized as follows. In Section 2, the STS metric used in 2D video quality assessment is described. The details of the proposed stereoscopic video quality assessment method are introduced in Section 3. Section 4 presents the experimental results. Finally, conclusion and future works are given in Section 5.

2. SPATIAL-TEMPORAL STRUCTURE METRIC

STS metric is proposed by Wang *et al.* in [9]. They apply it to predict the quality of conventional 2D video clips in LIVE database and VQEG FR-TV Phase-I database and find it not only robust to a variety of distortions appears in video but also of low computational complexity.

The idea of STS is based on two widely accepted phenomenons of HVS. One is that edges are the basic features that arouse the greatest visual stimulation within a frame. In another word, degradation of edges in the spatial domain influences the perceived quality seriously. Another phenomenon is that motion along the temporal domain affects the quality mostly. Wang *et al.* think that motion can be represented by the position variation of edges between adjacent frames which would stretch out a plane along the motion trajectory. They set the direction perpendicular to the plane formed by spatial edge and its motion trajectory as the primary direction. Distortions within and between frames would change the localized spatial edge and motion trajectory respectively and finally resulting in a different primary direction. Therefore, the similarity between primary directions of corresponding pixels in the original video and the distorted video can reflect the visual quality.

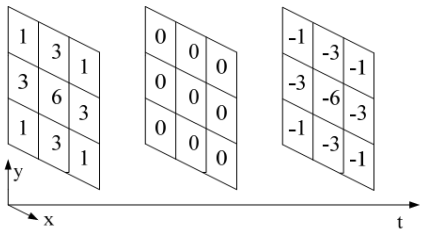


Fig. 1. 3D Sobel kernel for *t* direction [9].

Specifically, the STS metric firstly calculates the local gradients of each pixel along *x, y* and *t* directions using 3D Sobel kernels. The kernel for *t* direction is shown in Figure 1 and kernels for *x* and *y* directions are acquired by rotating it. As can be seen from Figure 1, this step utilizes not only pixels around the central one in the current frame, but also both pixels in the same spatial regions of the forward and backward frames.

Next, saliency of each pixel is computed according to the following equation [9]:

$$sal(p) = \sqrt{gx^2(p) + gy^2(p) + gt^2(p)} \quad (1)$$

where gx, gy and gt are the local gradients of pixel p along *x, y* and *t* directions respectively. If both pixels at the same location in the original video and the distorted video have a great enough saliency value, they are marked as salient pixels. These salient pixels are more likely to be edge or motion regions which play important roles in HVS as previously discussed.

The third step of the method is to generate 3D structure tensor from local gradients for each salient pixel and perform eigenvalue decomposition. The format of 3D structure tensor for a salient pixel p is [9]:

$$str(p) = \nabla g(p) \cdot \nabla g^T(p) = \begin{bmatrix} \sum_W gx^2(p) & \sum_W gx(p) \cdot gy(p) & \sum_W gx(p) \cdot gt(p) \\ \sum_W gx(p) \cdot gy(p) & \sum_W gy^2(p) & \sum_W gy(p) \cdot gt(p) \\ \sum_W gx(p) \cdot gt(p) & \sum_W gy(p) \cdot gt(p) & \sum_W gt^2(p) \end{bmatrix} \quad (2)$$

where W is the local integration window. Jacobi method [10] is performed to decompose the largest eigenvalue and its corresponding eigenvector of the 3D structure tensor. The obtained eigenvector represents the primary direction mentioned above and the largest eigenvalue reflects the strength of variation along the direction [9]. After that, each salient pixel can be represented by a largest eigenvalue λ and its corresponding eigenvector $\hat{\xi}$. The eigenvector $\hat{\xi}$ is a space vector as Equation (3).

$$\hat{\xi} = < x, y, t >, \quad x^2 + y^2 + t^2 = 1 \quad (3)$$

Then the similarity between 3D structure tensors of salient pixels in the reference video and the distorted video is calculated as Equation (4) as follows:

$$m(p') = \frac{2 \cdot \lambda_{ref}(p) \cdot \lambda_{dis}(p')}{\lambda_{ref}^{\ 2}(p) + \lambda_{dis}^{\ 2}(p')} \times \cos < \hat{\xi}_{ref}(p), \hat{\xi}_{dis}(p') > \quad (4)$$

where p and p' are corresponding salient pixels at the same location in the reference video and the distorted video respectively, λ_{ref} and λ_{dis} are the largest eigenvalues for 3D structure tensors of pixels p and p' respectively, while $\hat{\xi}_{ref}$ and $\hat{\xi}_{dis}$ denote the corresponding eigenvectors.

Finally, local scores for all salient pixels in every frame of the distorted video are averaged and thus produce a single score of the video.

3. PROPOSED ALGORITHM

In the proposed algorithm, the STS metric is applied to assess the quality of stereoscopic videos. The joint descriptor which takes into consideration of the correlation between two views is constructed and used to evaluate the video quality. Details of the proposed metric are described in the following sections.

3.1. Representation of joint descriptor

As discussed in Section 2, the eigenvector corresponding to the largest eigenvalue represents the local primary direction and the largest eigenvalue reflects the strength of variation along the direction [9]. In other words, eigenvectors indicate the directions of local variation of edge and motion information, while eigenvalues measure the intensities of the variation.

Similar to the STS metric, an initial descriptor can be obtained for each pair of pixels at the same position in the left view and the right view. The format is:

$$d_{initial} = < \lambda_L, \hat{\xi}_L, \lambda_R, \hat{\xi}_R > \quad (5)$$

where λ_L and λ_R are the largest eigenvalues of local 3D structure tensors of corresponding pixels in the left view and the right view respectively, $\hat{\xi}_L$ and $\hat{\xi}_R$ are the corresponding eigenvectors.

From the initial descriptor, we construct a joint descriptor as follows:

$$d_{joint} = < X, Y, T, \lambda_R, \Delta\alpha, \Delta\beta > \quad (6)$$

$$< X, Y, T >= \lambda_L \cdot \hat{\xi}_L =< \lambda_L x_L, \lambda_L y_L, \lambda_L t_L > \quad (7)$$

where $< X, Y, T >$ is the product of eigenvalue and eigenvector of the pixel in the left view as Equation (7), λ_R is the largest eigenvalue of the pixel in the right view, $\Delta\alpha$ is the orientation difference of $\hat{\xi}_L$ and $\hat{\xi}_R$ projected to the spatial domain (the *xoy* plane) and $\Delta\beta$ is their orientation difference in the temporal domain (the *yot* plane). An illustration of how $\Delta\alpha$ and $\Delta\beta$ are calculated is shown in Figure 2. Both of them are measured in radians.

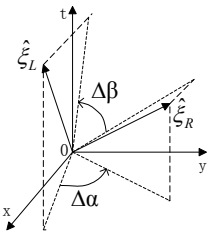
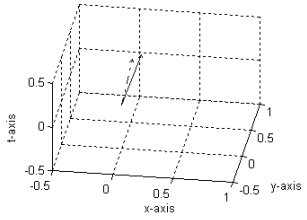


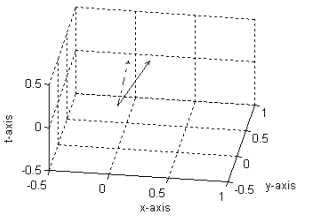
Fig. 2. Illustration of the joint descriptor. $\hat{\xi}_L$ and $\hat{\xi}_R$ are the eigenvectors for pixels at the same position in the left view and the right view respectively. $\Delta\alpha$ is the orientation difference of $\hat{\xi}_L$ and $\hat{\xi}_R$ projected to the spatial domain (the *xoy* plane). $\Delta\beta$ is the orientation difference of $\hat{\xi}_L$ and $\hat{\xi}_R$ in the temporal domain (the *yot* plane).

(a) The fourth frames of original left view and right view. Pixels located at the position of (116, 285, 4) are marked by a circle.



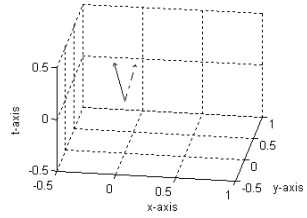
(b) Eigenvectors when QP for the left view equals to 0 and QP for the right view equals to 0.

$< \lambda_R, \Delta\alpha, \Delta\beta > =$
$< 4.9\text{x}10^5, 0.083, -0.066 >$.



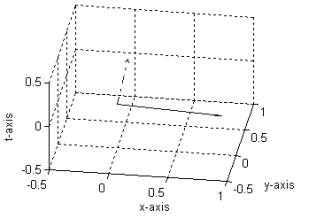
(c) Eigenvectors when QP for the left view equals to 0 and QP for the right view equals to 32.

$< \lambda_R, \Delta\alpha, \Delta\beta > =$
$< 5.4\text{x}10^5, 0.196, -0.017 >$.



(d) Eigenvectors when QP for the left view equals to 0 and QP for the right view equals to 42.

$< \lambda_R, \Delta\alpha, \Delta\beta > =$
$< 9.5\text{x}10^4, -0.198, -0.017 >$.



(e) Eigenvectors when QP for the left view equals to 0 and QP for the right view equals to 52.

$< \lambda_R, \Delta\alpha, \Delta\beta > =$
$< 4.3\text{x}10^4, 1.718, -0.017 >$.

Fig. 3. Illustration of *Book Arrival* [11] and eigenvectors of pixels located at (116, 285, 4) of the left view and the right view under different QP values. The dotted line is for the left view and the solid line is for the right view. λ_R is the largest eigenvalue of the pixel in the right view. $\Delta\alpha$ and $\Delta\beta$ are illustrated in Figure 2.

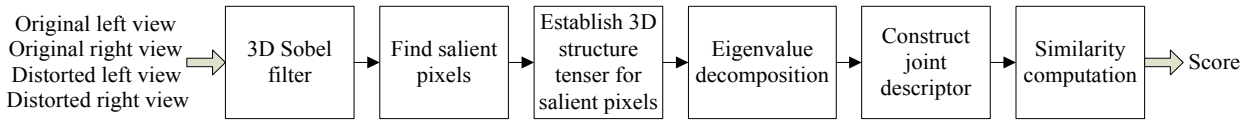


Fig. 4. Flow chart of the proposed method.

As can be seen from Equation (6), the joint descriptor contains two kinds of information. The first three elements are features of the left view and the last three elements are features of the right view relative to the left view. We regard the left view as the reference view in HVS and the right view as the auxiliary view. Therefore, the features of the left view are constructed independently. However, the features of the right view contain the inter-view correlation information represented by orientation difference between eigenvectors of pixels in the left view and the right view.

According to our test, the features of the left and right views correlate with the degree of distortion. For example, we encode the left view and right view of *Book Arrival* [11] (512x384) independently using the Joint Scalable Video Model (JSVM) reference software with different QP values (32, 42 and 52) and calculate the eigenvectors for each pixel. Figure 3(a) shows the 4th frames of original left view and right view and Figure 3(b)-(e) shows the eigenvectors of pixels located at the coordinates of (116, 285, 4) under different QP values. Eigenvectors of the pixel in the left view are in dotted line and eigenvectors of the pixel in the right view are in solid line. The values of features of the right view used in the joint descriptor are also demonstrated in Figure 3. We can see that all of them change with greater QP values. This phenomenon confirms that the proposed joint descriptor can be used to evaluate quality.

3.2. Similarity measurement

Similarity of joint descriptors of pixels in the reference video and the distorted video is computed to get the quality score at the pixel level. We assume that the joint descriptors for pixels located at the same position in the reference video and the distorted video are:

$$\begin{aligned} d_{joint}^{ref} &= < X, Y, T, \lambda_R, \Delta\alpha, \Delta\beta > \\ d_{joint}^{dis} &= < X', Y', T', \lambda_R', \Delta\alpha', \Delta\beta' > \end{aligned} \tag{8}$$

The local quality of the pair of pixels is computed according to Equation (9), where q_L donates the quality of the pixel in the left view and q_R donates the quality of pixel in the right view. Local quality q equals to the product of q_L and q_R. As left view is regarded as the reference view in HVS, the quality of pixel in left view only considers the similarity of the first three elements in the joint descriptor which are X, Y and T. However, the quality of pixel in the right view considers both the similarity of eigenvalue $q_{\lambda R}$ and the degree of change of inter-view correlation q_{ori}. The latter is calculated as one minus the average of changes of $\Delta\alpha$ and $\Delta\beta$ divided by their maximum value π, because we find that the values of changes of $\Delta\alpha$ and $\Delta\beta$ increase with greater QP values as in Figure 3.

$$\begin{aligned} q &= q_L \cdot q_R \\ q_L &= 1 - \frac{\sqrt{(X-X')^2 + (Y-Y')^2 + (T-T')^2}}{\sqrt{X^2+Y^2+T^2} + \sqrt{(X')^2 + (Y')^2 + (T')^2}} \\ q_R &= q_{\lambda R} \cdot q_{ori} \\ q_{\lambda R} &= \frac{2 \cdot \lambda_R \cdot \lambda_R'}{(\lambda_R)^2 + (\lambda_R')^2} \\ q_{ori} &= 1 - \frac{1}{2} \cdot \left(\frac{|\Delta\alpha - \Delta\alpha'|}{\pi} + \frac{|\Delta\beta - \Delta\beta'|}{\pi} \right) \end{aligned} \tag{9}$$

As can be seen from Equation (9), all of these terms are normalized to the range of [0, 1]. Greater value indicates better quality.

3.3. Procedure of the proposed algorithm

The process of the proposed algorithm is presented in Figure 4.

First of all, gradient magnitudes are calculated for all pixels using the 3D Sobel kernel introduced in Section 2.

Second, classify each pixel into salient pixel or non-salient pixel. If the saliency calculated according to Equation (1) is greater than a predefined threshold (Th = 900) for a certain pixel located at the coordinates of (x, y, t), all pixels located at the same positions in the original left and right views and the distorted left and right views are regarded as salient pixels. On the contrary, if none of the pixels located at the coordinates of (x, y, t) has a greater saliency value than the threshold, these pixels are all marked as non-salient pixels.

Third, 3D structure tensors for each salient pixel are established. Eigenvalue decomposition is performed on each 3D structure tensor.

From the above step, the initial descriptor can be obtained and then converted to the joint descriptor as mentioned in Section 3.1.

In the next step, similarity of joint descriptors of salient pixels in the original video and the distorted video is computed as Equation (9) to get the quality score at the pixel level.

Finally, we compute the video quality score as the average of the scores of every salient pixel in all frames. Each salient pixel has the same weight.

4. EXPERIMENTAL RESULTS

In order to test the performance of the proposed algorithm, we choose four stereoscopic videos to conduct subjective experiments. Details of the database and the performance of our metric compared with several previously proposed algorithms are presented below.

4.1. Test sequences and subjective experiments

Four stereoscopic videos in the format of uncompressed YUV 4:2:0 are chosen to establish our database. They are *Poznan Street* [12], *Tsinghua Classroom*, *Balloons* [13] and *Pantomime* [13]. The first frames of original left views of each sequence are illustrated in Figure 5 and more details of the sequences are demonstrated in Table 1. We cut all of these sequences into 250 frames and encode them using the Joint Multiview Video Model (JMVM) 2.1 software which is based upon the Joint Scalable Video Model (JSVM) reference software at different QP values (20, 30, 40 and 50) for both left views and right views independently and asymmetrically. Thus for each view, there are 5 distortion levels including the reference. For each sequence, there are 25 (5x5) test sequences. Therefore, there are totally 100 (5x5x4) stereoscopic video clips in our database.

In the subjective experiments, all sequences are displayed on a computer with NVIDIA GeForce GTS 450 display card and rendered by NVIDIA's 3D Vision. Observers watch these sequences by wearing wireless glasses contacted to the IR emitter. 18 observers between the ages of 20-35 score all of the sequences on the five-grade impairment scale using the single-stimulus (SS) method according to ITU-R BT.500-11 standard [14]. Each observer is trained to get familiar with the scoring scale before subjective experiments. To avoid visual fatigue, each observer rates 50 sequences for about 30 minutes without any break and then rates the remaining 50 sequences after a short rest of ten minutes. All video clips are displayed in the random order. By outlier detection [15][16][17], for each sequence, 2 observers are discarded and the remaining 16 scores are averaged to obtain the final Mean Opinion Score (MOS) of our database.

4.2. Results and analysis

In order to evaluate the performance of proposed algorithm, we compare it with two widely used 2D quality metrics which are PSNR and SSIM [18], and three previously published 3D quality metrics including PQM [7], PHVS-3D [8] and SFD [6]. For 2D quality metrics, PSNR and SSIM [18] are calculated independently for left view and right view and then averaged to obtain the final objective score of stereoscopic video. PSNR value in the undistorted case is set to be 50 instead of positive infinity, because the maximum value under distortion circumstances is around 45

Fig. 5. The first frames of original left views of all sequences in our database.

Table 1. Information of original sequences

Sequences	Resolution	Frame rate (fps)	View point	
			Left	Right
Poznan street	1920x1088	25	4	3
Tsinghua classroom	1280x720	25	0	1
Balloons	1024x768	25	0	1
Pantomime	1280x960	25	20	21

Table 2. Performance comparison for considered metrics on the overall database

Metrics	CC	SROCC	RMSE	BP
PSNR	0.9091	0.9329	0.4618	3
SSIM	0.8506	0.8731	0.5829	8
PQM	0.8610	0.8935	0.5638	4
PHVS-3D	0.7796	0.7832	0.6943	15
SFD	0.6900	0.7049	0.8023	16
Proposed	**0.9488**	**0.9398**	**0.3500**	**0**

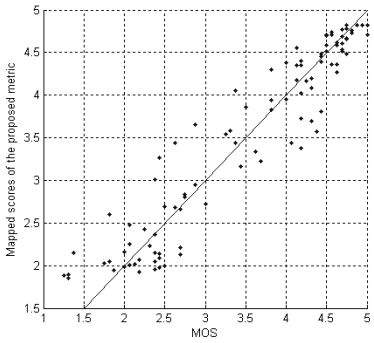


Fig. 6. The scatter plot of the MOS against mapped objective scores calculated by the proposed metric.

in our database. For SFD [6], the threshold to MAD is set to be 50 in our experiments. We apply a nonlinear mapping between subjective scores and objective scores using a 4-parameter logistic function as suggested by the VQEG Group as follows.

$$f(x) = \frac{\tau_1 - \tau_2}{1 + \exp(-\frac{x - \tau_3}{\tau_4})} + \tau_2 \tag{10}$$

Pearson correlation coefficient (CC), Spearman rank order correlation coefficient (SROCC), root mean square error (RMSE) and the number of bad point (BP) are

computed between MOS and mapped objective scores. If the difference between mapped objective score and its corresponding MOS is greater than the standard deviation of all MOSs, the objective score is regarded as a bad point.

Results on the overall database are listed in Table 2. It can be seen from Table 2 that the proposed algorithm outperforms the other metrics in all performance criteria with Pearson correlation coefficient equals to 0.9488 and SROCC equals to 0.9398. Although the PQM metric shows great alignment to MOS when applied to 3D videos in the format of color plus depth [7], it is not quite suitable for stereoscopic videos with binocular view. The PHVS-3D [8] metric does not perform well because the object motion in our database is faster than that in their own database. However, their algorithm does not take into account of motion distortion. An interesting fact in Table 2 is that the performance of the PSNR metric ranks the second. We find that its performance depends on the parameter settings in our experiments. As mentioned above, we set the maximum value of PSNR to be 50 in our experiments. Further trial is done to set greater values and the performance declines obviously. For instance, CC and SROCC equals to 0.8265 and 0.8463 respectively when the maximum value of PSNR metric is set to be 80 in the undistorted case.

Figure 6 shows the scatter plot of the MOS against mapped objective scores calculated by the proposed metric. It can be seen from the scatter plot that our metric correlates well with subjective scores.

5. CONCLUSION AND FUTURE WORKS

In this paper, we propose an objective quality evaluation method for 3D videos utilizing spatial-temporal structural information. The algorithm jointly represents and evaluates two views. In particular, we firstly select salient pixels based on the results of 3D Sobel filter. Then eigenvalues and eigenvectors are obtained from local 3D structure tensors of each salient pixel. Next, the similarity of joint descriptors constructed from eigenvalues and eigenvectors of pixels in the left view and the right view is calculated at the pixel level. Finally, all of the local scores are pooled into one global score. The experimental results show that our proposed metric correlates well with subjective quality.

Future works could be done in two aspects. One is to explore an appropriate way to distinguish different degrees of the influence of salient pixels on HVS and thus improve the performance of proposed algorithm further. The other aspect is to extend the proposed metric to evaluate multiview videos in various disparities.

6. ACKNOWLEDGEMENTS

This work was supported in part by Major State Basic Research Development Program of China (973 Program, 2009CB320903), National Science Foundation (61121002, 61103088) and National High-tech R&D Program of China(863 Program, SS2012AA010805).

REFERENCES

[1] L. Xing, J. You, T. Ebrahimi and A. Perkis, "A perceptual quality metric for stereoscopic crosstalk perception," *IEEE ICIP 2010*, pp. 4033-4036, Sept. 2010.

[2] A. Mittal, A. K. Moorthy, J. Ghosh and A. C. Bovik, "Algorithmic assessment of 3D quality of experience for images and videos," *DSP/SPE 2011*, pp. 338-343, Jan. 2011.

[3] Z. M. P. Sazzad, S. Yamanaka, Y. Kawayokeita and Y. Horita, "Stereoscopic image quality prediction," *QoMEx 2009*, pp. 180-185, July 2009.

[4] X. Mao, M. Yu, X. Wang, G. Jiang, Z. Peng and J. Zhou, "Stereoscopic image quality assessment model with three-component weighted structure similarity," *ICALIP 2010*, pp. 1175-1179, Nov. 2010.

[5] X. Wang, S. Kwong and Y. Zhang, "Considering binocular spatial sensitivity in stereoscopic image quality assessment," *IEEE VCIP 2011*, Nov. 2011.

[6] F. Lu, H. Wang, X. Ji and G. Er, "Quality assessment of 3D asymmetric view coding using spatial frequency dominance model," *3DTV-Con,* May 2009.

[7] P. Joveluro, H. Malekmohamadi, W. A. C. Fernando and A. M. Kondoz, "Perceptual video quality metric for 3D video quality assessment," *3DTV-Con*, June 2010.

[8] L. Jin, A. Boev, A. Gotchev and K. Egiazarian, "3D-DCT based perceptual quality assessment of stereo video," *IEEE ICIP 2011*, pp. 2521-2524, Sept. 2011.

[9] Y. Wang, T. Jiang, S. Ma and W. Gao, "Novel spatio-temporal structural information based video quality metric," *IEEE TCSVT 2012*, Feb. 2012.

[10] G. H. Golub and C. F. V. Loan, "Matrix computations," Jason Hopkins University Press, 1983.

[11] Fraunhofer HHI Mobile3DTV project, http://sp.cs.tut.fi/mobile3dtv/

[12] M. Domanski, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski and K. Wegner, "Poznan multiview video test sequences and camera parameters," ISO/IEC JTC1/SC29/WG11, M17050, Oct. 2009.

[13] MPEG-FTV Test Sequence Download Page, http://www.tanimoto.nuee.nagoya-u.ac.jp/~fukushima/mpegftv/

[14] Methodology for the subjective assessment of the quality of television pictures, ITU, Document ITU-R BT.500-11, Geneva, Switzerland, 2002.

[15] L. Ma, W. Lin, C. Deng and K. N. Ngan, "Study of subjective and objective quality assessment of retargeted images", *ISCAS 2012*, May 2012.

[16] IVP video quality database, http://ivp.ee.cuhk.edu.hk/research/database/subjective/index.shtml

[17] Image retargeting subjective quality database, http://ivp.ee.cuhk.edu.hk/projects/demo/retargeting/index.html

[18] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, pp. 600-612, April 2004.