SIViP
DOI 10.1007/s11760-015-0802-4

ORIGINAL PAPER

Stereoscopic video quality assessment based on visual attention and just-noticeable difference models

Feng Qi[1] · Debin Zhao[1] · Xiaopeng Fan[1] · Tingting Jiang[2]

Received: 25 September 2014 / Revised: 8 July 2015 / Accepted: 16 July 2015

Abstract With the consideration that incorporating visual saliency information appropriately can benefit image quality assessment metrics, this paper proposes an objective stereoscopic video quality assessment (SVQA) metric by incorporating stereoscopic visual attention (SVA) to SVQA metric. Specifically, based upon the multiple visual masking characteristics of HVS, a stereoscopic just-noticeable difference model is proposed to compute the perceptual visibility for stereoscopic video. Next, a novel SVA model is proposed to extract stereoscopic visual saliency information. Then, the quality maps are calculated by the similarity of the original and distorted stereoscopic videos' perceptual visibility. Finally, the quality score is obtained by incorporating visual saliency information to the pooling of quality maps. To evaluate the proposed SVQA metric, a subjective experiment is conducted. The experimental result shows that the proposed SVQA metric achieves better performance in comparison with the existing SVQA metrics.

Keywords Stereoscopic video quality assessment · Just-noticeable difference · Stereoscopic visual attention · Binocular masking

✉ Xiaopeng Fan
fxp@hit.edu.cn

Feng Qi
fqi@jdl.ac.cn

Debin Zhao
dbzhao@hit.edu.cn

Tingting Jiang
ttjiang@jdl.ac.cn

[1] Harbin Institute of Technology, 92 West Dazhi Street, Harbin 150001, China

[2] National Engineering Lab for Video Technology, Peking University, Beijing 100087, China

1 Introduction

Stereoscopic video quality assessment (SVQA) is one of the most fundamental yet challenging issues in 3D video processing technology. Lots of efforts have been devoted to the study of SVQA in the last decade. Ha et al. [1] designed a quality assessment method by considering the factors of temporal variation and disparity distribution. Based on the associated binocular energy and the binocular signal generated by simple and complex cells, Bensalma et al. [2] proposed a binocular energy quality metric to assess quality for stereoscopic images. Shao et al. [3] proposed a quality assessment metric for stereoscopic images by considering binocular perception and combination properties. Based on objective metrics of 2D video, Joveluro et al. [4] proposed a perceptual quality metric (PQM) for SVQA. Jin et al. [5] proposed a novel SVQA method based on 3D-DCT transform. Lu et al. [6] proposed a spatial frequency dominance (SFD) model by considering the observed phenomenon that spatial frequency determines view domination under the action of HVS. Han et al. [7] proposed a 3D spatial–temporal structural (3D-STS) metric to evaluate the inter-view correlation of spatial-temporal structural information extracted from adjacent frames. Inspired by these prior works, based on binocular visual properties, we establish a visual perception model for SVQA.

HVS has complicated visual characteristics, and it is still an up-to-date sealed book in physiology and psychology. In psychophysics, just-noticeable difference (JND) is a significant approach to detecting the smallest difference between starting and secondary levels of a particular sensory stimulus. Since its good approximation of many sensory dimensions [2], JND has been an active research in the study of visual perception [8–11]. As one of the most important visual characteristics, visual attention makes human focusing certain

Published online: 06 August 2015

salient regions in the visual field. Thus, the distortions in these salient regions would affect the subject's judgment on the overall quality of the stereoscopic video. To reflect this visual characteristic, a stereoscopic visual attention (SVA) model is proposed to extract the visual saliency information from stereoscopic videos. In the proposed SVQA metric, the stereoscopic video's quality maps are first calculated by the similarity of perceptual visibility between the reference and distorted stereoscopic videos. Then, the visual saliency information is introduced as a weighting function in the pooling of quality maps.

The main contributions of our work are listed as follows:

(1) A novel SVQA metric is proposed based upon the perceptual visibility of human binocular visual system and the visual saliency information of stereoscopic videos.
(2) A stereoscopic JND (SJND) model is proposed to estimate perceptual visibility of human binocular visual system, in which four visual characteristics are taken into account, e.g., sensitivity of luminance contrast, spatial masking, temporal masking and binocular masking.
(3) A SVA model is proposed to extract the visual saliency information of the stereoscopic video, including intraframe's saliency, interframe's saliency and binocular saliency.
(4) A subjective experiment is conducted to establish the ground-truth database for stereoscopic videos.

2 Related work and motivations

2.1 JND model

JND reveals the limitation of the human visual perception, and it is widely used in video or image quality evaluation [12]. In 3D image processing, only few stereoscopic JND models are available for the human binocular visual perception. Zhao et al. [13] proposed binocular JND (BJND) model to mimic the basic binocular vision properties in response to asymmetric noises in a pair of stereoscopic images. Silva et al. [14] derived a mathematical model to explain the depth JND. Based on the idea that human has different visual perception for the objects with different depths, Li et al. [15] proposed a depth perception-based joint JND model for stereoscopic images.

Human binocular visual system allows us to perceive stereoscopic spatiotemporal visual information from the outside world. However, what people see is not a direct translation of retinal stimuli; it involves complicated psychological inference [16]. Based on the free energy theory in brain theory and neuroscience [17], human visual system (HVS) adaptively excludes the disorder tendency information in a continued movement scene and endeavors to focus on the definite content of the perceived image. It indicates that HVS exerts the minimum noticeable difference to perceive the visual information. As human binocular visual system, the minimum noticeable difference decides various visibility limitations, such as luminance contrast, disparity, binocular rivalry and binocular masking [18]. Most of the previous stereoscopic JND models focused on the influences between the disparity and depth perception. In this paper, we try to characterize the binocular visual perception by considering the multiple visual masking characteristics.

2.2 Stereoscopic visual attention (SVA) model

Visual attention models have been well investigated in the last decade, which gave rise to a series of important theories and models for image quality assessment [19]. Most of the existing SVA models are the extensions of traditional single-view attention modeling methods by considering depth information [20–22]. Based on multiple perceptual stimuli, Zhang et al. [20] proposed a bottom-up SVA model to simulate stereoscopic vision in HVS. Dittrich et al. [21] detected 3D saliency of stereoscopic video by three components: salient colors in individual frames, salient information derived from camera and object motion, and depth saliency. Wang et al. [22] proposed a depth-saliency-based model for 3D visual attention and conducted a binocular eye-tracking experiment to create ground-truth database. All of their 3D visual attention models need disparity map to generate the stereoscopic saliency map.

Wang et al.'s [19] and Zhang et al.'s [23] studies indicated that information content weighting plays a significant role in pooling stage and leads to consistent improvement in the performance of IQA algorithms. Therefore, inspired by their work, we propose a novel stereoscopic visual attention model and incorporate it to our SVQA metric.

2.3 Motivation

To evaluate the quality of stereoscopic video pairs is a psychophysical process. Subjects make evaluation through the relationship between human sense and image stimulus. This evaluation process can be decomposed into three stages (as shown in Fig. 1). The first is visual perception, in which the subject perceives visual information from the given stereoscopic video. The second is visual attention, in which the subject focuses on the significantly local regions of the stereoscopic video. The third is evaluation, in which the sub-

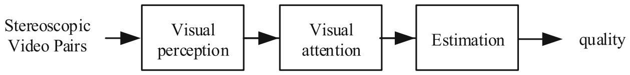

Fig. 1 The flowchart of the evaluating process

ject judges the degraded level between the reference and distorted stereoscopic videos. Therefore, according to the evaluation process, the proposed SVQA metric consists of three parts and it will be elaborated in the next section.

3 The proposed SVQA

The framework of the proposed SVQA metric is shown in Fig. 2. The original and distorted stereoscopic videos' SJNDs are firstly computed. Secondly, stereoscopic saliency information is extracted from the original stereoscopic video. Next, the quality maps are calculated by the similarity of the original and distorted stereoscopic videos' SJNDs. And then, the stereoscopic saliency information is used as a weighting function in the pooling of the quality maps. Finally, the final quality score is obtained.

3.1 SJND model

According to the previous studies [8,9,13,15], four major masking effects have been validated to influence the perceptual visibility of stereoscopic videos.

(1) *Luminance adaptation* As indicated by Weber's law, human visual perception is sensitive to luminance contrast rather than absolute luminance value.
(2) *Spatial masking* The reduction in the visibility of the stimuli is induced by the increase in the spatial nonuniformity of the background luminance.
(3) *Temporal masking* The masking effect in the time domain is known as temporal masking, which has reduction peculiarity when watching a video.
(4) *Binocular masking* When dissimilar stimuli are presented in the corresponding retinal locations of the two eyes, one eye's stimulus is influenced by the other eye's.

The four masking effects are characterized by SJND [11], which consists of TJND and BPJND. TJND corresponds to the first three factors; BPJND corresponds to the factor of binocular masking. The element of TJND is a classic spatial JND model [8], in which luminance contrast and spatial masking are the two factors that determine the JND of the image. The perceptual model simplifies the complex process of estimating visibility by JND, which is defined as:

$$\mathrm{JND}(i) = \max\left\{f_1\left(\mathrm{bg}(i), \mathrm{mg}(i)\right), f_2(bg(i))\right\}, \tag{1}$$

where $f_1(\mathrm{bg}(i), \mathrm{mg}(i))$ and $f_2(\mathrm{bg}(i))$ give the spatial masking effect and the visibility threshold due to background luminance around the pixel i at (x, y), respectively. $\mathrm{bg}(i)$ and $\mathrm{mg}(i)$ are the average background luminance and the maximum weighted average of luminance differences around the pixel i, respectively.

It is generally acceptable that bigger interframe difference (caused by motion) can lead to larger temporal masking [9]. Then, TJND is defined as:

$$\mathrm{TJND}(i,t) = \max\left\{f_3\left(\mathrm{bg}(i,t), \mathrm{mg}(i,t)\right), f_4\left(bg(i,t)\right)\right\}, \tag{2}$$

where

$$f_3\left(\mathrm{bg}(i,t), \mathrm{mg}(i,t)\right) = \arg\max\left((P_t - P_{t-1}), \Delta\bar{P}\right), \tag{3}$$

$$f_4\left(\mathrm{bg}(i,t)\right) = \arg\max\left((Q_t - Q_{t-1}), \Delta\bar{Q}\right), \tag{4}$$

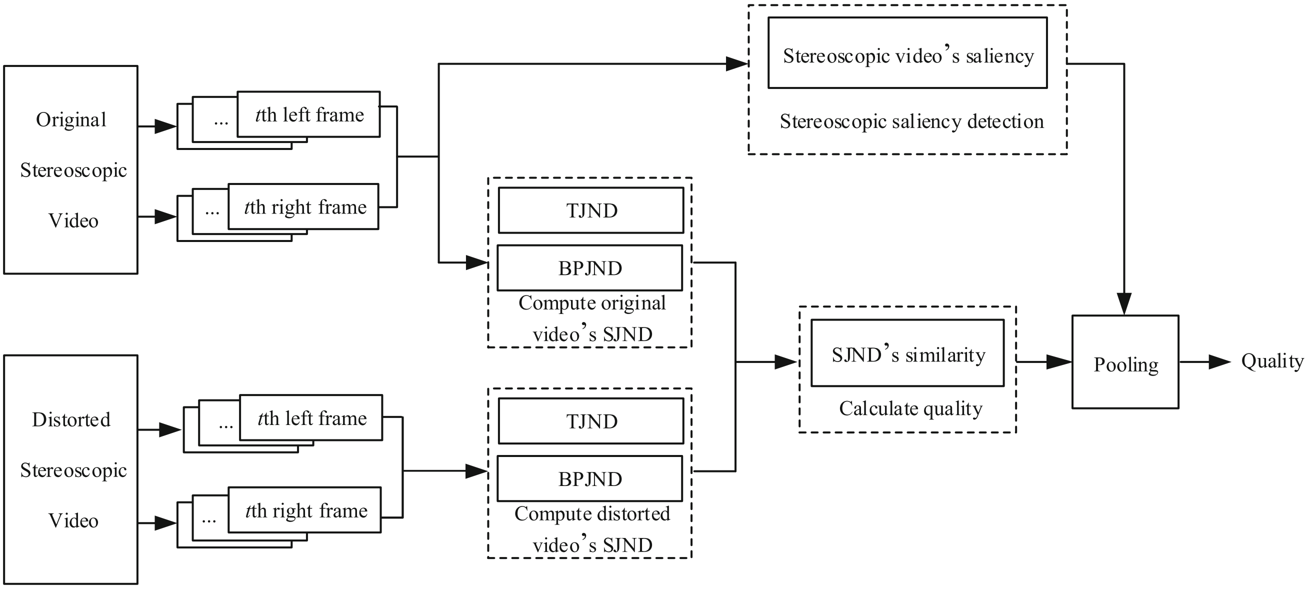

Fig. 2 The framework of the proposed SVQA metric

P_t and Q_t denote $f_1(\mathrm{bg}(i), \mathrm{mg}(i))$ and $f_2(\mathrm{bg}(i))$ of $\mathrm{JND}(i)$ at pixel i in the frame $t(t \geq 2)$, respectively. $\Delta \bar{P}$ and $\Delta \bar{Q}$ denote the mean difference between the two adjacent frames of the whole video's P and Q. For each view's video, there are the corresponding $\mathrm{TJN}D_{\mathrm{L}}(i,t)$ and $\mathrm{TJN}D_{\mathrm{R}}(i,t)$. According to the relationship of the reference (right) and auxiliary (left) views [24], TJND is defined as:

$$\mathrm{TJND}(i,t) = \frac{3}{8}\left[\mathrm{TJN}D_{\mathrm{L}}(i,t)\right] + \frac{5}{8}\left[\mathrm{TJN}D_{\mathrm{R}}(i,t)\right]. \tag{5}$$

Physiologists suggest that disparity-sensitive neurons in the visual cortex of mammals are encoded to perceive stereopsis [25]. These neural mechanisms are directly represented as binocular rivalry and binocular fusion. Binocular rivalry occurs when dissimilar monocular stimuli present to the corresponding retinal locations of the two eyes; in contrast, binocular fusion occurs in similar monocular stimuli. Therefore, we divide each view frame of stereoscopic video into rivalry stimuli and fusion stimuli, and they correspond to occlusion and nonocclusion pixels, respectively.

Occlusion pixels indicate that the pixels only present to one eye. They cannot be seen superimposed in the both eyes, and they are seen for a random moment. Based on the concept of contrast sensitivity function (CSF), only the luminance contrast is adopted in the BPJND model, which is defined as:

$$\begin{aligned}\mathrm{BPJN}D_{\mathrm{O}}(i,t) = {} & p(t) \cdot f_{3L}\left(\mathrm{bg}(i,t), \mathrm{mg}(i,t)\right) \\ & + (1-p(t)) \cdot f_{3\mathrm{R}}\left(\mathrm{bg}(i,t), \mathrm{mg}(i,t)\right),\end{aligned} \tag{6}$$

where $p(t)$ is a random number <1 which varies about the time. $f_{3\mathrm{L}}$ and $f_{3\mathrm{R}}$ are luminance differences of the interframe between the left view and the right view, respectively. Here, $p(t)$ is a sawtooth value between [0, 1].

For nonocclusion pixels, besides the two factors affecting the visibility in the spatial domain, another factor affecting the binocular visibility is the left and right view's consistency of luminance. It is expressed as:

$$\begin{aligned}\mathrm{BPJN}D_{\mathrm{N}}(i,t) = \max\{ & f_3\left(\mathrm{bg}(i,t), \mathrm{mg}(i,t)\right), \\ & f_4\left(\mathrm{bg}(i,t)\right), f_5\left(\mathrm{bg}'(i,t)\right)\},\end{aligned} \tag{7}$$

where $f_5(\mathrm{bg}'(i,t))$ represents the luminance visibility of one view's tth frame relative to the other view's.

A psychophysical experiment is conducted to verify the luminance visibility of binocular masking effect [11]. The experimental result is shown in Fig. 3. Then, the binocular masking function is approximately defined as:

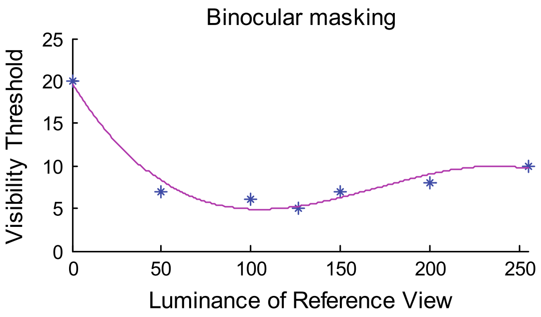

Fig. 3 Binocular masking effect

$$f_5\left(\mathrm{bg}'(i,t)\right) = \begin{cases} a \times \left(1 - \left(\mathrm{bg}'(i,t)/127^{\frac{1}{2}}\right)\right) + b, \\ \quad \text{if} \quad \mathrm{bg}'(i,t) \leq 127 \\ c \times \left(\mathrm{bg}'(i,t) - 127\right) + d, \\ \quad \text{otherwise} \end{cases} \tag{8}$$

where a denotes the visibility when the other view's gray level is 0, and c denotes the slope of the line that models the function at higher luminance of the other view. b and d are the minimum amplitudes of visibility due to binocular masking effect. A function fitting is used to parameterize the four parameters $a = 15$, $b = 5.08$, $c = 0.04$, $d = 5.08$.

Combining TJND with BPJND, SJND is defined as:

$$\mathrm{SJND}(i,t) = [\mathrm{TJND}(i,t)]^{\mu} \cdot [\mathrm{BPJND}(i,t)]^{\eta}, \tag{9}$$

where μ and η denote the weights to adjust the balance of TJND and BPJND. They will be discussed in Sect. 5.

3.2 SVA model

According to the selective processing mechanism of human visual system, we develop a novel stereoscopic visual attention model and incorporate it into the SVQA metric. The framework of the proposed SVA is shown in Fig. 4.

In monocular attention, with the histogram-based contrast algorithm [26], the pixel i's saliency value $S_{\mathrm{m}}(i)$ is computed by:

$$S_m(i) = \frac{1}{(m-1)T} \sum_{j=1}^{n} \left(T - D(c, c_j)\right) S(c_j), \tag{10}$$

where T is the sum of distances between color c and its m nearest neighbors c_j. $(T - D(c, c_j))$ assigns larger weights to colors closer to c in the color feature space. $S(c_j)$ is the image saliency value of a color value c_j.

In binocular attention, based on binocular stereopsis adaptation model that HVS separates adaptable channels for summation and difference of the neural signals to perceive

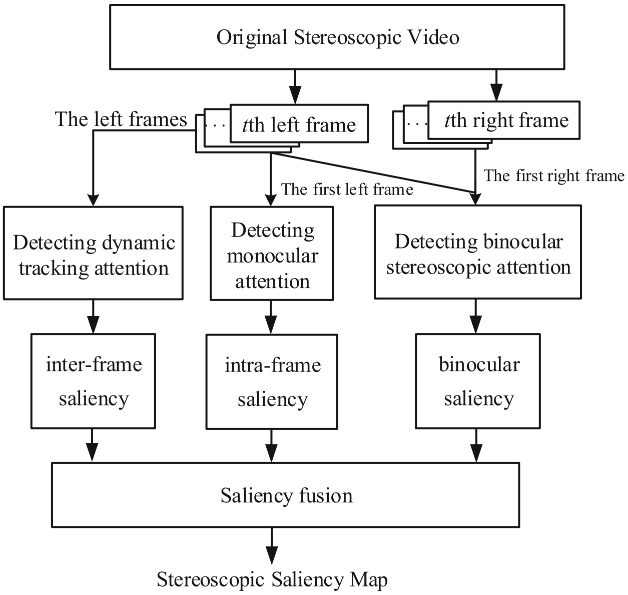
Original Stereoscopic Video
The left frames
tth left frame
tth right frame
The first right frame
The first left frame
Detecting dynamic tracking attention
Detecting monocular attention
Detecting binocular stereoscopic attention
inter-frame saliency
intra-frame saliency
binocular saliency
Saliency fusion
Stereoscopic Saliency Map

Fig. 4 The framework of the proposed SVA model

stereopsis from the two eyes [27], the two views first frame's summation and difference are computed as:

$$I_{\text{sum}} = \frac{I_{\text{L}}}{2} + \frac{I_{\text{R}}}{2}, \tag{11}$$

$$I_{\text{dif}} = |I_{\text{L}} - I_{\text{R}}| + |I_{\text{R}} - I_{\text{L}}|, \tag{12}$$

where I_{L} and I_{R} are the left and right view first frame, respectively. The feature self-resemblance algorithm [28] is exploited to extract the saliency map from I_{sum} and I_{dif}. The pixel i's saliency value $S_{b_s}(i)$ is calculated from I_{sum} by:

$$S_{b_s}(i) = \frac{1}{\sum_{j=1}^{Q} \exp\left(\frac{-1+\rho(F_k,F_l)}{\sigma^2}\right)}, \tag{13}$$

where $\rho(F_k, F_l)$ is called the matrix cosine similarity and is defined as the Frobenius inner product. Q is the number of features in the center+surrounding region, and σ is a parameter controlling the fall-off of weights. The pixel i's saliency value $S_{b_d}(i)$ is calculated from I_{dif} in the same way. Then, the saliency value $S_b(i)$ in binocular stereoscopic saliency map can be obtained as:

$$S_b(i) = \frac{1}{2} \cdot S_{b_s}(i) + \frac{1}{2} \cdot S_{b_d}(i). \tag{14}$$

In dynamic tracking attention, due to the high consistence between both views' videos, only the consistent motions of the prominent objects with enough amplitude in left view's video are popped out as the indicators of salient region. Using an optical flow function [29], the indicators are calculated as:

$$\begin{aligned} &\arg\min_{u,v,n} E\left(u, v, \hat{u}, \hat{v}\right) \\ &= \sum_{i,j} \Big\{ f_{\text{D}} \Big[\sum_{r \le n} \left(I_{\text{L}}(x, y, t) - I_{\text{L}}\left(x + r u_{x,y}, y + r v_{x,y}, t+m\right)\right) \Big] \\ &\quad + \lambda_1 \left[f_{\text{S}}(\Delta u_x) + f_{\text{S}}\left(\Delta u_y\right) + f_{\text{S}}\left(\Delta v_x\right) + f_{\text{S}}\left(\Delta v_y\right)\right]\Big\} \\ &\quad + \lambda_2 \left(\|u - \hat{u}\| + \|v - \hat{v}\|\right) + \lambda_3 f_{\omega}\left(\hat{u}, \hat{v}\right), \end{aligned} \tag{15}$$

where u, v are the horizontal and vertical components of the optical flow field to be estimated from the tth left frame $I_{\text{L}}(t)$ and the $(t+m)$th left frame $I_{\text{L}}(t+m)$. $\hat{u}, \hat{v}$ denote an auxiliary flow field. f_{D} is the brightness constancy constraint function, and f_{S} is the smooth penalty function. λ_1 is a regularization parameter. λ_2, λ_3 are scalar weights. f_{ω} is the state similarity function. Then, the interframe's saliency is expressed as:

$$S_t(i) = N\left(\sqrt{u_i^2 + v_i^2}\right), \tag{16}$$

where $N(\cdot)$ is a normalized function.

In the saliency map fusion, the weights of the three saliency maps are determined by the distribution of salient pixels in each saliency map. If the salient pixels of the saliency map converge at one region, a larger weight will be set for this saliency map. While the salient pixels disperse among the saliency map, a smaller weight will be assigned. The intraframe, interframe and binocular saliency maps are fused as:

$$S_F = w_m \cdot S_m + w_t \cdot S_t + w_b \cdot S_b, \tag{17}$$

where w_m, w_t, w_b are the normalized weights of monocular saliency map w'_m, binocular saliency map w'_b and temporal saliency map w'_t. The w'_m is calculated as:

$$w'_m = \frac{1}{\mathrm{e}^{\frac{\|S_m\|_0}{W \times H}}}, \tag{18}$$

where $\|S_m\|_0$ is ℓ_0 norm, which counts the nonzero entries of the salient pixels in saliency map S_m. W, H are the width and height of the saliency map S_m , respectively. Similarly, the w'_b and w'_t are calculated in the same way. Then, the normalized weighting of w'_m is expressed as:

$$w_m = \frac{w'_m}{w'_m + w'_b + w'_t}, \tag{19}$$

w_t, w_b are calculated in the same way.

3.3 Quality calculation

The quality maps of the original and distorted SJND maps are calculated as:

Fig. 5 The first frames of nine sequences in the subjective test

$$q(i,t) = \frac{2 \cdot \mathrm{SJN}D_{\mathrm{o}}(i,t) \cdot \mathrm{SJN}D_{\mathrm{d}}(i,t) + \varepsilon}{\mathrm{SJND}_{\mathrm{o}}^2(i,t) + \mathrm{SJND}_{\mathrm{d}}^2(i,t) + \varepsilon}, \tag{20}$$

where $\mathrm{SJN}D_{\mathrm{o}}(i,t)$ and $\mathrm{SJN}D_{\mathrm{d}}(i,t)$ denote tth frame's SJND map values at pixel i of the original and the distorted stereoscopic videos, respectively, and here we take $\varepsilon = 0.1$ empirically. Since degradations in the nonsaliency regions still affect subjects' evaluation, we adopt different weights in saliency and nonsaliency regions to pool the quality maps as a quality score:

$$Q = \sum_{i,t} q(i,t) \times [w_{3\mathrm{D}} \cdot S_F + (1 - w_{3\mathrm{D}}) \cdot (1 - S_F)], \tag{21}$$

where $w_{3\mathrm{D}}$ is the weight of salient region in our SVQA model and is discussed in Sect. 5.

4 Subjective experiment

To the best of our knowledge, there is only one public database [30] in the studies of 3D video quality assessment. To better evaluate the proposed SVQA metric, a subjective experiment is conducted to construct a ground-truth database. Nine stereoscopic videos are chosen to establish our database (Fig. 5). The subjective test setting is shown in Table 1. For more details, please visit our website [31].

5 Experimental results

5.1 Parameter optimization

To determine the three parameters in the proposed metric, we compare its performances at the different parameters. Four evaluation criteria are chosen in the performance evaluation, e.g., PLCC, SROCC, KRCC and RMSE. In Eq. (9), μ, η denote the weights to adjust the balance of TJND and BPJND, there exists $\mu + \eta = 1$. In Eq. (21), $w_{3\mathrm{D}}$ is the weighting of saliency, there exists $w_{3\mathrm{D}} \in [0.5, 1]$. Here, we firstly fix $\mu = \eta = 0.5$ and compare the performance of different $w_{3\mathrm{D}}$ in Table 2. The criterion that achieves the best performance is highlighted in bold.

From Table 2, the metric achieves the best performance when $w_{3\mathrm{D}} = 0.7$. Then, we fix $w_{3\mathrm{D}} = 0.7$, and compare the performance of different μ, η in Table 3.

Table 1 Subjective test setting

Stereoscopic video encoder	JMVM 2.1
QP	0, 20, 30, 40, 50
Distortion type	GaussBlur (OpenCV)
Sigma	0, 1, 3, 5, 7
Frame rate	25 fps
Display	ViewSonic VX2268wm
Display resolution	1680 × 1050
Refresh rate	120 Hz
Glasses	Nvidia 3D vision shutter glasses
Glasses refresh rate	60 Hz
Subjective test standard	ITU-R BT.500-11
Test method	SSCQE
Observers	18
Age range	20–35
Viewing distance	1 m

Table 2 Performance of different $w_{3\mathrm{D}}$

$w_{3\mathrm{D}}$	PLCC	SROCC	KRCC	RMSE
0.5	0.8289	0.8271	0.6489	0.5512
0.6	0.8353	0.8352	0.6558	0.5453
0.7	**0.8378**	**0.8356**	**0.6565**	**0.5429**
0.8	0.8319	0.8321	0.6542	0.5486
0.9	0.8112	0.8101	0.6336	0.5677
1	0.7915	0.7918	0.6169	0.5849

Table 3 Performance of different μ, η

μ, η	PLCC	SROCC	KRCC	RMSE
(0.1, 0.9)	0.5926	0.5795	0.4190	0.8010
(0.2, 0.8)	0.6336	0.6219	0.4573	0.7693
(0.3, 0.7)	0.7648	0.7695	0.5913	0.6407
(0.4, 0.6)	0.8240	0.8177	0.6432	0.5634
(0.5, 0.5)	0.8378	0.8356	0.6565	0.5429
(0.6, 0.4)	**0.8415**	**0.8379**	**0.6650**	**0.5372**
(0.7, 0.3)	0.8370	0.8348	0.6558	0.5434
(0.8, 0.2)	0.8187	0.8144	0.6391	0.5710
(0.9, 0.1)	0.6948	0.6962	0.5248	0.7152

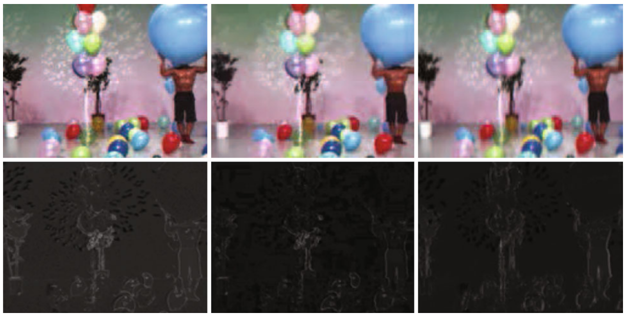

Fig. 6 Comparison of SJND maps. Images in the *first line* are the first frames of left view's sequence with original, and distortion by H.264 (QP = 50), GaussBlur (sigma = 7). Images in the *second line* are the corresponding SJND maps

From Table 3, the metric achieves the best performance when $\mu = 0.6$, $\eta = 0.4$.

5.2 Performance of SJND

Figure 6 shows the comparison of SJND maps of *Balloons* with original, and distortion by H.264, GaussBlur. From the second and the third SJND maps in the last line of Fig. 6, it can be found that block artifacts are existed in flat regions and texture details are lost in edge regions, respectively.

If we close the SVA model, the performance of the proposed SVQA metric only using SJND model is listed in the first line of Table 2. In comparison with the best performance in Table 3, the performance of the proposed metric only using SJND model decreases slightly.

5.3 Performance of proposed SVA model

Since no public 3D video eye-tracking database is provided for evaluation, we choose the 3D image eye-tracking database [22] for performance evaluation. Three criteria are used to measure the similarity between the fixation density maps and the computed saliency maps, e.g., PLCC, KLD and AUC. Note that higher PLCC, AUC and lower KLD score mean a better performance. Here, we attempt to develop two state-of-the-art SVA models [20,21] for comparison. The interframe's saliency of the three models is all use optical flow algorithm. If the performance contributions of interframe's saliency in these three models are same, the performance of SVA model depends on the intraframe's and binocular saliency. The comparison results without interframe's saliency are listed in Table 4.

Table 4 Comparison with three stereoscopic saliency models

SVA model	PLCC	KLD	AUC
Zhang et al. model	0.158	0.759	0.567
Dittrich et al. model	0.342	0.552	0.619
Proposed	**0.389**	**0.474**	**0.657**

5.4 Performance of different distortion types

We also performed the proposed SVQA metric on the distorted stereoscopic videos with different distortion types. Table 5 lists the performance results on NAMA3DS database [30].

Table 5 Performance of different distortion types

Distortion type	PLCC	SROCC	KRCC	RMSE
H.264	0.5834	0.6810	0.4890	0.6672
JPEG2000	0.8062	0.6901	0.5029	0.5079
Downsampling and sharpening	0.6153	0.5071	0.4247	0.7209

It can be seen that the proposed metric has better prediction performance for JPEG2000 than the other two distortion types.

5.5 Comparison with state-of-the-art metrics

We choose four representative metrics to compare, including PQM [4], PHVS-3D [5], SFD [6] and 3D-STS [7]. Note that 3D-STS is the state-of-the-art metric in SVQA. The performance comparison results on our database are listed in Table 6.

As shown in Table 6, the proposed and 3D-STS metric achieve the better performance than the other three metrics.

We also performed the experiments on NAMA3DS database. Table 7 provides the performance comparison results on the NAMA3DS database.

Table 6 Performance comparison of SVQA metrics on our database

Metrics	PLCC	SROCC	KRCC	RMSE
PQM	0.7852	0.8165	0.6365	0.6158
PHVS-3D	0.7082	0.7195	0.5353	0.7021
SFD	0.6483	0.6633	0.5021	0.7571
3D-STS	0.8311	0.8338	0.6553	0.5520
Proposed	**0.8415**	**0.8379**	**0.6650**	**0.5372**

Table 7 Performance comparison of SVQA metrics on NAMA3DS database

Metrics	PLCC	SROCC	KRCC	RMSE
PQM	0.6340	0.6006	0.4391	0.8784
PHVS-3D	0.5480	0.5146	0.3572	0.9501
SFD	0.5965	0.5896	0.4025	0.9117
3D-STS	0.6417	0.6214	0.4544	0.9067
Proposed	**0.6503**	**0.6229**	**0.4575**	**0.8629**

It can be seen from Table 7 that the performance results of PQM, 3D-STS and the proposed metric are similar. Although the stereoscopic video contents, the display systems, the subjects and the subjective experiments are different in the two databases, the proposed metric can still work well.

6 Conclusion

This paper proposed a novel SVQA metric in three stages, e.g., visual perception, visual attention and estimation. To evaluate the proposed metric, a subjective test is conducted to construct a ground-truth database. The experimental result shows that the proposed SVQA metric has a competitive performance comparing with the existing SVQA metrics.

Acknowledgments This work has been supported in part by the Major State Basic Research Development Program of China (973 Program 2015CB351804), the National Science Foundation of China under Grant (Nos. 61472101, 61272386).

References

1. Ha, K., Kim, M.: A perceptual quality assessment metric using temporal complexity and disparity information for stereoscopic video. In: Proceedings of the ICIP, pp. 2525–2528 (2011)
2. Bensalma, R., Larabi, M.C.: A perceptual metric for stereoscopic image quality assessment based on the binocular energy. Multidimens. Syst. Signal Process. **24**(2), 281–316 (2013)
3. Shao, F., Lin, W., Gu, S., Jiang, G., Srikanthan, T.: Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics. IEEE Trans. Image Process. **22**(5), 1940–1953 (2013)
4. Joveluro, P., Malekmohamadi, H., Fernando, W.A.C., Kondoz, A.M.: Perceptual video quality metric for 3d video quality assessment. In: Proceedings of the 3DTV-CON, pp. 1–4 (2010)
5. Jin, L., Boev, A., Gotchev, A., Egiazarian, K.: 3D-DCT based perceptual quality assessment of stereo vide. In: Proceedings of the ICIP, pp. 2521–2524 (2011)
6. Lu, F., Wang, H., Ji, X., Er, G.: Quality assessment of 3D asymmetric view coding using spatial frequency dominance model. In: Proceedings of the 3DTV-CON, pp. 1–4 (2009)
7. Han, J., Jiang, T., Ma, S.: Stereoscopic video quality assessment model based on spatial–temporal structural information. In: Proceedings of the VCIP, pp. 119–125 (2012)
8. Chou, C.-H., Li, Y.-C.: A perceptually tuned sub-band image coder based on the measure of just-noticeable-distortion profile. IEEE Trans. Circuits Syst. Video Technol. **5**(6), 467–476 (1995)
9. Yang, X., Lin, W., Lu, Z., Ong, E.P., Yao, S.: Motion-compensated residue preprocessing in video coding based on just-noticeable-distortion profile. IEEE Trans. Circuits Syst. Video Technol. **15**(6), 742–752 (2005)
10. Zhang, X., Lin, W., Xue, P.: Just-noticeable difference estimation with pixels in images. J. Vis. Commun. Image Represent. **19**, 30–41 (2008)
11. Qi, F., Jiang, T., Fan, X., Ma, S., Zhao, D.: Stereoscopic video quality assessment based on stereo just-noticeable difference model. In: Proceedings of the ICIP, pp. 34–38 (2013)
12. Lin, W., Jay Kuo, C.-C.: Perceptual visual quality metrics: a survey. J. Vis. Commun. Image Represent. **22**(4), 297–312 (2011)
13. Zhao, Y., Chen, Z., Zhu, C., Tan, Y., Yu, L.: Binocular just-noticeable-difference model for stereoscopic images. IEEE Signal Process. Lett. **18**(1), 19–22 (2011)
14. De. Silva, D., Fernando, W.A.C., Worrall, S.T., Yasakethu, S.L.P., Kondoz, A.M.: Just noticeable difference in depth model for stereoscopic 3D displays. In: Proceedings of the ICME, pp. 1219–1224 (2010)
15. Li, X., Wang, Y., Zhao, D., Jiang, T., Zhang, N.: Joint just noticeable difference model based on depth perception for stereoscopic images. In: Proceedings of the VCIP, pp. 1–4 (2011)
16. Zhai, G., Wu, X., Yang, X., Lin, W., Zhang, W.: A psychovisual quality metric in free-energy principle. IEEE Trans. Image Process. **21**(1), 41–52 (2012)
17. Friston, K.: The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. **11**(2), 127–138 (2010)
18. Howard, I.P., Rogers, B.J.: Binocular Vision and Stereopsis. Oxford University Press, Oxford (1995)
19. Wang, Z., Li, Q.: Information content weighting for perceptual image quality assessment. IEEE Trans. Image Process. **20**(5), 1185–1198 (2011)
20. Zhang, Y., Jiang, G., Yu, M., Chen, K.: Stereoscopic visual attention model for 3-D video. In: Proceedings of the Multimedia Modeling, pp. 314–324 (2010)
21. Dittrich, T., Kopf, S., Schaber, P., Guthier, B., Effelsberg, W.: Saliency detection for stereoscopic video. In: Proceedings of the 4th ACM Conference on Multimedia Systems, pp. 12–23 (2013)
22. Wang, J., Perreira, M., Silva, D., Callet, P.L., Ricordel, V.: A computational model of stereoscopic 3D visual saliency. IEEE Trans. Image Process. **22**(6), 2151–2165 (2013)
23. Zhang, L., Shen, Y., Li, H.Y.: VSI: a visual saliency induced index for perceptual image quality assessment. IEEE Trans. Image Process. **23**(10), 4270–4281 (2014)
24. Aflaki, P., Hannuksela, M.M., Hakkinen, J., Lindroos, P., Gabbouj, M.: Subjective study on compressed asymmetric stereoscopic video. In: Proceedings of the ICIP, pp. 4021–4024 (2010)
25. Fleet, D.J., Wagner, H., Heeger, D.J.: Neural encoding of binocular disparity: energy models, position shifts and phase shifts. J. Vis. Res. **36**(12), 1839–1857 (1996)
26. Cheng, M.M., Zhang, G.X., Mitra, N.J., Huang, X., Hu, S.M.: Global contrast based salient region detection. In: Proceedings of the CVPR, pp. 409–416 (2011)
27. May, K.A., Li, Z.P., Hibbard, P.B.: Perceived direction of motion determined by adaptation to static binocular images. Curr. Biol. **22**(1), 28–32 (2012)
28. Seo, H.J., Milanfar, P.: Visual saliency for automatic target detection, boundary detection, and image quality assessment. In: Proceedings of the ICASSP (2010)
29. Zhong, S.H., Liu, Y., Ren, F.F., Zhang, J.H., Ren, T.W.: Video saliency detection via dynamic consistent spatio-temporal attention modelling. In: Proceedings of the AAAI Conference on Artificial Intelligence (2013)
30. Urvoy, M., Gutirrez, J., Barkowsky, M., Cousseau, R., Koudota, Y., Ricordel, V., Callet, P.L., Garca, N.: NAMA3DS1-COSPAD1: subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences. In: Fourth International Workshop on Quality of Multimedia Experience (2012)
31. Qi, F.: The Illumination of SVQA Subjective Test. http://www.escience.cn/people/qifeng/index.html