3D Res (2016) 7:17
DOI 10.1007/s13319-016-0079-6

3DR EXPRESS

Using Saliency-Weighted Disparity Statistics for Objective Visual Comfort Assessment of Stereoscopic Images

Wenlan Zhang · Ting Luo · Gangyi Jiang · Qiuping Jiang · Hongwei Ying · Jing Lu

Received: 24 November 2015 / Revised: 31 December 2015 / Accepted: 3 January 2016 / Published online: 26 May 2016

Abstract Visual comfort assessment (VCA) for stereoscopic images is a particularly significant yet challenging task in 3D quality of experience research field. Although the subjective assessment given by human observers is known as the most reliable way to evaluate the experienced visual discomfort, it is time-consuming and non-systematic. Therefore, it is of great importance to develop objective VCA approaches that can faithfully predict the degree of visual discomfort as human beings do. In this paper, a novel two-stage objective VCA framework is proposed. The main contribution of this study is that the important visual attention mechanism of human visual system is incorporated for visual comfort-aware feature extraction. Specifically, in the first stage, we first construct an adaptive 3D visual saliency detection model to derive saliency map of a stereoscopic image, and then a set of saliency-weighted disparity statistics are computed and combined to form a single feature vector to represent a stereoscopic image in terms of visual comfort. In the second stage, a high dimensional feature vector is fused into a single visual comfort score by performing random forest algorithm. Experimental results on two benchmark databases confirm the superior performance of the proposed approach.

Keywords Quality of experience (QoE) · Three-dimensional (3D) · Visual comfort assessment (VCA) · 3D Visual saliency · Random forest (RF)

W. Zhang · T. Luo (✉) · J. Lu
College of Science and Technology, Ningbo University, Ningbo 315211, China
e-mail: luoting@nbu.edu.cn

T. Luo · G. Jiang · Q. Jiang · H. Ying
Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China

1 Introduction

Recent decades have witnessed a booming development of data transmission and display technologies and users' demand for video services with high quality of experience (QoE) has become considerably urgent. With the additional depth sensation provided by stereoscopic three-dimensional (3D) visual media, a growing body of attention has been drawn to advanced 3D videos due to the enhanced viewing experience to viewers [1–4]. However, increasing complaints on the experienced visual discomfort (also termed as visual fatigue in some other literatures) have also become the focus that is extensively concerned by the researchers in both industrial and academic communities [1–4]. Especially, it is of great significance to evaluate the degree of experienced visual discomfort when viewing stereoscopic images. As known, the most reliable way to measure visual discomfort is the subjective assessment conducted by human observers (the ultimate receiver of 3D contents). However, subjective

assessment is always labor-consuming and non-systematic. Therefore, how to develop objective visual comfort assessment (VCA) approaches that can automatically predict the degree of experienced visual comfort is both meaningful and desirable.

In the literature, it has been discovered that factors including binocular disparity, conflict between accommodation and vergence, binocular mismatch, spatial frequency, crosstalk, and depth motion are all relevant with visual discomfort [5–7]. Especially, the excessive binocular horizontal disparity and unnatural accommodation-convergence conflict are identified as the most influential ones among these factors. The position shift between two projected retinal images, which is known as binocular horizontal disparity, contributes largely to the stereoscopic depth perception. That is, binocular horizontal disparity provides a direct depth clue that modifies the visual perception of the immediate 3D environment by inducing convergence movements, which are deeply related to visual discomfort.

Based on this principle, several disparity statistics-based VCA approaches have been proposed over the past several decades. In [2], mean and range of disparities were calculated from an entire image frame in a video sequence for VCA. The range of disparities was computed by the difference between the maximum and minimum 10 % of disparity magnitudes over all pixels in an entire image. Yano et al. [8] computed the ratio of absolute disparity magnitude summation between the regions near and far from the screen for VCA. Kim et al. [9] calculated the horizontal and vertical disparity as the predictive features to estimate the degree of visual comfort. Choi et al. [10] presented a VCA model for 3D video by computing spatial and temporal complexity of depth image as the predictive features. Recently, Jiang et al. [11] learned a preference learning model for VCA. Three types of features including zone of comfort, depth of focus, and spatial frequency are extracted from disparity map to represent a stereoscopic image in terms of visual discomfort. One point should be emphasized is that the visual comfort-aware features of all the above-mentioned schemes are derived from a global perspective. The global perspective corresponds to that each pixel or location in a scene is treated equally for global feature descriptor construction, which ignores the important visual attention mechanism of human visual system (HVS). It has been well discovered that the HVS can exhibit high selectivity towards raw input visual signals, implying that different pixels or locations are of different visual sensitivity and perceptual importance values [12].

Inspired by the visual attention mechanism of HVS, there have been some early attempts to improve the performance of the traditional global-based VCA schemes by considering this important visual property. For instance, Sohn et al. [13] extracted salient object-dependent disparity characteristics to predict the visual comfort of stereoscopic images. The relative disparity values between adjacent objects and the stimulus size of foreground object are extracted as the salient object-dependent features. Lee et al. [14] proposed a VCA model by combining the width of foreground object with the disparity statistics from the observation that smaller stimulus width tended to induce larger visual discomfort. Yong et al. [15] derived an objective VCA model by taking human visual attention into account. The used 3D visual saliency map is obtained by linearly combining the 2D saliency map and normalized depth map with equal weights. The most significant problem is that whether the assigned equal weights to 2D saliency and depth maps are rational and optimal to reflect the human attention fixation distribution under 3D viewing condition.

In this paper, we propose a new VCA approach using saliency-weighted disparity statistics fused with random forest (RF). The main contributions of this paper are threefold: (1) instead of extracting the features from a global perspective, we extract saliency-weighted disparity statistics (i.e., disparity magnitude, disparity contrast, disparity dispersion, and disparity skewness) as the predictive features for VCA; (2) we propose an adaptive 3D visual saliency detection model to derive pixel-wise saliency maps. In particular, the 2D saliency and depth maps are fused with adaptive weights, which are determined by the entropy of depth map; (3) RF [16] is applied to learn a visual comfort predictor, providing the latent mapping from high dimensional feature space to low dimensional quality score space. The organization structure of this paper is sketched as follows. In Sect. 2, an overview of the proposed approach is presented. Section 3 illustrates the details of the proposed approach. Experimental results and analyses are presented in Sect. 4. Finally, we draw the conclusions in Sect. 5.

2 Overview of the Proposed Approach

The problem of VCA is similar to the recent focused blind image quality assessment (BIQA) which aims to automatically measure the perceptual quality of distorted images without the reference for comparison [17]. Generally speaking, mainstream of the learning-based BIQA approaches work in the following three steps:

(a) Quality-aware feature extraction;
(b) Model training with machine learning (ML) algorithms;
(c) Model testing on new samples (training and testing samples are independent).

Similar with the mainstream of learning-based BIQA framework, the proposed VCA approach is also composed of feature representation, model training, and model testing modules. In particular, for feature representation, 3D saliency-weighted disparity statistics are extracted to represent the visual discomfort of a stereoscopic image. Especially, the 3D saliency map is derived by fusing 2D saliency and depth maps with adaptive weights, which are determined by the entropy of depth map. Then, we apply the RF algorithm to learn a regression model from multi-dimensional statistical features to a single visual comfort score. RF is shown to be an effective tool to learn the relationship between visual comfort-aware feature vectors and quality score. Finally, the learned regression model is tested on new samples for performance benchmarking. Note that the samples used for training and testing are rigorously independent. Figure 1 presents the overall flowchart of our proposed three-stage VCA framework. Obviously, the key modules in the framework lie in how to derive the 3D saliency map and to extract visual-comfort features. We will elaborate these procedures in the next section.

3 The Proposed VCA Approach

3.1 3D Saliency Detection Model

Human visual system (HVS) employs an attentional mechanism to perceive the raw visual signals by allocating limited visual computational resources to those perceptual important regions. The perceptual important regions are termed as salient regions which are usually different from their surrounding regions in terms of low-level attributes such as intensity, color, texture and orientation, etc. In order to better understand where human looks, there have been many studies devoted to predicting human fixations under free-viewing conditions. As the most influential work, Itti et al. [18] proposed a well-known saliency model, which first computes feature maps of luminance, color and orientation using a center-surround operator across different scales, and then performs normalization and summation to generate the saliency map. Salient regions showing high local contrast with their surrounding regions in terms of any of the three

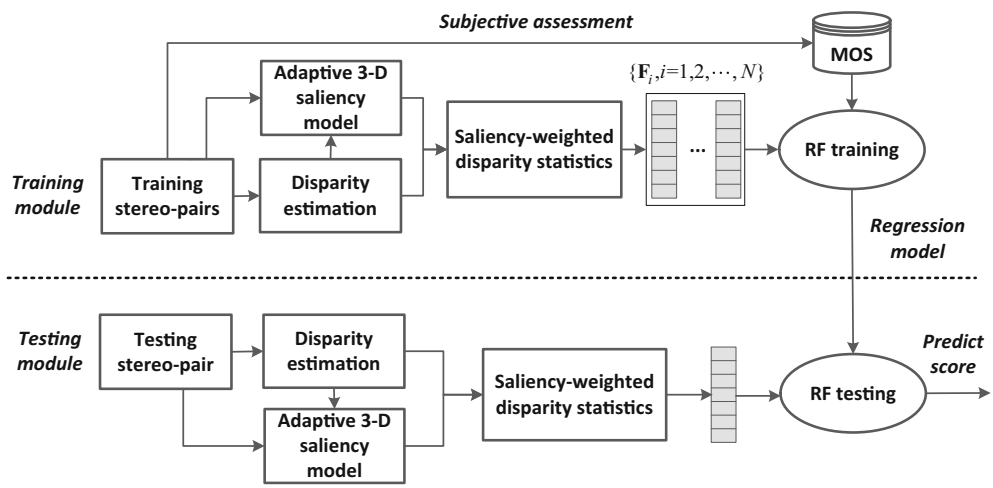

Fig. 1 The framework of the proposed VCA approach

features are highlighted in the saliency map. Based on this milestone work, many relevant saliency computational models were successively proposed based on the center-surround mechanism while implemented using a variety of features including local contrasts of color, texture and shape features, oriented sub-band decomposition based energy, ordinal signatures of edge and color orientation histograms. Later, Harel et al. [19] developed a graph-based visual saliency (GBVS) model which extended Itti's model by using a more accurate measure of dissimilarity. In [20], Goferman et al. proposed a context-aware saliency detection model that can extract salient regions to represent a scene. They claimed that people tend to describe the scene rather than the single salient object. Hou et al. [21] introduced a simple image descriptor referred to as the image signature for human fixation prediction. They demonstrated that image signature preferentially contains information about the foreground of an image-a property which is useful for detecting salient image regions. Li et al. [22] proposed a new bottom-up paradigm for detecting visual saliency, characterized by a scale-space analysis of the amplitude spectrum of natural images. Most recently, Martinel et al. [23] introduced a saliency computation approach named Kernelized Graph-Based Visual Saliency (KGBVS) that extends the standard GBVS algorithm by using different kernels in the computation of transition probabilities. For a more comprehensive overview on this field, researchers can refer to [12].

Note that the above saliency models are all proposed for 2D images, which may not completely suitable for cases of 3D visual saliency detection. Compared with various saliency detection models proposed for 2D images, how to understand the role of additional depth information on the deployment of human fixations under 3D viewing condition is of great importance. Previous studies have found that both low-level appearance features (e.g., intensity, color, texture and orientation) and depth cues are related with human fixation behaviors when viewing stereoscopic images [24]. We apply off-the-shelf 2D saliency models to estimate 2D saliency maps to reflect the effect of low-level appearance features to human fixation behaviors. For depth information, we use the binocular disparity map for depth-aware saliency estimation since disparity map provide a direct cue to create depth sensation in current stereoscopic imaging system. On the basis of the common sense that objects near the observer will achieve more attention so as to induce more severe visual discomfort compared with objects far away from the observer, we formulate the 3D visual saliency as

$$S_{3D} = (1-\alpha) \cdot S_{2D} + \alpha \cdot S_{Depth} \quad (1)$$

where S_{2D} is a standard GBVS map and S_{Depth} is a normalized disparity map. The parameter α is an adaptive value that controls the relative importance between 2D saliency and depth saliency.

By analyzing the ground truth fixation data of stereoscopic images with different depth structures, we have the following observations. First, scenes with obvious salient object tend to have relatively compact fixation distribution while scenes with non-salient object tend to have relatively dispersive fixation distribution. Second, the fixation data of salient object scenes tend to focus on the salient object while the fixation data of non-salient object scenes tend to spread over the whole image. That is to say, depth information will have different impacts on human fixation distribution of stereoscopic images with different depth structures. Examples of scenes with salient object and non-salient object along with their corresponding fixation density maps can be found in Fig. 2.

Inspired by this observation, we intuitively think that the determination of parameter α should be adaptive with depth structure so as to better characterize the influence of depth saliency. In particular, we propose to compute the entropy of disparity map to reflect the deployment of depth structure, assuming that scenes with complex depth structure will have lower disparity entropy values and vice versa. To measure the entropy of a disparity map, we first quantize the gray-scale into K bins in the range of [0,255], and then compute the entropy as Eq. (2).

$$E = -\sum_{l=1}^{k} p[d(l)] \, ln \, (p[d(l)]) \quad (2)$$

where $d(l)$ represents the disparity value of the l-th bin and $p[d(l)]$ is the probability of the l-th bin. Since we have known that a disparity map with uniform distribution in terms of the depth gray-scale will have the maximum entropy value, the maximal entropy value is a constant which can be denoted by E_{max}. We

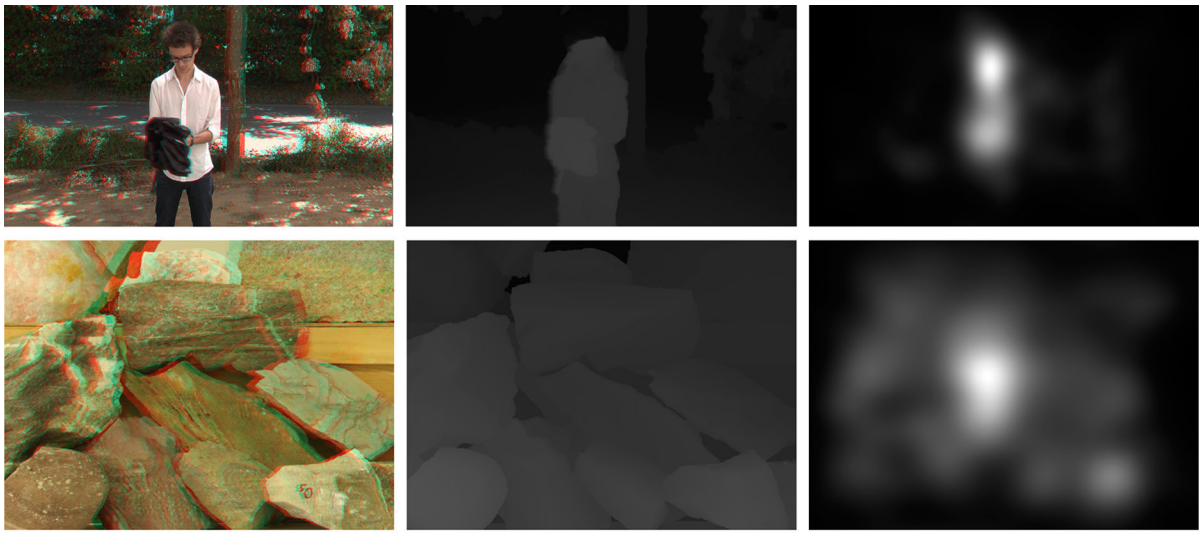

Fig. 2 Examples of stereoscopic images (in *red-green* anaglyph formats) with salient object and non-salient object along with their corresponding fixation density maps. The first row shows the scene with salient object while the second row shows the scene with non-salient object

assume that parameter α is linearly proportional to the disparity entropy value. Based on this assumption, we compute the ratio of E and E_{max} to adaptively adjust parameter α

$$\alpha = \frac{E}{E_{\max}} \tag{3}$$

It is worthy emphasizing that, by using the adaptive parameter to combine 2D saliency and depth saliency, the relative importance between them is well characterized. Examples of some stereoscopic images associated with their 3D saliency maps are shown in Fig. 3. As compared with the ground truth fixation density maps, the estimated saliency maps can well predict human fixations due to the consideration of adaptive weights between 2D saliency and depth saliency.

3.2 Saliency-Weighted Disparity Statistics

This section involves extracting visual comfort-aware statistics from disparity maps by using the previously estimated 3D saliency maps as weights. The used disparity statistics include disparity magnitude, disparity contrast, disparity dispersion, and disparity skewness, which have been demonstrated to be deeply relevant with visual comfort [25]. Specifically, given a stereoscopic image $I_{3D}(x, y)=\{I_L(x, y), I_R(x, y)\}$ and its corresponding disparity map, $d(x, y)$ we first compute its 3D saliency map $S_{3D}(x, y)$ by using the above described adaptive 3D saliency detection model, then the related saliency-weighted disparity statistics are given by

(a) 3D saliency-weighted disparity magnitude:

$$f_1 = \frac{1}{d_m} \cdot \left(\sum_{i=1}^{M} \sum_{j=1}^{N} S_{3D}(i,j) \cdot |d(i,j)| \right) \Bigg/ \left(\sum_{i=1}^{M} \sum_{j=1}^{N} S_{3D}(i,j) \right) \tag{4}$$

(b) 3D saliency -weighted disparity contrast:

$$f_2 = \frac{1}{d_m} \cdot \left(\sum_{i=1}^{M} \sum_{j=1}^{N} S_{S3D}(i,j) \cdot |d_c(i,j)| \right) \Bigg/ \left(\sum_{i=1}^{M} \sum_{j=1}^{N} S_{S3D}(i,j) \right) \tag{5}$$

(c) 3D saliency-weighted disparity dispersion:

$$f_3 = \frac{1}{d_m} \cdot \sqrt{\left(\sum_{i=1}^{M}\sum_{i=1}^{M} S_{3D}(i,\, j) \cdot d(i,\, j)^2\right) \Big/ \left(\sum_{i=1}^{M}\sum_{i=1}^{M} S_{3D}(i,\, j)\right)} \tag{6}$$

(d) 3D saliency-weighted disparity skewness:

$$f_4 = \left(\sum_{i=1}^{M}\sum_{j=1}^{N} S_{3D}(i,\, j) \cdot d(i,\, j)\right) \Big/ \left(\sum_{i=1}^{M}\sum_{j=1}^{N} S_{3D}(i,\, j) \cdot d(i,\, j)\right) \tag{7}$$

where $\{d_c(x, y)\}$ is the disparity contrast map calculated by using center-surrounding operator, M and N are the width and height of $d(x, y)$, respectively, and d_m is the maximum disparity magnitude as a normalized factor.

In addition, we also utilize the fact that excessive binocular disparity magnitude tends to induce visual discomfort. In general, stereoscopic images with even a small amount of excessive binocular disparities may still be perceived as uncomfortable, which motivates us to take the percentages of maximum and minimum disparity values into account for VCA. The average disparity values of the maximum and minimum p % disparity values are given by

(e) the average value of the maximum p % disparities:

$$f_5 = \frac{1}{d_m} \cdot \left(\frac{1}{N(\Omega_p^+)} \sum_{(i,\, j) \in \Omega_p^+} d(i,\, j)\right) \tag{8}$$

(f) the average value of the minimum p % disparities:

$$f_6 = \frac{1}{d_m} \cdot \left(\frac{1}{N(\Omega_p^-)} \sum_{(i,\, j) \in \Omega_p^-} d\,(i,\, j)\right) \tag{9}$$

where Ω_p^+ and Ω_p^- represent the sets of pixels whose disparities belong to the maximum and minimum p % disparities over all pixels in $d\,(x, y)$, $N(\Omega_p^+)$ and $N(\Omega_p^-)$ are the number of pixels in Ω_p^+ and Ω_p^-, respectively. In our experiment, the number of p % is empirically set to 5 %.

As a result, a 6-dimensional feature vector can be obtained by combining all the 3D saliency-weighted disparity statistics: $\mathbf{F}_\mathrm{p} = [f_1, f_2, f_3, f_4, f_5, f_6]$.

3.3 Learning with Random Forest (RF)

In order to predict a single visual comfort score for a stereoscopic image, we learn a regression model from a set of training samples. Through feature extraction, each stereoscopic image can be represented as a 6-dimensional feature vector. Given a training set $\Omega_a = \{I_1, I_2, \ldots, I_N\}$, a set of visual comfort-aware feature vectors $F_{\Omega_a} = \{\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_N\}$ can be acquired. Then, with F_{Ω_a} and their corresponding subjective scores, visual comfort regression model $\phi\,(\cdot)$ is constructed by using RF algorithm. At the testing stage, given a to-be-assessed stereoscopic image, by extracting the feature vectors, i.e., $\mathbf{F_t}$, and feeding them into the learned regression model, the quality scores are predicted by: $Q_t = \phi\,(\mathbf{F}_t)$.

Here, of course, other machine learning (ML) algorithms also can be adopted. However, the motivations that we choose the RF algorithm as the regression model in our approach are two folds. First, it has been demonstrated that the RF algorithm is the best one among 179 ML algorithms arising from 17 families when testing on 121 datasets. Although the difference is not statistically significant with the second best, i.e., the SVM with Gaussian kernel, we experimentally found that the RF algorithm is slightly better than support vector regression in our approach. Second, as we all known, in subjective VCA experiments, the ultimate visual comfort score is obtained by averaging the evaluation from different subjects. Inspired by this fact, we adopt the RF algorithm, the training and testing procedures of which are similar to the subjective VCA process, as the regression model in our approach. Actually, this method combines the "bagging" theory and the random selection of features. In the implementation of RF algorithm, there are two most important hyperparameters namely the number of trees *ntree* and the number of variables to split on at each node *mtry*. We set $ntree = 1000$ and $mtry = 2$ in our experiment. In addition, the impact of different RF hyperparameters will be presented and analyzed in Sect. 4.4.3.

Fig. 3 Examples of stereoscopic images (in *red-green* anaglyph formats) associated with their estimated 3D saliency maps and fixation density maps

4 Experimental Results

4.1 Databases Description

In order to evaluate the performance of our proposed approach, two benchmark databases are used: NBU S3D-VCA database [11] and IVY database [13]. The NBU database contains 82 indoor and 118 outdoor stereoscopic images with a wide range of horizontal disparities. The MOS of visual comfort for each stereoscopic image is provided, which is obtained via a large scale standard human subjective studies. More details about this database can be found in [11]. The IVY database contains a total number of 120 stereo image pairs with a Full-HD resolution (i.e., 1920×1080 pixels). All these images were captured using a 3D digital camera with dual lenses (Fujifilm FinePix 3D W3). The magnitude of maximum crossed horizontal disparity of each image pair ranges from 0.11 to 5.07 degrees in their experimental

environments. A large-scale standard subjective assessment was also conducted on these images to generate the associated subjective scores (i.e., MOS) to serve as the ground truth.

4.2 Performance Criteria

For performance evaluation, four performance criteria including Pearson linear correlation coefficient (PLCC), Spearman rank order correlation coefficient (SRCC), Kendall rank-order correlation coefficient (KRCC), and Root mean squared error (RMSE), between the predicted visual comfort scores and MOSs are computed. Among them, PLCC and RMSE are used to measure the prediction accuracy, and SRCC and KRCC are used to measure the prediction monotonicity. For a well-defined model, we have PLCC = SRCC = KRCC = 1 and RMSE = 0. As recommended by the Video Quality Experts Group [26], we perform a nonlinear regression using the following logisitic function on the objective visual comfort scores given by VCA models before computing these criteria. The used logistic mapping function is defined as

$$MOS_{pi} = \beta\left(\frac{1}{2} - \frac{1}{1 + \exp\left(\beta_2\left(s_i - \beta_3\right)\right)}\right) + \beta_4 x + \beta_5 \tag{10}$$

where $\beta_j\,(j = 1, 2, \ldots, 5)$ are free model parameters to be fitted, MOS_{pi} is the mapped visual comfort score, and s_i is the objective score given by the VCA model.

4.3 Experimental Protocol

The performance of a VCA model is evaluated by a 200 times 10-fold cross validation on each individual database. Specifically, we first randomly divide the each database into 10 non-overlapped subsets. For each fold, 9 subsets are used for training, and the remaining one subset is used for testing. This process will be repeated 10 folds so that each subset is used as the testing set only once. In addition, to ensure that the performance evaluation results were not dependent on a specific split, this kind of 10-fold cross validation was iterated 200 times by randomly splitting. Finally, the overall performance of a VCA model was computed as the average results over 2000 iterations (= 200 times × 10 folds).

4.4 Performance Evaluation

4.4.1 Overall Performance Comparison

We compare the proposed approach with five state-of-the-art schemes, i.e., Kim's [9], Choi's [10], Sohn's [13], and Jung's [15] schemes. As shown in Table 1, the PLCC, RMSE, SRCC, and KRCC results of these schemes on NBU database in terms of the mean values measured over 2000 iterations are presented. Note that a higher mean value corresponds to a better performance. It is obvious that, the proposed model shows the best performance against all the other compared schemes in terms of all the performance criteria. As shown in Table 2, the similar observation can also be observed from the results on IVY database in terms of SRCC and KRCC, which further demonstrate the promising performance of our proposed approach on reflecting human subjective perception of visual discomfort.

Although the scheme proposed by Jung et al. [15] also takes the human visual attention into account, it also exhibits worse performance than our proposed approach since the used 3D saliency map in [15] is generated by directly combining 2D saliency map and depth map with equal weights, which is not rational and cannot well address the issue of human fixation prediction under stereo viewing condition. Moreover, only two statistics including disparity magnitude and disparity gradient are extracted to represent a stereoscopic image in terms of visual comfort, which is usually inadequate to fully account for the sensation of experienced visual comfort of observers when viewing stereoscopic images. Another interesting observation is that the proposed approach is also slightly better than the Sohn's scheme in [13], which is also designed from the perspective of object-dependent feature

Table 1 Mean values of different VCA schemes on NBU database measured over 2000 iterations

Schemes	PLCC	SRCC	KRCC	RMSE
Kim's [9]	0.7350	0.6672	0.4987	0.5363
Choi's [10]	0.7046	0.6474	0.4832	0.5676
Sohn's [13]	0.7868	0.7608	0.5781	0.4820
Jung's [15]	0.7784	0.7647	0.5824	0.5035
Proposed	**0.8049**	**0.7769**	**0.5954**	**0.4628**

Table 2 Mean values of different VCA schemes on IVY database measured over 2000 iterations

Schemes	PLCC	SRCC	KRCC	RMSE
Kim's [9]	0.7715	0.7436	0.5537	0.5062
Choi's [10]	0.7608	0.7322	0.5414	0.5146
Sohn's [13]	0.8464	0.8178	0.6251	0.4285
Jung's [15]	0.8088	0.7837	0.5934	0.4616
Proposed	**0.8505**	**0.8208**	**0.6285**	**0.4247**

representation. However, this scheme only focuses on the statistics of most salient object while ignoring all the remaining parts of the scene. It is somewhat unreasonable because excessive disparity or crosstalk occurred in other regions can also lead to visual discomfort. Overall speaking, the proposed VCA approach can achieve a high consistency with human subjective perception due to the consideration of 3D visual saliency and the proposed adaptive 3D saliency detection model can well reflect the influence of depth structure on human fixation deployment under stereo viewing.

4.4.2 *Effectiveness of 3D Saliency-Weighted Disparity Statistics*

Since the performance of the proposed approach is highly dependent on the extracted visual comfort-aware features, it is quite interesting to examine whether these features will also be effective if incorporate them into the features used by off-the-shelf VCA schemes. Without loss of generality, we only conduct the experiments on the NBU database. Table 3 and Fig. 4 show the average PLCC, SRCC, KRCC and RMSE results over 2000 iterations by incorporating the proposed visual comfort-aware features into the off-the-shelf VCA schemes (i.e., Kim's [9], Choi's [10], Sohn's [13], and Jung's [15] schemes) on this database. It is obvious that, compared with the original results in Table 1, the performances are significantly improved for all schemes, which further demonstrate effectiveness of the proposed feature extraction strategy and also the necessity of the consideration of visual attention mechanism for visual comfort assessment.

Table 3 The PLCC, SRCC, KRCC and RMSE results over 2000 iterations by incorporating the proposed visual comfort-aware features into the off-the-shelf VCA schemes (the values in parentheses are the achieved increments)

Schemes	PLCC	SRCC	KRCC	RMSE
Kim's [9]	0.7562 (+0.0212)	0.6736 (+0.0064)	0.5132 (+0.0145)	0.5224 (−0.0139)
Choi's [10]	0.7418 (+0.0372)	0.6702 (+0.0228)	0.4936 (+0.0104)	0.5392 (−0.0284)
Sohn's [13]	0.7925 (+0.0057)	0.7729 (+0.0121)	0.5845 (+0.0064)	0.4735 (−0.0085)
Jung's [15]	0.7934 (+0.0150)	0.7745 (+0.0098)	0.5881 (+0.0057)	0.4721 (−0.0314)

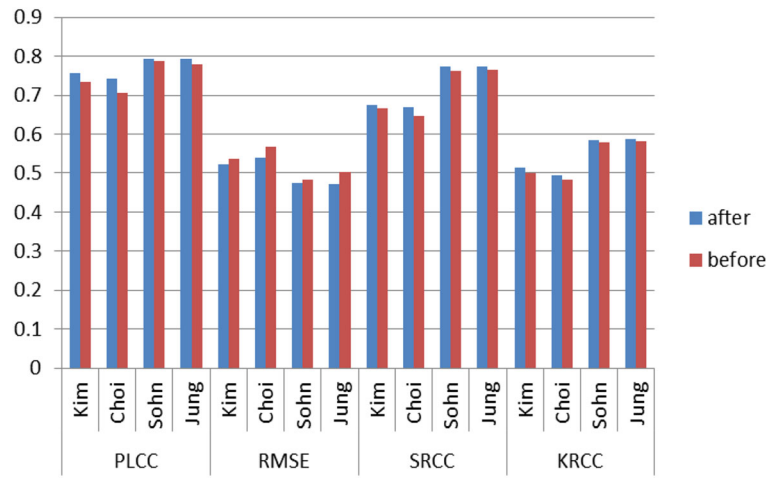

Fig. 4 Performance comparison between the previous schemes before and after incorporating the proposed 3D saliency-weighted visual comfort-aware features. It can be observed that the performances of all the schemes are improved by incorporating the proposed features

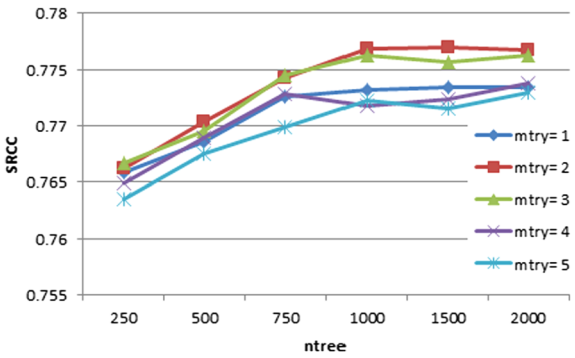

Fig. 5 Impact of different RF hyperparameter configurations

4.4.3 Impact of RF Hyperparameters

As stated, there are two most important hyperparameters namely the number of trees *ntree* and the number of variables to split on at each node *mtry* in the implementation of RF algorithm. It is of great interest to investigate how the performance of our proposed approach is influenced by different parameter configurations. In the experiments, we propose to consider the following parameters: $ntree \in \{250, 500, 750, 1000, 1500, 2000\}$, and $mtry = \{1, 2, 3, 4, 5\}$. It is worth noting that the value of *mtry* is always smaller than the dimension of features. Figure 5 shows how the performance on the NBU dataset in terms of SRCC value varies with different settings of *ntree* and *mtry*. It is clearly that almost all the criteria first sharply increase and then gently ascend as the number of constructed trees *ntree* becomes greater. This is consistent with the generally recognized fact that better prediction accuracy can be obtained as the tree number increased. However, higher computational complexity will be a major concern when a greater number of trees are constructed. As can be observed from the figure, the combination of $ntree = 1000$ and $mtry = 2$ can provide a best tradeoff between the prediction accuracy and computational efficiency.

5 Conclusions

In this paper, we have presented a novel objective visual comfort assessment (VCA) approach for stereoscopic images by taking human visual attention mechanism into account. The main contribution of this work is threefold. First, we propose a 3D saliency detection model that can adaptively fuse 2D saliency and depth saliency based on the property of depth structure. Second, visual comfort-aware features from 3D saliency-weighted disparity statistics are considered for VCA. Third, RF algorithm is applied to learn a visual comfort predictor, providing the latent mapping from high dimensional feature space to low dimensional quality score space. Experimental results on NBU and IVY databases confirm the superior performance of the proposed approach. In the future work, we plan to construct a more comprehensive 3D image database that simultaneously accounts for various perceptual modalities (e.g., image distortion, depth perception, and visual comfort) to advance the evaluation of QoE.

Acknowledgments This work was supported in part by Natural Science Foundation of China (Grant No. 61501270), in part by Zhejiang Provincial Natural Science Foundation of China (Grant No. LY14F010004), in part by Open fund of Zhejiang Provincial Key Academic Project (first level), in part by College Students Science and Technology Innovation Project (Xin Miao Talent Project) of Zhejiang Province (2014R40 5077), in part by Ningbo Natural Science Foundation (2016A610071), and in part by the Scientific Research Foundation of Ningbo University (XYL15025).

References

1. Hur, N., Lee, H., Lee, G., Lee, S., Gotchev, A., & Park, S. (2011). 3DTV broadcasting and distribution systems. *IEEE Transactions on Broadcasting, 57*(2), 395–407.
2. Lambooij, M., Ijsselsteijn, W., Fortuin, M., & Heynderickx, I. (2009). Visual discomfort and visual fatigue of stereoscopic displays: A review. *Journal of Imaging Science and Technology, 53*(3), 1–14.
3. Tam, W., Speranza, F., Yano, S., Shimono, K., & Ono, H. (2011). Stereoscopic 3D-TV: Visual comfort. *IEEE Transactions on Broadcasting, 57*(2), 335–346.
4. Urvoy, M., Barkowsky, M., & Le Callet, P. (2013). How visual fatigue and discomfort impact 3D-TV quality of experience: a comprehensive review of technological, psychophysical, and psychological factors. *Annals of Telecommunications-Annales Des Télécommunications, 68*(11–12), 641–655.
5. A. Mittal, A. Moorthy, J. Ghosh, and A.C. Bovik (2011) Algorithmic assessment of 3D quality of experience for images and videos, Proceedings of IEEE Digital Signal Processing Workshop, pp. 338–343
6. Lambooij, M., IJsselsteijn, W., & Heynderickx, I. (2011). "Visual discomfort of 3D TV: Assessment methods and modeling". *Displays, 32*(4), 209–218.
7. Ukai, K., & Howarth, P. (2008). Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations. *Displays, 29*(2), 106–116.

8. Yano, S., Ide, S., Mitsuhashi, T., & Thwaites, H. (2002). A study of visual fatigue and visual comfort for 3D HDTV/HDTV images. *Displays, 23*(4), 191–201.
9. Kim, D., & Sohn, K. (2011). Visual fatigue prediction for stereoscopic image. *IEEE Transactions on Circuits System and Video Technology, 21*(2), 231–236.
10. Choi, J., Kim, D., Choi, S., & Sohn, K. (2010). Visual fatigue modeling and analysis for stereoscopic video. *Optical Engineering, 51*(1), 017206.
11. Jiang, Q., Shao, F., Jiang, G., Yu, M., & Peng, Z. (2015). Three dimensional visual comfort assessment via preference learning. *Journal of Electronic Imaging, 24*(4), 043002.
12. Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207.
13. Sohn, H., Jung, Y., Lee, S., & Ro, Y. (2013). Predicting visual discomfort using object size and disparity information in stereoscopic images. *IEEE Transactions on Broadcasting, 59*(1), 28–37.
14. Lee, S., Jung, Y., Sohn, H., Speranza, F., & Ro, Y. (2013). Effect of stimulus width on the perceived visual discomfort in viewing stereoscopic 3D-TV. *IEEE Transactions on Broadcasting, 59*(4), 580–590.
15. Jung, Y., Sohn, H., Lee, S., Park, H., & Ro, Y. (2013). Predicting visual discomfort of stereoscopic images using human attention model. *IEEE Transactions on Circuits System and Video Technology, 23*(12), 2077–2082.
16. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
17. Gu, K., Zhai, G., Yang, X., & Zhang, W. (2015). Using free energy principle for blind image quality assessment. *IEEE Transactions on Multimedia, 17*(1), 50–63.
18. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.
19. Harel, J., Koch, C., Perona, P. (2006) Graph-based visual saliency, Proceedings of Advances in Neural Information Processing Systems.
20. Goferman, S., Zelnik-Manor, L., & Tal, A. (2012). Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(10), 1915–1926.
21. Hou, X., Harel, J., & Koch, C. (2012). Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(1), 194–201.
22. Li, J., Levine, M., An, X., Xu, X., & He, H. (2013). Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(4), 996–1010.
23. Martinel, N., Micheloni, C., & Foresti, G. L. (2015). Kernelized saliency-based person re-identification through multiple metric learning. *IEEE Transactions on Image Processing, 24*(12), 5645–5658.
24. Lang, C., Nguyen, T., Katti, H., Yadati, K., Kankanhalli, M., Yan, S. (2012). Depth matters: Influence of depth cues on visual saliency. In *Proceeding of 12th European Conference on Computer Vision (ECCV)*.
25. Park, J., Lee, S., & Bovik, A. (2014). 3D visual discomfort prediction: Vergence, foveation, and the physiological optics of accommodation. *IEEE Journal of Selected Topics in Signal Processing, 8*(3), 415–426.
26. Final report from the video quality experts group on the validation of objective models of video quality assessment VQEG, 2000. http://www.vqeg.org