Sparse representation based stereoscopic image quality assessment accounting for perceptual cognitive process

Sparse Representation Based Stereoscopic Image Quality Assessment Accounting for Perceptual Cognitive Process

Jiachen Yang[a], Bin Jiang[a], Yafang Wang[a], Wen Lu[b,*], Qinggang Meng[c]

[a]School of electrical and Information Engineering, Tianjin university, Tianjin, 300072, P. R. China
[b]School of Electronic Engineering, Xidian University, Xi'an 710071, Shanxi Province, China
[c]Department of Computer Science, School of Science at Loughborough University, UK

Abstract

In this paper, we propose a sparse representation based Reduced-Reference Image Quality Assessment (RR-IQA) index for stereoscopic images from the following two perspectives: 1) Human visual system (HVS) always tries to infer the meaningful information and reduces uncertainty from the visual stimuli, and the entropy of primitive (EoP) can well describe this visual cognitive progress when perceiving natural images. 2) Ocular dominance (also known as binocularity) which represents the interaction between two eyes is quantified by the sparse representation coefficients. Inspired by previous research, the perception and understanding of an image is considered as an active inference process determined by the level of "surprise", which can be described by EoP. Therefore, the primitives learnt from natural images can be utilized to evaluate the visual information by computing entropy. Meanwhile, considering the binocularity in stereo image quality assessment, a feasible way is proposed to characterize this binocular process according to the sparse representation coefficients of each view. Experimental results on LIVE 3D image databases and MCL database further demonstrate that the proposed algorithm achieves high consistency with subjective evaluation.

*Corresponding author

Email addresses: yangjiachen@tju.edu.cn (Jiachen Yang), jiangbin@tju.edu.cn (Bin Jiang), WangYF0739@tju.edu.cn (Yafang Wang), luwen@mail.xidian.edu.cn (Wen Lu), q.meng@lboro.ac.uk (Qinggang Meng)

1. Introduction

3D multimedia greatly enriches the way we perceive the world and attracts much more attention than ever before because of better psychological perception. More and more scholars pay close attention to the research of visual information processing to promote the development of 3D media technology [1, 2]. However, stereoscopic image quality assessment (SIQA, also known as 3D quality of experience (QoE) in some literatures) is still a challenging area in image processing due to the complex and non-intuitive interactions of multiple 3D visual cues including depth perception, visual comfort, and 2D-IQA related issues [3]. Meanwhile, since the 3D images we perceive are synthetic images of the brain, how to effectively simulate the visual system for image synthesis also poses a challenge to the 3D research and more researches are developed in the field of image synthesis [4, 5, 6, 7, 8].

There have been a number of extraordinary metrics that concentrate on different aspects of 2D-IQA. For example, structural similarity index (SSIM) focused on image formation [9], and visual information fidelity (VIF) making full use of information theory [10]. Natural scene statistics (NSS) based metrics analyze how different distortions affect statistical properties, as introduced in [11, 12]. Orientation selectivity mechanism based visual pattern was proposed for visual information representation [13]. Meanwhile, a number of deep learning based blind IQA frameworks also achieve fairly good performances [14, 15].

However, owing to the remarkable distinction between human monocular and binocular vision, the quality of the stereopair is not a simple combination of individual views' quality [16]. When viewing a stereopair, HVS does not examine the left and right view individually; instead, complex binocular fusion and rivalry processes take effect and generate a synthetic 3D mental view [17]. In this process, the quality of the stereopair is affected not only by the degradation level of each view, but also the complicated binocular interaction.

In SIQA, the left and right views can be symmetrically distorted, which reduces the difficulties in evaluating the quality of stereopairs, and correspondingly, SIQA methods always show higher consistency with subjective

2

assessment. In these cases, the interaction between the two views is relatively simple and accounts for a less dominant role. In this case, applying 2D-IQA methods into the prediction of 3D image quality seems a reasonable way and achieves good performance. Later on, supposing that depth/disparity information also contributes to the 3D image quality, You *et al.* employed depth/disparity to tackle with this issue [18]. However, there were some key points should not be ignored. At first, the ground truth disparity is not always available and these methods can only work on estimated disparity from distorted stereopairs. Thus the accuracy of corresponding disparity estimation algorithms has a huge influence on SIQA. Goldmann *et al.* have also demonstrated the fact that 3D quality is more than just the combination of two individual views and disparity [19].

Despite the shortcomings of the techniques mentioned above, they can still perform fairly well in predicting the quality of symmetrically distorted 3D images. However, predicting the quality of asymmetrically distorted images is much more challenging due to the asymmetrical visual response of the two views [20]. Binocular vision is much more complex and shows different properties under various distortion types. In [3], the relationship between different types of distortion and their binocular quality perception was explored. According to the research, the distortion can be divided into information-loss distortion (e.g. blur) and information-additive distortion (e.g. blockiness). The perceived quality of a stereopair is dominated by the high-quality component of information-loss distortion, while it is not true with information-additive distortion [21]. To model the complex process of binocular vision, binocular visual behaviors are described as visual input integration or combination. Some biological models were proposed to address this issue, such as Eye-Weighting [22], Gain-Control [23], Neural Network [24] and Vector Summation [25]. In most cases, the monocular stimulus with strong visual response is always given a heavier weight and considered to play a dominant role in judging 3D perception. However binocular fusion and rivalry properties are not fully considered or correctly solved in these methods, and the evaluation of asymmetrically distorted images is still a challenge. Hopefully, cyclopean image provides a better way to account for binocular visual properties.

Chen *et al.* developed an intermediate image which has a perceived quality close to that of cyclopean image [17]. The cyclopean image is regarded as a simulation of binocular single vision perceived in the brain which takes complex binocular interactions into consideration. Inspired by the cyclopean

image, SIQA has been further developed. In [26], an RR-IQA method was proposed based on the statistical properties of the stereoscopic image in the reorganized discrete cosine transform domain. Shao *et al.* trained two deep neural networks from monocular images and cyclopean images to model the process of quality prediction [20]. Although the cyclopean image achieves better consistency with subjective prediction, there are still more efforts to explore about how it is formulated and its relationship with the extent of various degradation.

In addition to the achievements mentioned above, there is also research on binocular summation and difference channels [27]. Different from the traditional thought that stereo perception was achieved by combining signals from two views, HVS has separately adaptable channels for adding and subtracting the neural signals from the two views which was demonstrated that encoding the summation and difference between two stereo-halves can be used for stereopsis. In our previous research, we found that binocular difference has strong correlation with correlated disparity [28]. In other words, binocular difference map can be regarded as an alternative of binocular disparity. Inspired by this observation, a quality index for stereoscopic images by separately evaluating adding and subtracting signals was proposed. In this paper, we further explore the binocularity (also known as ocular dominance) with regard to binocular summation and difference channels. To be specific, the proportion of sparse coefficients from binocular difference and summation channels is highly related to distortions and binocular visual properties.

From neuro-biological research, the receptive fields of simple cells in mammalian primary visual cortex can be reasonably characterized by a coding strategy that maximizes sparseness. To understand the behavior of visual perception, sparse coding would be a more efficient representation for later stages of visual cortex [29]. Motivated by these conclusions, many sparse coding based methods have been applied into predicting image quality. Xu *et al.* [30] exploited sparse component to handle tail and researched the multi-label learning problem under the low-rank principle. Guha *et al.* [31] proposed a sparse representation-based approach to encode the information of an image using the information from other images, and used the sparse of the representation as a measure of its compressibility with respect to other images. Considering the importance of the feature selection for the learning performance, Liu *et al.* [32] proposed cost-sensitive feature selection (CSFS) method based on sparse optimization. Ma *et al.* proposed the concept of EoP in evaluating the visual information and found that the changing tendency of

EoP in image sparse representation was highly relevant to the procedure of perceiving natural images. Shao *et al.* proposed a full-reference (FR) SIQA method by learning binocular receptive field properties [33]. However, there was less work focusing on the sparse representation in 3D domain, and it failed to address the binocular visual properties by directly using 2D-based spare coding model into SIQA. To solve these problems, we extend the EoP into evaluating the quality of stereoscopic images. At first, EoP is utilized to describe the visual cognitive progress when perceiving natural images. Secondly, perceptual information (termed as entropy) is encoded by visual primitives, and all the converging curves of the entropy were different from each other in terms of various distortion. Thirdly, the EoP values of the difference channel and the summation channel are used to describe the stereo perception related issues, especially under the circumstance of asymmetric distortion.

In conclusion, we argue that the coefficients after the primitive encoding are not only consistent with the perception procedure, but also reasonably related with binocularity. Therefore, we propose an image primitive coding based RR-SIQA framework accounting for perceptual cognitive process. The main contributions are as follows.

1) Image primitives are used to capture the latent structure of reference image dictionary in the training phase.

2) EoP, which shows a similar trend of hierarchical perceptual cognitive process of human eyes, is applied into the prediction of image quality.

3) Binocularity (ocular dominance) is further explored using the sparse coding coefficients from difference and summation channels, to fill the gap between stereo perception related issues and asymmetric distortion.

The rest of the paper is organized as follows. Section II introduces the related work. Section III presents the framework of the proposed SIQA. The experimental results and analysis are given in Section IV, and finally conclusions are drawn in Section V.

2. Related Work

2.1. Image Primitives

The concept of image primitive coding is established on sparseland, which assumes that natural signals, such as images, can be adaptively represented with a small number of elementary signals [34]. As for an image, a set of image primitives in the learned dictionary can effectively describe its visual

information. To be specific, for a given image, the basic units of sparse representation are patches with size $\sqrt{b} \times \sqrt{b}$ at location $[i, i = 1, 2, 3..., n]$ that are mathematically denoted by $X = [x_1, x_1, ..., x_n] \in R^{B \times N}$ and $x_i \in R^b$, where N is the number of sample images, b is the size of each path vector, and n is the number of patches [35, 36].

Using the collected vectors X as input, we learn an over-complete dictionary $D = [d_1, d_2, ..., d_n] \in R^{b \times m} (m > n)$ with m basic image primitives by seeking a sparse representation for each patch. In sparse representation, the dictionary learning process can be formulated as

$$D\{a_i\} = \underset{D_i\{a_i\}}{argmin} \sum_i ||x_i - Da_i|| \ s.t. ||a_i||_0 < T \tag{1}$$

The vector $a_i \in R^m$ is coefficient vector for the patch $\{x_i\}$. $||a_i||_0$ is the l_0 norm, and T is the constraint of the non-zero number of a_i.

For a patch x_i, sparse coding is the process of finding the sparse representation a_i with D. To obtain the sparse representation, sparse coding can be formulated as

$$a_i = \underset{a_i}{argmin} \sum_m ||x_i - Da_i||_2^2 \ s.t. ||a_i||_0 < L \tag{2}$$

where L controls the sparse level. Though the problem is NP-hard in general, it is solved by orthogonal matching pursuit (OMP) [35] algorithm for its simplicity and efficiency.

In the proposed framework, Eq. 1 can be effectively solved by the K-SVD algorithm [35]. m is set to 256, b is set to 64, T is set to 3, L is set to 14, and n is the number of patches collected from 30 reference images from LIVE 2D Image Database. The detailed process of dictionary learning is shown in Fig. 1
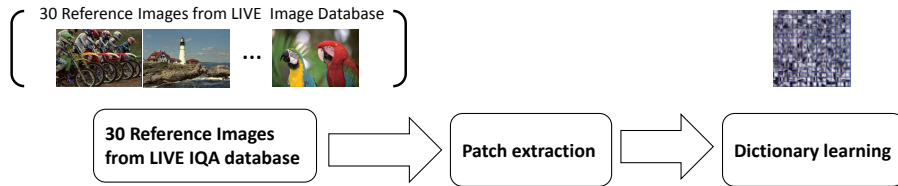

Figure 1: Illustration of dictionary learning.

2.2. Entropy of Primitives

According to the internal generative mechanism (IGM), HVS actively predicts visual information and tries to avoid uncertainty. An input image can be decomposed into the predicted portion which contains primary information (represented by structural information) and the disorder portion which mainly contains uncertain information (represented by non-structural information) [37]. In [38], a novel concept of EoP is developed to distinguish the two portions using the information conveyed by different types of image primitives. With this method, the visual information can be quantified to evaluate the "surprise" level known as the "free energy principle" introduced in [39], and determine the boundary of structural and non-structural layers. The mathematical expression of EoP is denoted as

$$EoP_i = -\sum\nolimits_{j=1}^{k} P^i(j) \log P^i(j) \tag{3}$$

$$P^i(j) = \frac{N_j^i}{\sum_t N_t^i} \tag{4}$$

$$N_j^i = \sum\nolimits_{t=1} n_j^t \tag{5}$$

where n_j^t indicates the number of the jth primitive used in the ith iteration during the OMP algorithm, N_j^i represents the total number of the ith primitive used in the previous iterations, and P is the probability density function (PDF).

Fig. 2 illustrates the hierarchical image representation and reconstruction process. The image is decomposed into multiple layers by OMP schemes according to perceptual importance. The reconstructed image by the first layer which is referred as the primary layer, maintains the most significant structure. The other layer in the structural layer part gradually recovers the detail. Since most perceptual information is captured by the former layers, the non-structural layer part is negligible. The reconstructed images with different layers (i.e. L) are also shown in Fig. 2. The reconstruction procedure corresponds with HVS, since the initial layers describe the most details of the object step by step, and HVS can hardly distinguish the differences between the last few layers.

Fig. 3 describes the tendency of EoP in terms of different iteration L. It can be observed that the EoP monotonously increases and converges to a constant level with the increasing value of L. At the same time, we also notice
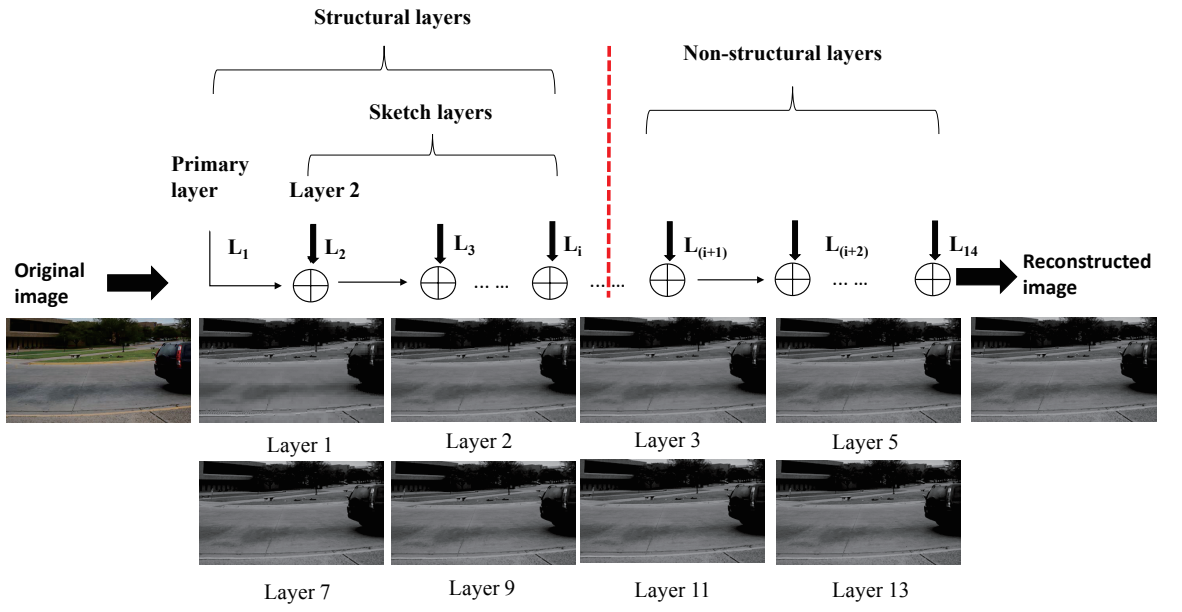
7

Figure 2: The hierarchical image representation and reconstruction process in terms of iteration L.

that the EoP value converges to a relatively stable level no matter what the image content is. Meanwhile, it also describes the converging curves of images with similar DMOS values, and from the differences among the curves, it can be observed that EoP values are closely related to distortion. Thus the EoP is applied into SIQA due to the fact that the type of distortion affects its tendency and maximum values. Further information about EoP is explored in [38].
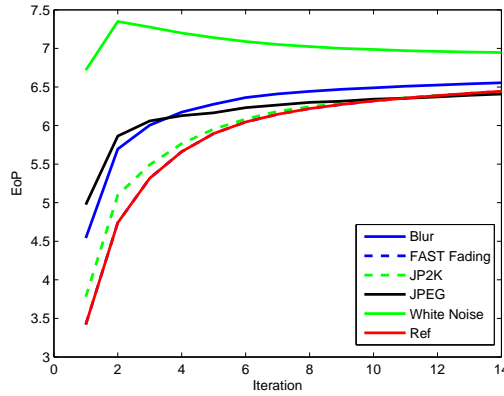

Figure 3: The EoP curves of images with similar DMOS values and different contents (around 20) under the circumstances of various distortion.

As mentioned above, EoP has similar converging tendency with corresponding reconstructed images, which indicates the saturated visual information. Consequently, we adopt EoP as an effective measurement of image quality.

2.3. Binocularity (Ocular Dominance)

As demonstrated in [27], HVS has separately adaptable channels for adding and subtracting the neural signals from the two views. Encoding the summation S and difference D between two stereo-halves can be used for stereopsis.

$$
\begin{aligned}
S &= L + R \\
D &= |L - R|
\end{aligned}
\tag{6}
$$

Fig. 4 shows examples of binocular difference and summation. Here, we further explore the role of binocular summation and difference. Following

9

the work introduced in the former subsection, we analyze binocularity issues with the help of image primitive coding. To be specific, the binocularity is defined as :

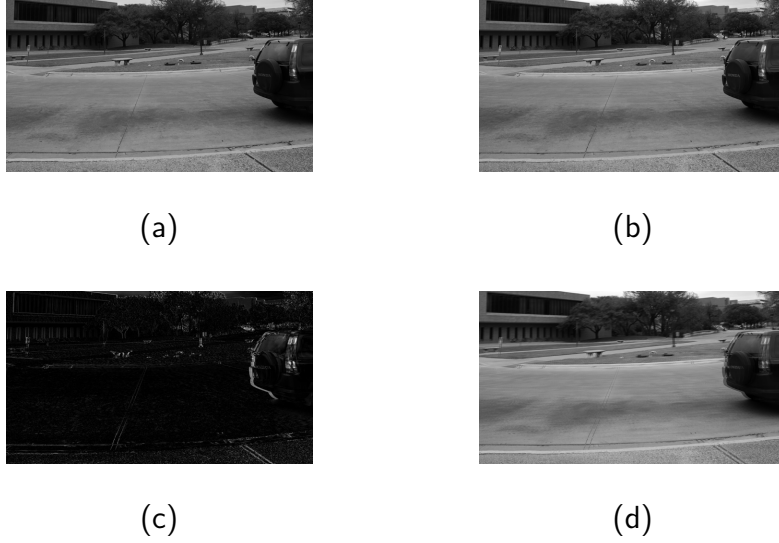$$b = \frac{|W_{left}| - |W_{right}|}{|W_{left}| + |W_{right}|} \tag{7}$$

Figure 4: Binocular summation and difference channels. (a) The left view. (b) The right view. (c) The difference channel. (d) The summation channel.

where W_{left} and W_{right} are the image primitives from each view, and b represents the degree of binocularity. Here, different values of b are used to characterize the properties from monocular response to binocular response. A large value of b represents a weak binocular response, and vice versa.

Fig. 5 shows that ocular dominance (to produce binocularity) of Fig. 4 changes with different visual primitives owning to the introduction of distortion. For example, visual responses of the 32th visual primitive from the reference, symmetrically and asymmetrically distorted stereo pairs differentiate with each other. In terms of asymmetric distortion, b is equal to 1, which represents highly monocular response-one view is in dominance while the other one is seriously suppressed. However, for reference and symmetric distortion, b equals to 0, which means that visual responses of both views are similar with each other. It is also noteworthy that ocular dominance happens

10

in a rarely high frequency in some cases, since there are a number of visual primitives are with relatively large b, even though symmetrically distorted. It may be concluded that once an eye percepts a visual signal, the other eye tend to weakly react to a similar one. As mentioned above, there seems not a consensus on how HVS react to visual stimuli, but ocular dominance is closely related to the introduction of distortion. Therefore it could be an index which characterizes stereo perception, and inspired by this idea, we find a reasonable way to describe binocularity using the image primitives.
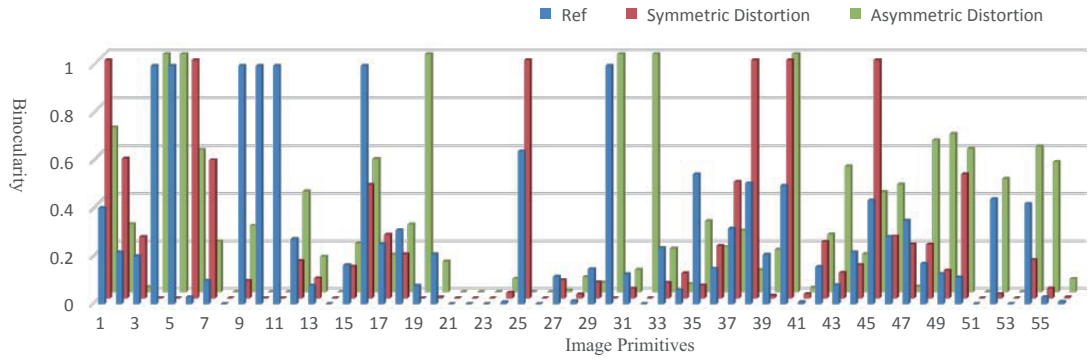

Figure 5: Binocularity of visual primitives under different distortion regiments. Note that only the last 56 visual primitives in the dictionary are shown for better display. The X axis represents different visual primitives and the Y axis denotes the level of binocularity. Green bars are the visual responses from the asymmetrically distorted stereo pair (JPEG compression for the left view and white noise for the other view). Brown bars come from symmetrically distorted stereo pairs (both JPEG compression).

2.4. The influence of various distortion on stereo perception

In terms of stereo perception, asymmetric distortion increases the difficulty to predict image quality. To further visualize its effects on binocularity and its EoP, examples are also shown in Fig. 7 which shows the EoP curves from binocularity, difference and summation channels. Note that the only difference between the symmetric and asymmetric stereo pairs is the change of distortion type in the right view. It can be seen that the white noise changes the tendency of EoP curves and therefore the EoP values from the three channels are considered as quality-aware indexes. Detailed examples are shown in Fig. 6 which describe the reconstructed images with regard to iteration number. Considering that most detailed information of the image

is recovered in the first few layers, only these examples in the 1st, 2nd, 3rd and 14th layers are shown. With the increase of iteration number, the visual degradation is centered at the back of the car and the visual uncertainty (surprise) is also reduced to a constant level.

In this section, we have made a review of related work, and demonstrated its reasonability. More importantly, these findings can be closely connected by sparse representation, and both monocular and binocular visual properties are taken into consideration. In addition, hierarchical perceptual cognitive process is also addressed by EoP.

3. The proposed Framework

Based on the physiological discoveries mentioned above, sparse representation will be a plausible way to address the critical issues related to SIQA into a unified framework. The proposed framework contains four parts: the training phase, the EoP extraction phase, the spatial and spectral entropy computation phase, and the prediction phase. The first one has been described in Section II, the second and third phase are illustrated in Fig. 8, and SVR is used in the last procedure to connect the features with image quality.

3.1. EoP from four channels

Image pre-processing: The first step of prediction is image pre-processing. DoG decomposition is employed to decompose an image into five different spatial frequency components [40]. Let $V(I)$ be the vector from representation of the DoG-bank-model of the image I, that is

$$V(I) = [V_0, \ ... \ , V_{n-1}, V_n] \tag{8}$$

In our proposed framework, n=5.

EoP of binocularity: To compute the entropy of binocularity, we still employ the way introduced in Section II, and the coefficients of each band component in the $V(I_L)$ and $V(I_R)$ are directly used as input, as expressed in Eq. 9, where $a_L^{i,j,k}$ represents the coefficient of the jth primitive used in the ith iteration in the kth patch, and the subscript L (R) denotes the left (right) view. Consequently, we get 14 values of EoP for each band after 14 iterations, and the total number of binocularity is 70 with 5 band components.

(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 6: Binocularity maps of reconstructed stereo images with regard to 1st,2nd,3rd and 14th iteration. Figures in the first row come from asymmetric distorted stereo images Figures in the first row come from symmetric distorted stereo images.

Figure 7: EoP curves of binocularity, difference channel and summation channel. The X axis denotes iterations, and the Y axis denotes EoP values.

Figure 8: The proposed framework.
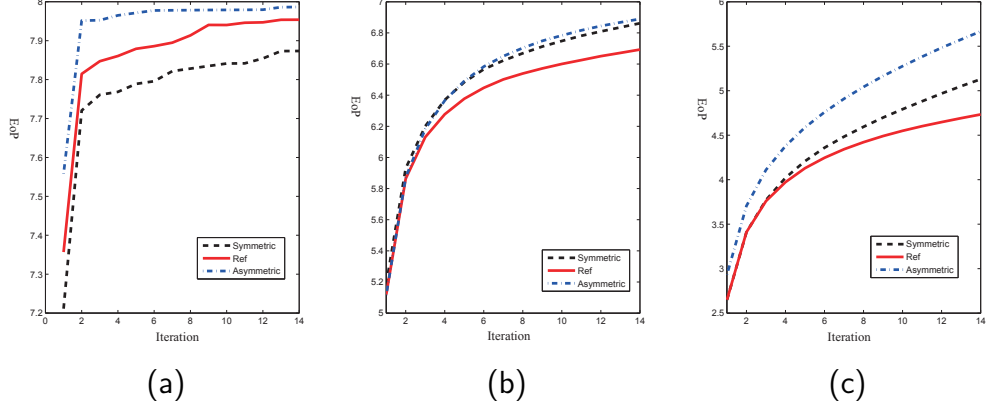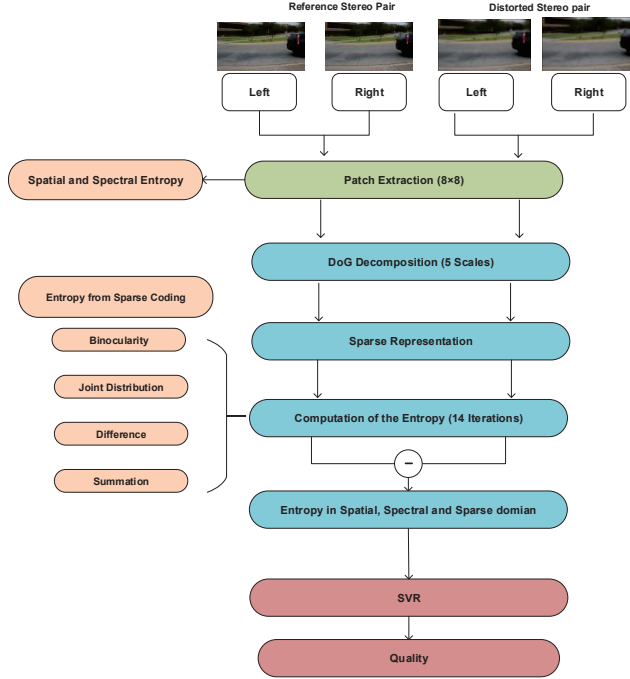
$$b^{i,j,k} = \frac{|a_L{}^{i,j,k} - a_R{}^{i,j,k}|}{a_L{}^{i,j,k} + a_R{}^{i,j,k}} \tag{9}$$

$$B_{i,k} = \sum_{j=1}^{m} b^{i,j,k} \tag{10}$$

$$P_{i,k}{}^{Binocularity} = \frac{B_{i,k}}{\sum\limits_{k=1}^{N} B_{i,k}} \tag{11}$$

$$EoP^{Binocularity}{}_i \stackrel{\Delta}{=} -\sum_{k=1}^{N} P_{i,k} \cdot log P_{i,k} \tag{12}$$

EoP of joint distribution: As defined in [33], given an image I and corresponding learnt dictionary, its sparse complexity is defined as the sparsity of sparse coefficient vectors averaged over all the patches, i.e.,

$$SC^{i,j}(I,D) = \frac{1}{n} \sum_{k=1}^{n} ||a^{i,j,k}||_2 \tag{13}$$

where $SC^{i,j}(I,D)$ denotes the sparse complexity value of the jth primitive and ith iteration. Considering that distortion affects the distribution of sparse coefficients, then the SC values are employed to be a weighting factor for binocular combination. Based on the sparse complexity, the coefficient of joint distribution is defined as:

$$c_j^i = W_L{}^{i,j} \cdot a_L{}^{i,j,k} + W_R{}^{i,j} \cdot a_R{}^{i,j,k} \tag{14}$$

$$W_L{}^{i,j} = \frac{SC_L{}^{i,j}(I_L,D)}{SC_L{}^{i,j}(I_L,D) + SC_R{}^{i,j}(I_L,D)} \tag{15}$$

$$W_R{}^{i,j} = \frac{SC_R{}^{i,j}(I_R,D)}{SC_L{}^{i,j}(I_L,D) + SC_R{}^{i,j}(I_L,D)} \tag{16}$$

The EoP of the joint can be defined as :

$$EoP^{Joint}{}_i \stackrel{\Delta}{=} -\sum_{j=1}^{m} P_j \cdot log P_j \tag{17}$$

15

$$P_j^{J\text{oint}} = \frac{C_j}{\sum\limits_{j=1}^{m} C_j} \qquad (18)$$

$$C_j = \sum_{i=1}^{n} c_j^i \qquad (19)$$

Similarly, the EoP of the summation and difference channels of the ith iteration are obtained, denoted as $EoP^{Summation}{}_i$ and $EoP^{Difference}{}_i$, respectively. The computation of the two channels is simply based on the summation $(a_L{}^{i,j,k} + a_R{}^{i,j,k})$ and difference $(|a_L{}^{i,j,k} - a_R{}^{i,j,k}|)$ of the sparse coefficients from the left and right views.

In total, there are four kinds of EoP, named: $EoP^{J\text{oint}}$, $EoP^{Binocularity}$, $EoP^{Summation}$ and $EoP^{Difference}$, and each of them is obtained from five band components with 14 iterations. Thus, there are $14 \times 5 \times 4 = 280$ features.

Considering that the loss of EoP between the original and distorted stereoscopic images reflects the distortion level, we also compute the EoP from original reference stereopairs and the deviations $EoP_{deviation}$ are considered to be related with SIQA.

3.2. Spatial and spectral entropy extracted from stereopairs

Previous research has proved that there is a close relationship between local image entropy and perceived image quality [10]. Likewise, the authors in [41] also found that entropy is highly consistent with the degrees and types of distortion. Therefore, spatial entropy ($E_{spatial}$) and spectral entropy ($E_{spectral}$) are utilized to describe the statistical characteristics of local regions rather than pixels which also contribute to the description of images' local structural information. In our framework, we also utilize spatial and spectral features extracted from the difference and binocularity channels, and thus there are another $12 \times 2 = 24$ features.

3.3. Prediction

In the prediction phase, all the features are listed in Table 1. As shown in Table 1, these features are divided into 5 groups among which 4 groups respectively are the entropies of the binocularity, joint distribution of the left and right image, difference image and summation image, and the last group features are the spatial and spectral entropy. In order to show the effectiveness of these features, 3 features are randomly selected from each group

16

Table 1: The Explanation of All The Features

feature vector	Feature description	number
$f_{entropy\_binocularity}$	entropy features of the binocularity	70
$f_{entropy\_joint}$	entropy features of joint distribution	70
$f_{entropy\_Difference}$	entropy features of the difference image	70
$f_{entropy\_Summation}$	entropy features of the summation image	70
$f_{entropy\_SS}$	spatial and spectral entropy features	24

features and the relationship between these features with dmos value of LIVE Phase-I was described in Fig.9. It is obvious that almost every feature is positively correlated or negatively related to the dmos value. Meanwhile, the features of stereopairs degraded by four distortion types (JP2K, JPEG, blur and FF) show similar relations with the corresponding dmos value. These relations are good evidence that these features can reflect image quality and distortion types.

These features are fed into a SVR model, and the output is the final quality of the stereopair. In the following section, we further explore the influence of EoP on prediction results.

$$Q = SVR\left(EoP_{deviation}, \left(E_{spatial}, E_{spectral}\right)\right) \tag{20}$$

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1. Stereo Database

To verify the performance of the proposed method, the LIVE 3D Image Quality Databases (Phase I and Phase II) of the University of Texas at Austin are used [42]. Database Phase I contains 365 stereopairs with symmetric distortion and Database Phase II contains 360 stereopairs with both asymmetric and symmetric distortion, including JPEG compression, JP2K compression, white noise (WN), Gaussian blur (Blur) and fast fading (FF).

To further verify the proposed method, the MCL 3D Image Database is also adopted [43]. This database consists of nine image-plus-depth original stereo scenes and corresponding 684 distorted stereopairs. Distortions applied to either the texture image or the depth image include: Gaussian blur, additive white noise, down sampling blur, JPEG and JP2K compression, and transmission error, and the subjective scores are in the form of MOS for all the stereopairs.
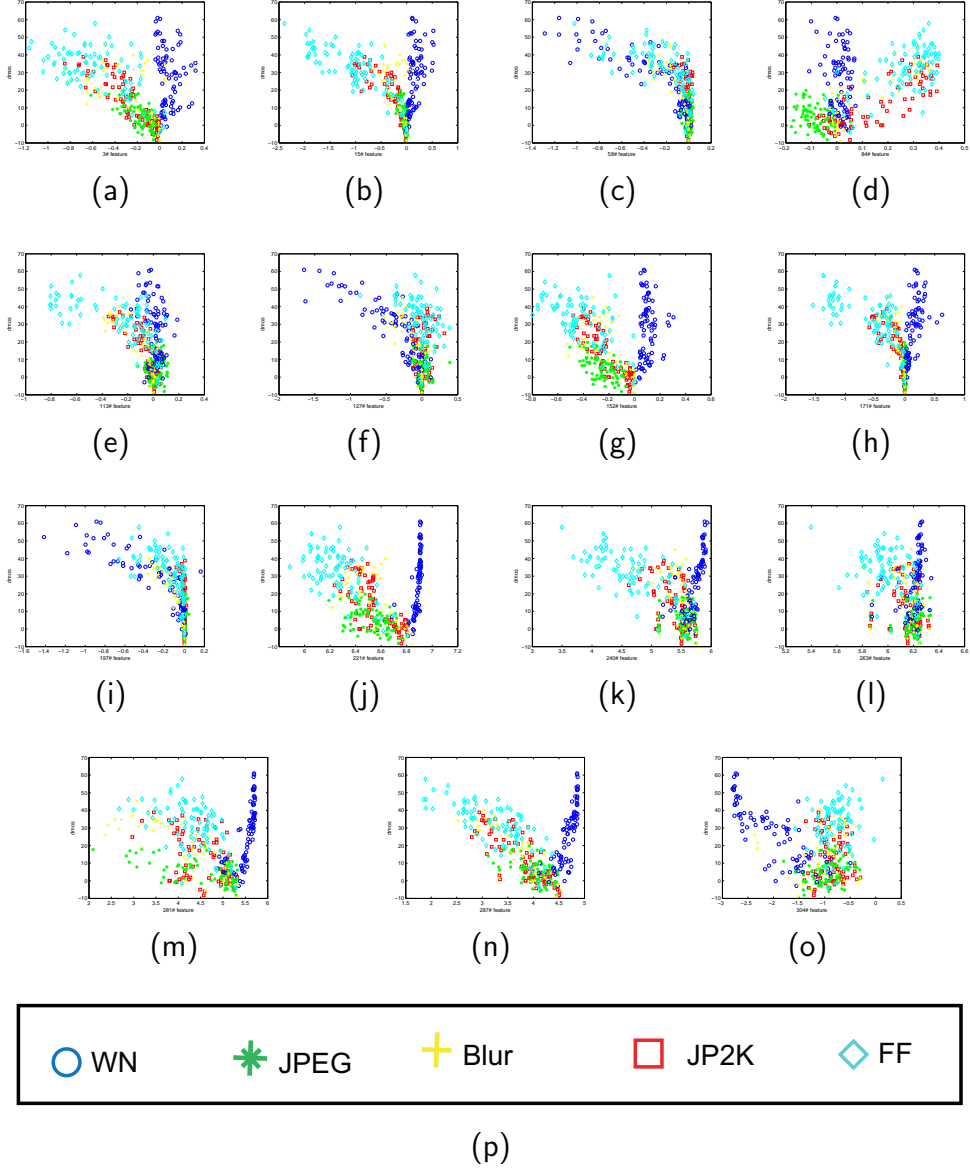
(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

(j)

(k)

(l)

(m)

(n)

(o)

○ WN   ✳ JPEG   ╋ Blur   ☐ JP2K   ◇ FF

(p)

Figure 9: The relationship between the features with the dmos value of LIVE Phase-I.

18

4.2. Performance Measure

For performance evaluation, three commonly used indicators are adopted: Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC), and Root Mean Squared Error (RMSE) between subjective scores and objective scores after nonlinear regression. For nonlinear regression, we use a 4-parameter logistic mapping function [44, 45]:

$$DMOS_P = \frac{\beta_1 - \beta_2}{e^{(Q-\beta_3)/|\beta_4|} + 1} + \beta_2 \tag{21}$$

where $\beta_1, \beta_2, \beta_3$ and β_4 are the parameters to be fitted. A better match is expected to have higher PLCC, SROCC,

In the prediction phase, the stereopairs in each database were randomly divided into two parts, with 80% for training and 20% for testing. In order to ensure that the proposed framework is robust, 1000 iterations of training are performed by varying the splitting of data over the training and testing sets, and the median value of all iterations is chosen as the final prediction quality. All the parameters of the SVR model in the proposed framework are the same for different databases.

In order to demonstrate its efficiency, the proposed method is compared with several existing state-of-art IQA metrics, including three 2D metrics (DIIVINE [46], BLIINDS-II [47] and BRISQUE [48]), and eight 3D metrics (Qi's scheme [16], Lin's scheme [40], Chen's scheme [17], Ma's scheme[26], Shao's scheme-A[49], Shao's scheme-B [33], Lv's scheme [50] and Shao's scheme-C [20]). Note that for the 2D-extended BIQA metrics, feature vectors are extracted separately from the left and right images, and weight-averaged to obtain the final feature vector for training, and SVR is used to train a regression function. For Qi's scheme, the parameter settings in the reference paper is adopted. For Lin's scheme, the FI-PSNR metric is adopted into the comparison. For Chen's scheme, the adopted 2D metric is MS-SSIM which performs the best. For Shao's schemes, experimental results in corresponding references are directly adopted.

4.3. Overall Performance on 3D Image Database

To better illustrate the experimental results, the values of PLCC, SROCC, and RMSE of each metric are shown in Table 2, where the indicators that give the best performance are highlighted in bold. As shown in the table, the overall performance of the proposed framework on Database Phase-I is

significantly better compared with other IQA metrics. Shao's scheme-C also achieves rather competitive performance. Note that Qi's scheme, Shao's scheme-A, Shao's scheme-B and ours are based on the sparse representation.

In addition to symmetric distortion, performance on Database Phase II which contains both symmetric and asymmetric distortion is also a concern in SIQA. It can be observed that the proposed framework overtakes the other metrics by a large extent. The significant difference further confirms the previous conclusion that our framework can effectively predict the quality of stereoscopically viewed images. However, it is worth noting that compared with symmetric distortion, the method predicting the quality of asymmetric distortion stereopairs still needs further improvement.

We also pay much attention to the role of EoP, the visual perception models (binocularity model, summation-difference channel) and the features of spatial and spectral entropy in determining image quality. The results respectively based on binocularity model, the summation-difference channel, the results of only using EoP and the results of only using spatial and spectral entropy (enropy-SS) are reported in the table 2. It can be observed that the EoP and the entropy-SS are closely related to the image quality and compared with the sparse representation based methods (Qi's scheme, Shao's scheme-A and Shao's scheme-B), their overall performance are also competitive, although they still lag behind the two best metrics (the proposed metric and Shao's scheme-C). Although the performance of the metric simulating summation-difference channel is not very good, it is better than most 3D models. The reason for that the metric only considering binocularity does not perform well is that binocularity mainly based on the ocular dominance, and the binocular fusion is neglected more or less. However, its performance is also much better than that of Lin's and Shao-A schemes, and this result shows the binocularity reflects the human visual mechanism to some extent. Meanwhile, Lv's and Shao-C schemes are based on deep learning, but the proposed metric performs better than the two metrics. This phenomenon shows that, although only considering the shallow features, our model is also good for simulating stereo vision systems, and at the same time, our metric also shows advantages under the condition of small samples.

To verify the demonstration mentioned above, we also report the performances on MCL 3D Image Database according to the corresponding settings. As shown in the Table 3, the proposed framework also achieves the most competitive performance, even only depending on EoP. Note that the MCL database contains two subsets: the FR database and the NR database. Con-

Table 2: Overall performances on LIVE 3D Image Database

Criteria	LIVE Phase I			LIVE Phase II		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
DIIVINE [46]	0.8317	0.8105	9.2291	0.7916	0.7596	6.8615
BLIINDS-II [47]	0.9361	0.9262	5.7781	0.8795	0.8682	5.3539
BRISQUE [48]	0.9540	0.9401	4.9043	0.9177	0.9084	4.4335
Lin [40]	0.8645	0.8559	8.2424	0.6584	0.6375	8.4956
Shao-A[49]	0.8913	0.8849	-	0.7843	0.8054	-
Ma [26]	0.9033	0.9034	7.0353	0.8431	0.8093	6.0694
Chen [17]	0.9161	0.9153	6.5740	0.9067	0.9068	4.7587
Qi [16]	0.9086	0.8783	6.7782	0.8798	0.8734	5.1425
Shao-B [33]	0.9350	0.9251	5.8155	0.8628	0.8494	5.7058
Lv [50]	0.901	0.8979	-	0.8700	0.8624	-
Shao-C [20]	0.9565	0.9449	4.7552	0.9265	0.9106	4.3381
entropy-SS	0.9274	0.9135	6.1189	0.8838	0.8545	5.2461
binocularity	0.8948	0.8796	7.2244	0.8231	0.8239	5.0854
summation-difference	0.9221	0.9083	6.1798	0.9021	0.8863	4.8289
Proposed-EoP	0.9312	0.9285	5.6370	0.8970	0.8711	4.9037
Proposed	**0.9601**	**0.9479**	**4.2774**	**0.9386**	**0.9249**	**3.7770**

Table 3: Overall performances on MCL 3D Image Database

Criteria	PLCC	SROCC	RMSE
Lin [40]	0.7854	0.6854	1.6100
Chen [17]	0.8278	0.8300	1.4596
Qi [16]	0.7480	0.6685	1.7766
Shao-C [20]	0.9138	0.9040	1.0233
Proposed	**0.9388**	**0.9225**	**0.9089**
Proposed-EoP	0.8987	0.8686	1.1769

sidering that the proposed metric is a RR metric, so the prediction results are based on the FR database, and the same is true of another two FR metrics (Chen's and Lin's Schemes). The results further confirm the demonstration that entropy of primitives from the four channels can effectively predict the quality of stereoscopically viewed images.

As mentioned above, results on individual database are reported, and the proposed framework achieves outstanding performance among all the metrics. However, the samples used for training and testing are selected from the same dataset and thus this approach is not sufficient to explain the generality and stability of the evaluation model. To exclude the effect of database dependence, cross-database experiments are also carried out in this part. Those metrics in training phase are trained on one database and tested on another database. Table 4 gives the detailed information of cross-database test. It can be observed from the table, the performance of each metric has a significant decline compared with the result on individual dataset because of the difference between databases, such as source images and distortion types. However, our metric still have a relatively better predictive ability among those metrics. When the metrics are trained on LIVE phase II and tested on LIVE phase I, those metrics perform better since Database Phase-II contains both symmetrically and asymmetrically distorted stereopairs. In contrast, when training is performed on Database Phase-I, the performance significantly declines because that there are only symmetric distortion types in LIVE phase I, and therefore, its ability to predict asymmetric distortion is relatively poor. Besides, whether these models are trained or tested on the MCL dataset, the results are not very good. The reason for this result is that most images in MCL are depth images and texture images, which are quite different compared with those images in LIVE phase I and II. This also shows the training and prediction ability of these metrics for depth map and texture map are still poor at present. However, our metric shows an advantage over the other SIQA no matter when it's trained on the MCL and tested on the other two datasets, or trained on two other datasets and tested on the MCL. Overall, the results on cross-database experiments show that SIQA still needs further exploration to overcome the database dependence.

The tendency of PLCC values on both databases when trained on LIVE-II is reported in Fig. 10. It can be observed that the prediction results become stable when the training ratio is larger than 0.4, and therefore the proposed framework could achieve better results on a relatively small training dataset.

The scatter plots of the proposed method on LIVE 3D Image Database

Table 4: Cross-database performance compared with others metrics

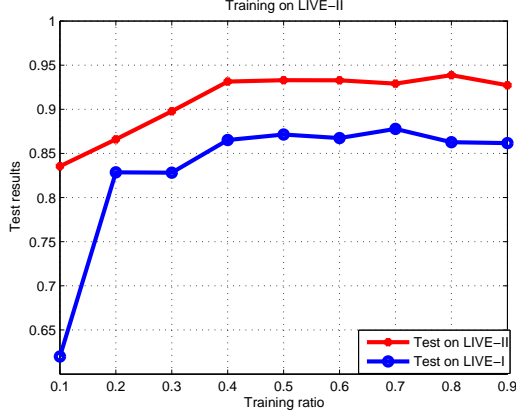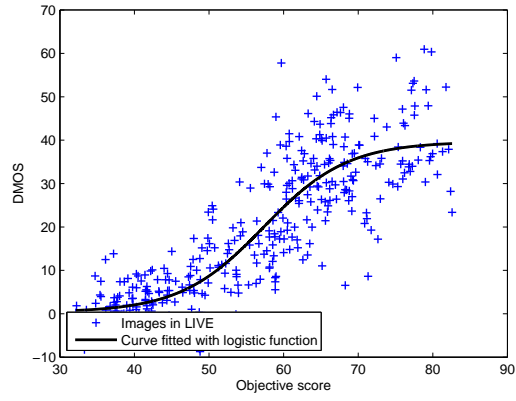| Criteria		LIVE I(testing)		LIVE II (testing)		MCL (testing)	
		PLCC	*SROCC*	*PLCC*	*SROCC*	*PLCC*	*SROCC*
LIVE I (training)	DIIVINE [46]	-	-	0.6161	0.5177	0.0461	0.1450
	BLIINDS-II [47]	-	-	0.6986	0.6716	0.5981	0.5221
	BRISQUE [48]	-	-	0.5949	0.4582	0.6063	0.3711
	Proposed	-	-	**0.8038**	**0.7898**	**0.6104**	**0.6450**
	Proposed-EoP	-	-	0.6900	0.5075	0.4749	0.3315
LIVE II (training)	DIIVINE [46]	0.5375	0.4954	-	-	0.5120	0.3653
	BLIINDS-II [47]	0.7652	0.7398	-	-	0.5827	0.5501
	BRISQUE [48]	0.5718	0.5564	-	-	0.5581	0.5599
	Proposed	0.8627	0.8541	-	-	**0.6035**	**0.6150**
	Proposed-EoP	**0.8914**	**0.8851**	-	-	0.5856	0.6550
MCL (training)	DIIVINE [46]	0.5115	0.1605	0.4651	0.1484	-	-
	BLIINDS-II [47]	0.6574	0.5768	**0.5753**	**9.2325**	-	-
	BRISQUE [48]	0.5213	0.4243	0.3478	0.2759	-	-
	Proposed	0.6433	0.5694	0.4488	0.4051	-	-
	Proposed-EoP	**0.7338**	**0.5473**	0.4812	0.3752	-	-

Figure 10: PLCC of the proposed metric on LIVE-I and LIVE-II in terms of training ratio (Training on LIVE-II).
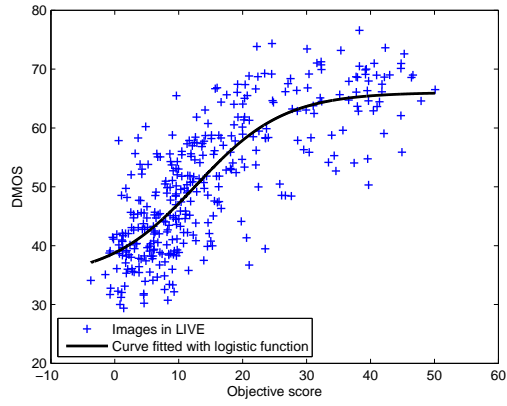
are shown in Fig. 11. The vertical axis denotes the human subjective scores of perceptual quality and the horizontal axis denotes the predicted scores. A better convergence of the points in the scatter plot to the fitting curve means a better consistency with the DMOS. The scatter plots demonstrate that the scatter points of the proposed method are concentrated to the fitting curve, which indicates a more prominent prediction.

In the previous subsection, we discussed the overall performance and the cross-database performance of the metric. However, good overall performance does not always mean good performance for individual distortion type. In this subsection, the performance of the metric compared with the other two 2D-extended metrics, BRISQUE and BLIINDS-II, on different distortion type will be analyzed and the results are listed on Table 5. The top indictors are highlighted with boldface. As shown in this table, the proposed metric has the best performance on all distortion types, whether for symmetric distortion stereopairs or asymmetric distortion stereopairs. For structural distortion, like JPEG and JP2K, all the three metrics do not have very good performance, but our metric is still superior to BRISQUE and BLIINDS-II.

To summarize, IQA is a perceptual cognitive process, in which entropy plays a crucial role in determining visual quality. Meanwhile, considering stereo perception is rather complex, ocular dominance (binocularity) could be an alternative to describe binocular interactions between the two views. The proposed method takes advantage of the two determinant factors to

24

(a)

(b)

Figure 11: Scatter plots of objective scores versus subjective data (DMOS). (a) Training on LIVE II/Test on LIVE I. (b) Training on LIVE I/Test on LIVE II.

Table 5: The performance of the proposed metric compared with several metrics on Individual distortion type

Criteria		BRISQUE [48]		BLIINDS-II [47]		Proposed	
		PLCC	*SROCC*	*PLCC*	*SROCC*	*PLCC*	*SROCC*
LIVE I	*JPEG*	0.7416	0.6630	0.7929	0.7231	**0.8248**	**0.7941**
	JP2K	0.9108	0.8361	0.9172	0.8915	**0.9656**	**0.9382**
	WN	0.9403	0.9059	0.9185	0.8914	**0.9593**	**0.9357**
	FF	0.8214	0.7613	0.8013	0.7015	**0.9060**	**0.8794**
	Blur	0.9501	0.8629	0.9373	0.8448	**0.9511**	**0.9167**
	All	0.9540	0.9401	0.9361	0.9262	**0.9601**	**0.9479**
LIVE II	*JPEG*	0.7953	0.7032	0.7699	0.6711	**0.8662**	**0.8110**
	JP2K	0.7392	0.6732	0.8256	0.7750	**0.9158**	**0.8901**
	WN	0.8401	0.7950	0.8333	0.7956	**0.9681**	**0.9473**
	FF	0.9319	0.9039	0.9301	0.9099	**0.9680**	**0.9429**
	Blur	0.9473	0.8482	0.9504	0.8122	**0.9707**	**0.9473**
	All	0.9177	0.9084	0.8795	0.8682	**0.9386**	**0.9249**

predict perceived quality using the entropy from sparse representation. And the superiority has been further confirmed by experiments.

5. Conclusions

In this paper, a sparse coding based quality index for stereoscopic images is proposed, in which the entropy is used to describe the perceptual cognitive process. Based on the fact that EoP is closely related to distortion, we use the entropy information from four channels (namely, binocularity, the joint distribution, summation, and difference) to describe visual degradation. Experiments further confirm that the proposed framework is highly consistent with subjective test.

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

6. Acknowledgments

7. Reference

[1] Y. Wang, C. Xu, S. You, C. Xu, and D. Tao. Dct regularized extreme visual recovery. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, pages 1–1, 2017.

[2] Shuhang Gu, Qi Xie, Deyu Meng, Wangmeng Zuo, Xiangchu Feng, and Lei Zhang. Weighted nuclear norm minimization and its applications to low level vision. *International Journal of Computer Vision*, 121(2):183–208, 2017.

[3] Seungchul Ryu and Kwanghoon Sohn. No-reference quality assessment for stereoscopic images based on binocular quality perception. *Circuits and Systems for Video Technology, IEEE Transactions on*, 24(4):591–602, 2014.

[4] Nannan Wang, Xinbo Gao, Jie Li, Bin Song, and Zan Li. Evaluation on synthesized face sketches. *Neurocomputing*, 214(C):991–1000, 2016.

[5] Nannan Wang, Xinbo Gao, Leiyu Sun, and Li Jie. Bayesian face sketch synthesis. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, 26(3):1264–1274, 2017.

[6] Nannan Wang, Mingrui Zhu, Jie Li, Bin Song, and Zan Li. Data-driven vs. model-driven: Fast face sketch synthesis. *Neurocomputing*, 2017.

[7] Nannan Wang, Wenjin Zha, Jie Li, and Xinbo Gao. Back projection: An effective postprocessing method for gan-based face sketch synthesis. *Pattern Recognition Letters*, 2017.

[8] Nannan Wang, Xinbo Gao, Leiyu Sun, and Jie Li. Anchored neighborhood index for face sketch synthesis. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1–1, 2017.

[9] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, volume 2, pages 1398–1402. Ieee, 2003.

[10] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *Image Processing, IEEE Transactions on*, 15(2):430–444, 2006.

[11] Che-Chun Su, Lawrence K Cormack, and Alan C Bovik. Closed-form correlation model of oriented bandpass natural images. *Signal Processing Letters, IEEE*, 22(1):21–25, 2015.

[12] Wujie Zhou, Lu Yu, Weiwei Qiu, Yang Zhou, and Mingwei Wu. Local gradient patterns (lgp): An effective local-statistical-feature extraction scheme for no-reference image quality assessment. *Information Sciences*, s 397C398:1–14, 2017.

[13] Jinjian Wu, Weisi Lin, Guangming Shi, Leida Li, and Yuming Fang. Orientation selectivity based visual pattern for reduced-reference image quality assessment . *Information Sciences*, 351:18–29, 2016.

[14] Weilong Hou, Xinbo Gao, Dacheng Tao, and Xuelong Li. Blind image quality assessment via deep learning. *IEEE transactions on neural networks and learning systems*, 26(6):1275–1286, 2015.

[15] Fei Gao, Dacheng Tao, Xinbo Gao, and Xuelong Li. Learning to rank for blind image quality assessment. *IEEE transactions on neural networks and learning systems*, 26(10):2275–2290, 2015.

[16] Feng Qi, Debin Zhao, and Wen Gao. Reduced reference stereoscopic image quality assessment based on binocular perceptual information. *IEEE Transactions on Multimedia*, 17(12):2338–2344, 2015.

[17] Ming-Jun Chen, Che-Chun Su, Do-Kyoung Kwon, Lawrence K Cormack, and Alan C Bovik. Full-reference quality assessment of stereopairs accounting for rivalry. *Signal Processing: Image Communication*, 28(9):1143–1155, 2013.

[18] Junyong You, Liyuan Xing, Andrew Perkis, and Xu Wang. Perceptual quality assessment for stereoscopic images based on 2d image quality metrics and disparity analysis. In *Proc. of International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, AZ, USA*, 2010.

[19] Lutz Goldmann and Touradj Ebrahimi. 3d quality is more than just the sum of 2d and depth. In *IEEE International workshop on hot topics in 3D*, number EPFL-CONF-150334, 2010.

[20] Feng Shao, Weijun Tian, Weisi Lin, Gangyi Jiang, and Qionghai Dai. Toward a blind deep quality evaluator for stereoscopic images based on monocular and binocular interactions. *IEEE Transactions on Image Processing*, 25(5):2059–2074, 2016.

[21] Pieter Seuntiens, Lydia Meesters, and Wijnand Ijsselsteijn. Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric jpeg coding and camera separation. *ACM Transactions on Applied Perception*, 3(2):95–109, 2006.

[22] Willem JM Levelt. Binocular brightness averaging and contour information. *British journal of psychology*, 56(1):1–13, 1965.

[23] Jian Ding and George Sperling. A gain-control theory of binocular combination. *Proceedings of the National Academy of Sciences of the United States of America*, 103(4):1141–1146, 2006.

[24] Noboru Sugie. Neural models of brightness perception and retinal rivalry in binocular vision. *Biological Cybernetics*, 43(1):13–21, 1982.

[25] Gordon E Legge. Binocular contrast summationii. quadratic summation. *Vision Research*, 24(4):385–394, 1984.

[26] Lin Ma, Xu Wang, Qiong Liu, and King Ngi Ngan. Reorganized dct-based image representation for reduced reference stereoscopic image quality assessment. *Neurocomputing*, 215(C):21–31, 2016.

[27] Fredericka. A. Kingdom. Binocular vision: The eyes add and subtract. *Current Biology Cb*, 22(1):R22–4, 2012.

[28] Jiachen Yang, Yancong Lin, Zhiqun Gao, Zhihan Lv, Wei Wei, and Houbing Song. Quality index for stereoscopic images by separately evaluating adding and subtracting. *PloS one*, 10(12):e0145800, 2015.

[29] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[30] Chang Xu, Dacheng Tao, and Chao Xu. Robust extreme multi-label learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1275–1284, 2016.

[31] Tanaya Guha and Rabab K Ward. Image similarity using sparse representation and compression distance. *IEEE Transactions on Multimedia*, 16(4):980–987, 2014.

[32] Meng Liu, Chang Xu, Yong Luo, Chao Xu, Yonggang Wen, and Dacheng Tao. Cost-sensitive feature selection via f-measure optimization reduction. In *AAAI*, pages 2252–2258, 2017.

[33] Feng Shao, Kemeng Li, Weisi Lin, Gangyi Jiang, Mei Yu, and Qionghai Dai. Full-reference quality assessment of stereoscopic images by learning binocular receptive field properties. *IEEE Transactions on Image Processing*, 24(10):2971–2983, 2015.

[34] Lihuo He, Dacheng Tao, Xuelong Li, and Xinbo Gao. Sparse representation for blind image quality assessment. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1146–1153. IEEE, 2012.

[35] M Elad, M Aharon, and AM Bruckstein. The k-svd: an algorithm for designing of overcomplete dictionaries for sparse representations. *IEEE Trans. Image Process*, 15(12):3736–3745, 2006.

[36] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

[37] Jinjian Wu, Weisi Lin, Guangming Shi, and Anmin Liu. Perceptual quality metric with internal generative mechanism. *IEEE Transactions on Image Processing*, 22(1):43–54, 2013.

[38] Siwei Ma, Xiang Zhang, Shiqi Wang, Jian Zhang, Huifang Sun, and Wen Gao. Entropy of primitive: From sparse representation to visual information evaluation.

[39] Guangtao Zhai, Xiaolin Wu, Xiaokang Yang, Weisi Lin, and Wenjun Zhang. A psychovisual quality metric in free-energy principle. *IEEE Transactions on Image Processing*, 21(1):41–52, 2012.

[40] Yu-Hsun Lin and Ja-Ling Wu. Quality assessment of stereoscopic 3d image compression by binocular integration behaviors. *Image Processing, IEEE Transactions on*, 23(4):1527–1542, 2014.

[41] Lixiong Liu, Bao Liu, Hua Huang, and Alan Conrad Bovik. No-reference image quality assessment based on spatial and spectral entropies. *Signal Processing: Image Communication*, 29(8):856–863, 2014.

[42] Hamid R Sheikh, Zhou Wang, Lawrence Cormack, and Alan C Bovik. Live image quality assessment database release 2, 2005.

[43] Rui Song, Hyunsuk Ko, and C. C. Jay Kuo. Mcl-3d: a database for stereoscopic image quality assessment using 2d-image-plus-depth source. *Journal of Information Science and Engineering*, 31(5), 2014.

[44] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii. 2003.

[45] Seungchul Ryu, Dong Hyun Kim, and Kwanghoon Sohn. Stereoscopic image quality metric based on binocular perception model. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 609–612. IEEE, 2012.

[46] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, 2011.

[47] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE Transactions on Image Processing*, 21(8):3339–3352, 2012.

[48] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *Image Processing, IEEE Transactions on*, 21(12):4695–708, 2012.

[49] Feng Shao, Weijun Tian, Weisi Lin, Gangyi Jiang, and Qionghai Dai. Learning sparse representation for no-reference quality assessment of multiply-distorted stereoscopic images. *IEEE Transactions on Multimedia*, PP(99):1–1, 2017.

[50] Yaqi Lv, Mei Yu, Gangyi Jiang, Feng Shao, Zongju Peng, and Fen Chen. No-reference stereoscopic image quality assessment using binocular self-similarity and deep neural network. *Signal Processing Image Communication*, 47:346–357, 2016.