Blind Assessment for Stereo Images Considering Binocular Characteristics and Deep Perception Map Based on Deep Belief Network

Jiachen Yang[a], Yang Zhao[a], Yinghao Zhu[a], Huifang Xu[a], Wen Lu[b,*], Qinggang Meng[c]

[a] *School of Electrical and Information Engineering, Tianjin University, Tianjin, China*
[b] *School of Electronic Engineering, Xidian University, Xian, China*
[c] *Department of Computer Science, Loughborough University, Loughborough, UK*

Abstract

In recent years, blind image quality assessment in the field of 2D image/video has gained the popularity, but its applications in 3D image/video are to be generalized. In this paper, we propose an effective blind metric evaluating stereo images via deep belief network (DBN). This method is based on wavelet transform with both 2D features from monocular images respectively as image content description and 3D features from a novel depth perception map (DPM) as depth perception description. In particular, the DPM is introduced to quantify longitudinal depth information to align with human stereo visual perception. More specifically, the 2D features are local histogram of oriented gradient (HoG) features from high frequency wavelet coefficients and global statistical features including magnitude, variance and entropy. Meanwhile, the global statistical features from the DPM are characterized as 3D features. Subsequently, considering binocular characteristics, an effective binocular weight model based on multiscale energy estimation of the left and right images is adopted to obtain the content quality. In the training and testing stages, three DBN models for the three types features separately are used to get the final score. Experimental results demonstrate that the proposed stereo image quality evaluation model

[*] Corresponding author
Email address: luwen@xidian.edu.cn (Wen Lu)

has high superiority over existing methods and achieve higher consistency with subjective quality assessments.

Keywords: Blind stereoscopic image quality evaluation, deep belief network, wavelet transform, depth perception map, binocular weight

1. Introduction

With the development of 3D technology, 3D movies and TV have attracted global interest in many fields such as entertainment, medical treatment, games, architecture design and many others[10, 40, 43]. Human eyes as the final receiver of stereoscopic vision information, the quality of the display is critical to the quality of experience. However, in the process of image processing, transmission and reception, the quality of the stereo image will be degraded due to the introduction of various distortions. Therefore, it is necessary to raise an effective method for stereo image quality assessment.

Similar as the image quality assessment (IQA), the stereo image quality assessment (SIQA) can be divided into two categories: subjective quality evaluation by human eyes and objective quality evaluation by methods employed to simulate human subjective judgment. Since the human eye is the final acceptance, the subjective evaluation directly reflects the human visual system (HVS) and is regarded as the most precise evaluation method. However, the subjective evaluation requires many participants in the course of experiments and is time consuming. Therefore, it is unrealistic to implement subjective evaluation as a the real-time evaluation. Objective quality assessment can reliably predict quality, and has received great attention from numerous experts and scholars. According to the participation of the original image, objective quality assessment can be characterized as full-reference(FR), reduced-reference(RR) and no-reference(NR) approaches. In the FR metric, the quality is derived from the similarity between it and the distorted image. However, in most cases the original image cannot be available for the evaluation. In contrast, the RR model requires only a portion of the original information and no original image is needed in the NR model which is most challenging and promising in practical applications.

In recent decades, many state-of-art 2D-IQA metrics have been developed, such as structural similarity index measurement (SSIM)[36], the multi-scale Geometric Analysis method [12], visual information fidelity (VIF)[31], natural scene statistics based metrics [48, 18], and deep learning based metrics [8], etc. These metrics are quite effective on the quality evaluation of 2D images. However, when they are applied in the stereoscopic image evaluation, the later quality combination can not make the performance good.

Compared with the 2D image quality assessment, SIQA is more challening because of the depth information which is created by difference between left and right views. In earlier time, researchers tended to apply the 2D evaluation metric directly to the left and right viewpoint and disparity map, and then integrated them into the 3D quality score. Benoit *et al.* used the fusion of 2D quality metrics and the depth information to present a quality metric for the assessment of stereopairs [2], You *et al.* investigated the capability of some common 2D methods applying to the stereo image evaluation, while taking disparity into consideration [44]. However, these metrics can not obtain an satisfactory result for stereopairs evaluation, especially in the case of two views with asymmetric distortion. Subsequently, scholars found that the inaccurate evaluation was caused by ignoring of binocular visual characteristics, and then started research of image quality assessment based on HVS. Chen *et al.* designed a cyclopean image model which addressed binocular rivalry and disparity to evaluate the stereo images [5]. Shao *et al.* classified the stereoscopic image into non-corresponding, binocular fusion, and binocular suppression regions, and then extracted local amplitude and phase features from these regions to get the overall score [28]. In [11], the binocular integration behaviors (the binocular combination and the binocular frequency integration) are utilized as the bases of measuring the quality of 3D images. By verifying that the parameters of the GGD fit of luminance wavelet coefficients along with correlation values form excellent features, the STeReoscopic Image Quality Evaluator (STRIOE) was proposed in [17]. Yang *et al.* developed a evaluation model, that simulated the binocular fusion mechanism and the depth sensing mechanism respectively with the binocular “summation” channel and “difference” channel [41]. Based on the binocular fusion theory, Bensalma *et al.* [3] utilized simple and complex cells to reproduce the binocular signal and build the Binocular Energy Quality Metric (BEQM). Zhang *et al.* [46] proposed the 3D-MAD which estimated the quality degradation of the monocular views and the cyclopean image to obtained the quality score. Qi *et al.* predicted the quality score in terms of the binocular perceptual information (BPI), where the BPI is represented by the distribution statistics of visual primitives in left and right views’ images[22]. The results indicate that these metrics which are based on the HVS and take into consideration of both the stereo vision and the binocular characteristics have been have been verified to obtain effective index for stereoscopic images quality. However, all of them are FR metrics or RR metrics which need original images or part of the original information, and therefore limit their applications in most cases.

Recently, the NR SIQA attracted more interest because of its practicality owing to the no-need of the accessibility of the original 3D visual stimulus. To assess the perceptual quality, an SIQA method which deploys the binocular-rivalry related features was proposed in [4]. By exploring the relationship between the perceptual quality and the visual information, Ryu *et al.* proposed a no-reference metric based on blockiness, blurriness, visual saliency and the binocular perception model [25]. Appina *et al.* utilized the bivariate generalized Gaussian distribution (BGGD) of luminance and disparity coefficients of stereoscopic image to detect the features for training and testing [1]. Considering binocular energy response (BER) and binocular rivalry response (BRR), various binocular quality-predictive features are extracted which are applied in the NR-3D IQA. The complementary local patterns [49]. In [27], a blind quality assessment for stereoscopic images which constructed quality lookups to replace human opinion scores was proposed. This method is based on the characteristics of receptive fields (RFs) from perspective of dictionary learning. These

NR SIQA metrics have achieved comparable results with the FR metrics, and compensated the deficiency of the requirement of the original image.

Most of the above NR methods are based on well-known SVM or other machine learning methods which train and test the model by studying the characteristics of shallow layer. However, the structure of HVS is so complicated that hard to simulate by simple shallow learning structures. Recently, the deep learning is widely concerned and on account of its deep architecture similar to the human nervous system [13, 14] and its ability of learning features [45] to simulate the HVS, it has achieved some success in SIQA. Zhang *et al.* [47] employed different multiple convolutional neural network (CNN) with different inputs to get the CNN parameters by stacking three 3 × 3 convolution layers for training and being converted to quality score. Lv *et al.* [15] defined two indices, binocular Self-similarity and binocular integration, and then combined both of them with the five-layer DNN to predict quality. Shao *et al.* proposed a SIQA metric based on deep features resulting from four-layer-DNN, through investigating the interaction between monocular and binocular vision [29]. These metrics have shown an effective and robust performance compared with metrics using SVM, and thus can be the better models for learning and testing.

HVS is sensitive to different scales or orientations of the image, therefore many quality evaluation frameworks are proposed based on the wavelet transform. He *et al.* extracted NSS features in the wavelet domain and then represented them via sparse coding to obtain the final visual quality value [7]. Soroosh *et al.* proposed a RF IQA framework to obtain the quality perception scores in the discrete wavelet domain using the Haar wavelet [24]. Bovik *et al.* proposed other NSS-based distortion-generic approaches to NR IQA that statistically model images in the wavelet domain [18]. Considering the hybrid of curvelet, wavelet, and cosine transforms, Shen *et al.* proposed a no-reference image assessment model [32]. However, most of these metrics are 2D-IQA metrics and the wavelet transform is not applied in the SIQA widely.

In the real world, the visual perception of information and even depth is from a three-dimensional perspective. [23] has demonstrated the main cause of depth is binocular disparity, which reflects the horizontal positional difference between left and right retinal projections of a given point in space. In detail, the horizontal distance of human eyes is about 60mm, and therefore the two eyes view the image from slightly different angles. The early theory that the disparity induced depth sensing promoted the research for the depth information perception, and has dominated for a long period of time till now. The scholars study the perceived depth information from the parallax point of view [34, 35, 42]. Furthermore, ascribe the limitation of the research on the depth information, some scholars used the disparity map as an important depth perception factor in the stereoscopic image assessment [2, 44]. Based on the human depth perception, Jung *et al.* combined the disparity of salient object and the maximum for visual comfort (VC) prediction [9]. Chen *et al.* established the cyclopean image based on disparity and Gabor energy to simulate the binocular fusion and rivalry of human eyes [5]. The amount of quality evaluation models are derived on the strength of the cyclopean image and disparity map, some of them set up stereo visual model [11, 41, 22], and other metrics extracted features from the disparity map and the cyclopean image for training and testing to obtain the quality index [4, 15, 47, 1, 29].

The perception theory of depth information is still an open area, which is too complex to utilize a single theory for complete description. There are a lot of factors determining the depth information, for example the vanishing line, the size of the same object, and the occlusion relationship [21]. Therefore, the simple use of parallax to indicate the depth information is not precise enough. Moreover, all the disparity maps taken as the perception of the depth in the existing evaluation metrics reflect the horizontal disparity, which is only the cue of the depth information and not the direct reflection. Consequently, it is essential to develop a method which can directly reflect the depth perception.

Inspired by the previous work, a new NR IQA model is presented based on the wavelet transform and DBN to evaluate the quality of stereo images. The innovations of this paper are as follows:

(1) The vast majority of existing evaluation metrics regard absolute disparity

between two eyes as the depth perception, but ignore the subjective direct stereo perception impression caused by the positive and negative parallax of human eyes. We design our model inspired by the longitudinal stereoscopic perception and a novel depth perception map is derived to quantify longitudinal depth information to align with human eye perception, which can directly reflect the human intuitive feeling about the relative positions of the scenes and the screen generating the positive and negative parallax;

(2) The human eye is sensitive to the region of high contrast, such as edge, texture, and distortions often impact the high frequency components of stereo images. The conventional metrics usually extract the HoG features from the samples or the preprocessed images directly. To obtain the features which can reflect the stereo image quality more effectively, in this paper, the HoG features of the high frequency subband coefficients are extracted as the description of visual sensitivity. As far as we know, we are the first to use HoG features based on high frequency wavelet coefficients to evaluate the quality of the stereo image;

(3) Distinguishing the pre-existing works which directly take the average quality of two viewpoints to be the stereo image quality or simply study the binocular weight from single size images, a new binocular weighting system is proposed to obtain the content quality of the stereo image. Taking binocular characteristics into account, we study the multiscale perceptual characteristics of left and right images and a dynamic weighting system is designed.

The remaining sections of this paper are organized as follows. Section II presents a brief introduction of related work and the motivations. The overall 3D-IQA framework is described in Section III, Section IV presents the experiments conducted on the 3D-IQA databases and the performance analysis of the proposed model. Finally in Section V, the paper is conducted by a discussion and an outlook on the future work.

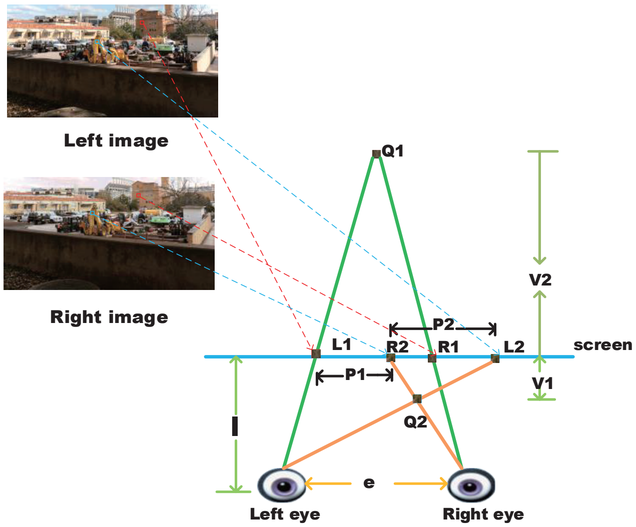


Figure 1: Longitudinal stereo perception theory.

2. Related Work

2.1. Longitudinal Depth Perception Theory

As is known to all, when watching movies, the human eye tends to regard the screen as the intermediate position. Therefore the most intuitive feeling is that some scenes are in the front of the screen, seem like the objects run out of the screen; while some scenes are in the back of the screen, seem like the objects run deep into the screen. The feeling of out screen and into screen is the intuitive perception of the depth. Inspired by this phenomenon, quantifying this perception is expected so as to transform the physical quantity into indices can be processed.

The parallax theory in [16] indicates that the stereoscopic perception of the human eye for the relative position between the object and the screen is derived from the longitudinal stereoscopic perception, as depicted in Fig.1.In this figure, L_1 and L_2 are the left pixel points respectively, R_1 and R_2 are the right pixel points respectively, V_1 and V_2 are the stereo depth respectively, P_1 is the horizontal disparity between L_1 and R_1, and P_2 is the horizontal disparity between L_2 and R_2. Additionally, l represents the viewing distance while e is the distance between the two eyes.

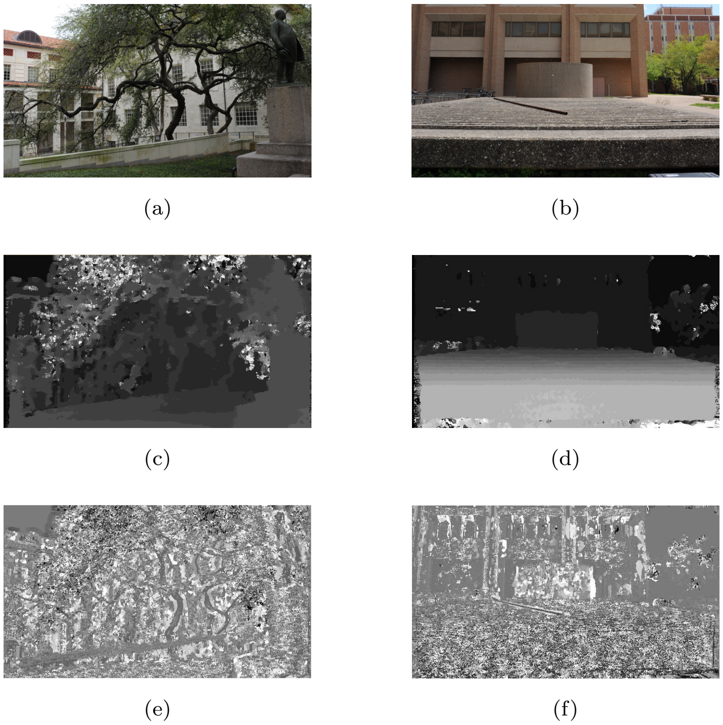
From Fig.1, it can be seen that the scene is in a relative back position such as the house, the right pixel point R_1 is located on the right of the left pixel point L_1, which is known as positive disparity. The object point has positive disparity with the negative stereo depth, which indicates the object point is located in the back of the screen, as the point Q_1. In another case, when the scene is in a relatively front position such as the crane, the right pixel point R_2 is located on the left of the left pixel point L_2, which is called as negative disparity. The object point has negative disparity with the positive stereo depth, which indicates the object point is located in the front of the screen, as the point Q_2. The stereo depth is calculated as follows:

$$V = \frac{lP}{P - e} \qquad P = \begin{cases} dis\tan ce\,(L, R) & R > L \\ -dis\tan ce\,(L, R) & L > R \end{cases} \tag{1}$$

where $R > L$ represents the right pixel point is located on the right of the left pixel point and $L > R$ is the converse of the meaning.

In this paper, the positive and negative disparity of the stereopairs can be computed based on the longitudinal stereo perception theory, which mainly inspired by the horizontal parallax algorithm. By translating the image to the left and right, the optimal disparity refers to the positive and negative parallax of the left and right images, which is relative to the traditional absolute disparity of the two views, can be found and the left and right pixels can be matched. When the optimal disparity is obtained through translating the right image to the right or translating the left image to the left, the optimal disparity is marked as negative value indicating negative parallax with positive depth. When the optimal disparity is obtained through translating the right image to the left or translating the left image to the right, the optimal disparity can be marked as positive value indicating positive parallax with negative depth.

On account of the unknown view distance, the positive and negative parallax are treated as the relative position indictor between the object and the display

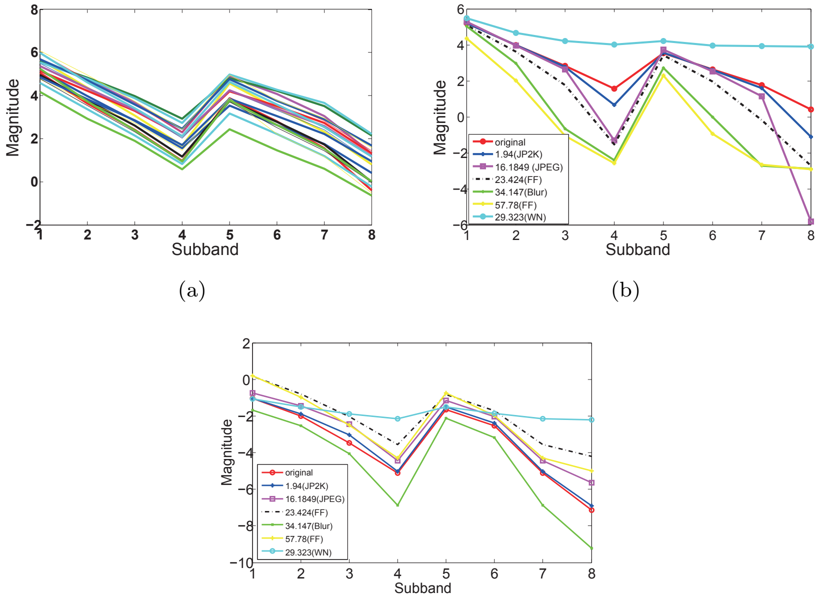


(a) (b) (c) (d) (e) (f)

Figure 2: (a) Original left image1 (b)Original left image2 (c) The corresponding disparity map of image1 (d) The corresponding disparity map of image2 (e)The corresponding DPM of image1 (f)The corresponding DPM of image2.

screen. After this process, the new disparity map can be contained whose values are normalized within $[0, 1]$. Closer to 1 means closer to the front of the screen while closer to 0 means closer to the back of the screen. Two original left images and corresponding disparity maps and DPM are shown in Fig.2.

In Fig.2(c) and (f), the lighter area indicates closer to the front of the screen while the darker area indicates closer to the back of the screen, and the depth perception map is consistent with subjective perception. Furthermore, compared with the corresponding disparity map (Fig.2(b) and (d)), the DPM shows more specific information in edge and texture regions.



(a)

(b)

(c)

Figure 3: (a) The magnitude characteristics of the 20 original left images (b) The magnitude characteristics of an original left image and corresponding distorted image with different subjective scores (c) The magnitude characteristics of an original DPM and corresponding distorted DPM with different subjective scores

2.2. NSS Features

The natural scene is the image or video which is obtained by using the high quality acquisition device in the the natural environment. Different from the text or the geometric figure obtained through artificial synthesis, the structure of the natural scene follows certain statistics characteristics which can be destructed by the introduction of distortions and thus many IQA metrics extract features for quality prediction based on this phenomenon. These methods mainly include: feature extraction of GGD and AGGD fitting based on frequency domain [18, 4], statistical properties of MSCN coefficients based on spatial domain [19], methods based on joint statistics of gradient magnitude and laplacian features [39] and others using MGGD to fit multivariate joint distribution statistical characteristics [1]. However, most of these methods are sensitive to different databases with different distortion types. Considering the drawbacks of the above approach, He *et al.* studied the exponential attenuation characteristics of the magnitude, variance and entropy in different wavelet subbands for 2D images based on sparse representation [7]. Specifically, the magnitude m_k is to encode the generalized spectral behavior, the variance v_k is to describe the fluctuations of the energy, and the entropy e_k is to represent the generalized information.

$$m_k = \frac{1}{N_k \times M_k} \sum_{j=1} \sum_{i=1} \log 2 \left| C_k \left(i,j \right) \right| \tag{2}$$

$$v_k = \frac{1}{N_k \times M_k} \sum_{j=1} \sum_{i=1} \log 2 \left| C_k \left(i,j \right) - m_k \right| \tag{3}$$

$$e_k = \sum_{j=1}^{N_k} \sum_{i=1}^{M_k} p \left[Ck \left(i,j \right) \right] \ln p \left[C_k \left(i,j \right) \right] \tag{4}$$

where $C_k\left(i,j\right)$ stands for the $\left(i,j\right)$ coefficient of the k-th subband; M_k and N_k define the size of the k-th subband,and represent the length and width respectively; and $p\left[\cdot\right]$ is the probability density function of the subband.

Motivated by this work, the m_k attenuation characteristics of the left image is investigated. First of all, the image is decomposed by wavelet transform in 4 scales and 3 orientations and the magnitude of all the subbands are calculated. Due to the similarity of the low-high and high-low subbands in the same scale, the mean value of the two subbands is taken as the magnitude of the low frequency subband. After that, the 8 magnitudes can be obtained as shown in Fig.3. Fig.3(a) shows the m_k of the 20 original left images. In the same orientation, the magnitudes of all the original left images decay exponentially. Fig.3(b) shows the m_k of an original left image and corresponding distorted images with different subjective scores. From Fig.3(b) it can be seen that in addition to the image with white noise distortion, the magnitude decay of almost images are accelerated with increasing DMOS value, especially in the fine scale. Due to the effect of distortion on the image is uniform, the magnitude for the image with white noise distortion shows slight oscillations. However, the introduction of distortion still changes the state of the original attenuation.

Through the analysis on magnitude, variance and entropy of stereoscopic images, it can be summarized that these features reflect the structure state of the image effectively.

2.3. The HoG Features

HoG features, which is composed of the gradient direction histogram of the local area of the image, are used for feature descriptor of object detection in computer vision, such as face recognition [6], vehicle identification [38] and pedestrian detection [46]. It results from that HoG features can be used for object detection is the representation and shape of a local object can be well described by a gradient or an edge. For better obtaining information and understanding the image, human eyes tend to observe areas with high contrast, such as edges, textures, and so on. Take into consideration that the image contrast is mainly caused by gradient changes, we utilize HoG algorithm to extract the gradient features.

Before the feature extraction, to reduce the influence of illumination, the image should be normalized. In this paper, the gamma correction is used as follows:

$$I\left(x,y\right)=\left|I\left(x,y\right)\right|^{\frac{1}{2}} \tag{5}$$

where I is the image, and (x, y) denotes the pixel location. In the following, the image should be divided into several blocks with several cells to calculate the gradient of the image in different directions. The gradient of pixel point (x, y) in horizontal direction and vertical direction are express as the Eq.(6) and (7) respectively:

$$G_x\left(x,y\right)=I\left(x+1,y\right)-I\left(x-1,y\right) \tag{6}$$

$$G_y(x,y) = I(x,y+1) - I(x,y-1) \tag{7}$$

Therefore the gradient amplitude and phase of the image respectively are:

$$G(x,y) = \sqrt{Gx(x,y)^2 + Gy(x,y)^2} \tag{8}$$

$$\theta(x,y) = \tan^{-}\left(\frac{Gy(x,y)}{Gx(x,y)}\right) \tag{9}$$

and the modification of the phase is

$$\theta(x,y) = \begin{cases} \theta(x,y) + \pi, \theta(x,y) < 0 \\ \theta(x,y), \theta(x,y) \geq 0 \end{cases} \tag{10}$$

At last, all the features in different directions and different blocks should be combined together to get the overall HoG features.

3. The proposed model

As discussed in previous sections, the perception of the human eyes for the stereoscopic image is based on the image content and the depth information. The assessment for image content reflects the quality level of the image. Meanwhile, the assessment for the depth information indicates the degree of the depth perception. As a result, a stereoscopic image evaluation metric is proposed based on wavelet transform as shown in Fig.4. As presented in this figure, we first set the depth perception map, and then extract 2D features from the left image and the right image to evaluate the image content separately. A novel binocular dynamic weighting system based on multiscale energy estimation of the two viewpoints is proposed to integrate into the overall stereo image content score. Moreover, the 3D features are extracted from the DPM to evaluate the stereo perception. It should be noted that all these works are based on wavelet transform. Finally, the two scores are combined to obtain the quality of the stereo image.

Table 1: The Explanation of All The Features

feature vector	Feature description
f_{NSS-l}	m_k,v_k,e_k coefficients of the left image;
f_{NSS-r}	m_k,v_k,e_k coefficients of the right image ;
f_{HOG-l}	HoG features of the left image;
f_{HOG-r}	HoG features of the right image;
f_{DPM}	m_k,v_k,e_k coefficients of the DPM;

3.1. Image Content Quality Aware Features

The left image and right image should initially be decomposed by wavelet transform in 4 scales and 3 orientations, and coefficients in 12 subbands can be gotten respectively. Taking the left image as an example, the magnitude in high-low, low-high, high-high subbands can be calculated by Eq.(2) for each scale. Due to the similarity of the high-low sub-band and low-high subband, the magnitude in low frequency is taken as the mean of the two subband coefficients, while the magnitude in high frequency is the result of the high-high subband and there are 8 magnitudes totally for the image. Similar to the magnitude, 8 variance coefficients and 8 entropy coefficients can also be calculate by Eq.(3) and (4) respectively. At last, there are 24 features for the left image and right image separately which are shown in Table 1 and are combined into a vector:

$$f_{NSS-l} = [m_{1l}, m_{2l}, ..., m_{8l}, v_{1l}, v_{2l}, ..., v_{8l}, e_{1l}, e_{2l}, ..., e_{8l}] \tag{11}$$

$$f_{NSS-r} = [m_{1r}, m_{2r}, ..., m_{8r}, v_{1r}, v_{2r}, ..., v_{8r}, e_{1r}, e_{2r}, ..., e_{8r}] \tag{12}$$

Human vision is a process of choice for that the human eye tends to concern areas of high contrast, such as edges or sharp regions, which include more visual information in the image. Consequently, if we desire to get valuable information

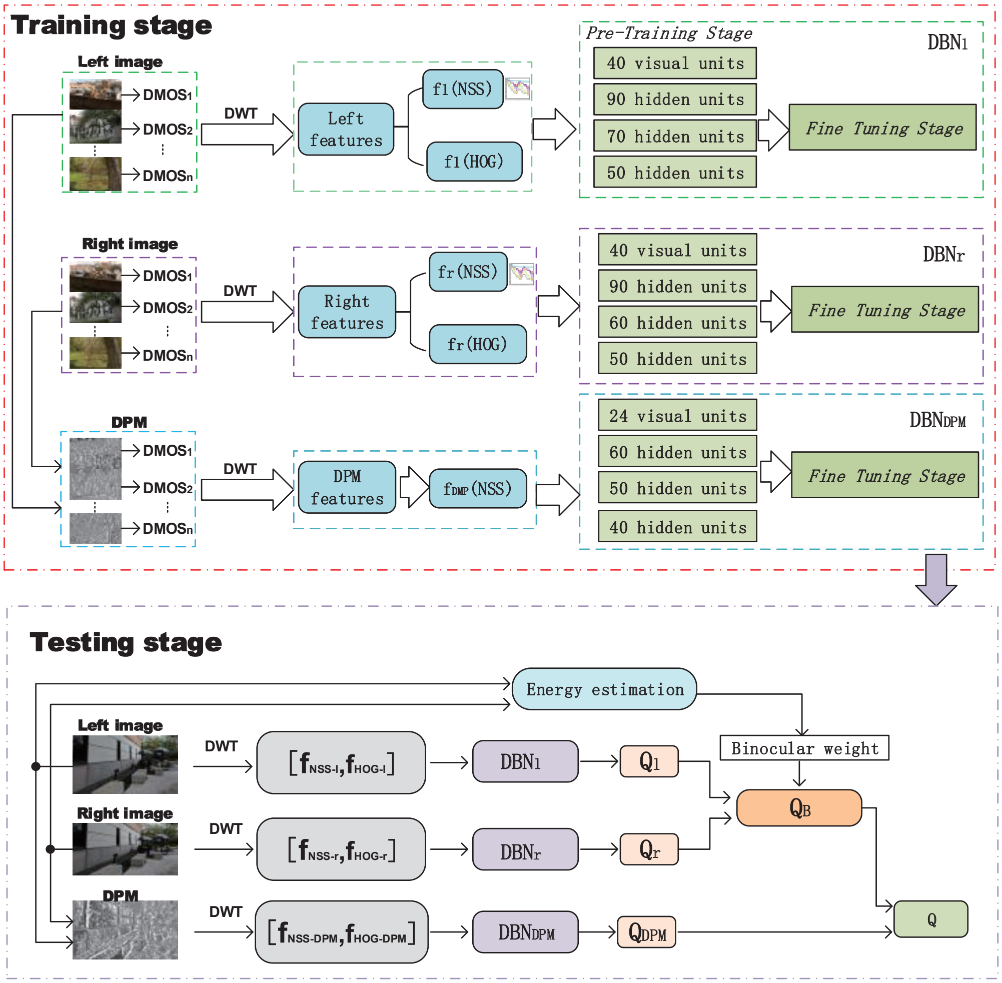


Figure 4: Framework of the proposed method

to describe the visual perception, the high frequency information should be paid more attention. As a local feature descriptor, HoG features are extracted to characterize the gradient statistics information of the image, and the gradient exists mainly in the edge region or the high frequency area. Therefore, HoG algorithm is utilized to extract features from the high frequency subband. The process has been described in section II. Here a cell element is composed of a block containing 8×8 pixels and the gradient direction is divided into 4 sections whose interval is $\frac{\pi}{2}$. For each cell, there are 4 HoG features in 4 angle intervals. To reduce the influence of the local light intensity and the contrast variation on the gradient intensity, a plurality of cell elements are combined into large and mutually connected blocks whose features should be normalized by Eq.(13):

$$f_{h-n} = \frac{f_h}{\sqrt{\|f_h\|_2^2 + C}} \tag{13}$$

where f_h is the HoG feature in each angle interal, $\|\cdot\|_2^2$ is the $L_2 - norm$, and C is a constant in case the denominator is zero.

Here 2×2 cells is taken as one block and thus there are 16 HoG features for each block. The HoG feature extraction is carried out in the high frequency subband in fine scale and the features are the average value of all blocks.

$$f_{hi} = \frac{1}{n_B} \sum_{j=1}^{n_B} f_{hi-n,j}, \qquad (hi = 1, 2, 3, ..., 16) \tag{14}$$

where f_{hi} is the HoG feature in each angle interval of every cell and n_B is the number of the blocks in each image. The 16 HoG features in the left image and right image are extracted respectively and they are depicted as

$$f_{HOG-l} = [f_{hl1}, f_{hl2}, f_{hl3}, ..., f_{hl16}] \tag{15}$$

$$f_{HOG-r} = [f_{hr1}, f_{hr2}, f_{hr3}, ..., f_{hr16}] \tag{16}$$

3.2. DBN Model Training

As shown in Fig.4, in the training stage, three 2D-DBN models including three hidden layers are trained by utilizing the left features, the right features and the DPM features respectively. Specifically, the DBN models learn the relationship between the three types of features and corresponding subjective scores through two stages: unsupervised pre-training stage and supervised fine tuning stage. In the first stage, for the left features and the right features, the structure of DBN is 40-90-70-50 nodes per layer, while there is a 24-60-50-40 nodes per layer for the DBN structure of the DPM, and the learning rate of all the three DBN model is set to 0.0005. The batch size is set to 1 due to the small number of training samples. Meanwhile, if the number of the epoch reaches to 200, the pre-training process is stopped. In order to obtain accurate prediction results, there must be a fine tuning stage to minimize the prediction error. In this stage, we refer to the fine tuning process in [29] and the human opinion scores are presented to be the supervised label to reduce the prediction error. More concretely, let f_i represents the output vector of the third hidden layer and s_i is the DMOS value. Then the purpose of the fine tuning is to minimize the cost function when given the training sets $\{(f_1, s_1), (f_2, s_2), ..., (f_N, s_N)\}$. Here the cost function is expressed as

$$\hat{\phi} = \arg\min \sum_{i=1}^{N} (s_i - \phi(f_i))^2 \tag{17}$$

where N is the number of the training samples. Meanwhile, in the fine-tuning phase, we implement a regression process to predict the quality of the image. Then the solution to minimize the cost function is follow as

$$\hat{\phi} = \sum_{i=1}^{N} w_i K(f, f_i) + b \tag{18}$$

where w_i is the weight matrix for the ith sample, b is the bias value, and $K(\bullet)$ is the reproducing kernel. Thus, the optimal w_i and b of the objective function are obtained by the way of back propagation error, and finally the objective scores which are closest to the subjective evaluation are obtained by several epoches and the deadline epoch number is 170. Finally, there are three 2D-DBN models (DBN_l, DBN_r and DBN_{DPM}) trained to the stereo image quality.

3.3. Image Content Quality Pooling

With the extracted features and corresponding subjective DMOS values, DBN is utilized to set regression model to obtain the objective score. Before use two models to predict quality scores of the left image and the right image separately, the NSS features and HoG features should be connected as a single feature vector. The quality scores of the two images are predicted by

$$Q_l = DBN_l([f_{NSS-l}, f_{HOG-l}]) \tag{19}$$

$$Q_r = DBN_r\left([f_{NSS-r}, f_{HOG-r}]\right) \tag{20}$$

The quality scores of the two images can be combined into an overall quality score. However, Due to the existence of the binocular characteristics, especially the binocular rivalry, the human vision system does not simply take the average quality of two viewpoints to be the stereo image quality. Therefore a new weight scheme is adopted.

According to the theory that high energy regions contain more visual information and are dominant in visual perception [37], the monocular image with higher energy should be more emphasized. Specifically, an image is divided into multiple scales, by employing an iterative low-pass filtering and downsampling procedure. Subsequently, the energy of every subband on the left and right images should be calculated. In the proposed method, the energy is obtained by summing the local variances using an 11×11 circular-symmetric Gaussian weighting function $w = \{\omega i | i = 1, 2, ..., N\}$, with standard deviation 1.5 samples, normalized to unit sum ($\sum_{i=1}^{N} \omega i = 1$) [36]. The local energy is then calculated by

$$e_{in} = \left(\sum_{i=1}^{N} \omega_i (x_i - \mu_{in})^2\right)^{\frac{1}{2}} \tag{21}$$

where

$$\mu_{in} = \sum_{i=1}^{N} \omega_i x_i \tag{22}$$

is the local mean value. The energy maps of the two images in different scales are shown in Fig.5. From Fig.5(b) and (c), obviously, the energy distribution of the two images in the same scale is different, especially the distorted level in these regions is different and thus triggering binocular rivalry. These two figures also give us another message that the difference of the two energy maps in the fine scale is clear. Meanwhile, it is evident that the difference between the left and right energy map on the fourth scale is very small and even cannot be distinguished. So there is few use with high-order scale energy map for calculating the weight factor. Moreover, if the number of scales is too small, it will result in the loss of difference information of the two views and the calculation is not accurate. So the number of scales we choose in this paper is 4. After calculating the local energies in different scales, the energy of the two images are computed as

$$e_l = \frac{1}{n_s \times M} \sum_{i=1}^{n_s} \sum_{j=1}^{M} e_{in,j-l} \tag{23}$$

$$e_r = \frac{1}{n_s \times M} \sum_{i=1}^{n_s} \sum_{j=1}^{M} e_{in,j-r} \tag{24}$$

where $e_{in,j-l}$ and $e_{in,j-r}$ are the local energies of the two image respectively, M is the number of pixels in the energy map and n_s is the number of scales.

Subsequently, the weight of the left and right image separately are

$$w_l = \frac{e_l^2}{e_l^2 + e_r^2} \tag{25}$$

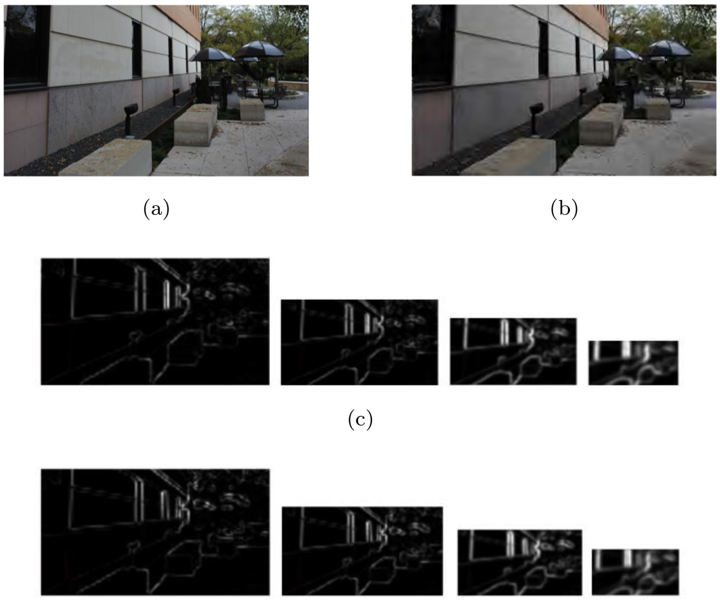
$$w_r = \frac{e_r^2}{e_l^2 + e_r^2} \tag{26}$$

then the quality of the image content is represented by

$$Q_B = w_l Q_l + w_r Q_r \tag{27}$$

3.4. Stereo Image Quality

As mentioned in earlier, in order to quantify the longitudinal depth information of human eye perception which expresses the most direct feeling of stereo vision, the concept of DPM is proposed and the process is described in section II. In the stage of evaluating the stereo perception, the magnitude, variance and entropy are utilized as stereo NSS features to predict the depth perception quality. Before the evaluation, the DPM is decomposed by wavelet transform in 4 scales and 3 orientations. The feature extraction procedure is the same as extracting NSS features on the left and right images and the m_k of an original DPM and corresponding distorted DPM with different subjective scores are



(a)

(b)

(c)

(d)

Figure 5: (a)The distorted left image (b)The distorted right image (c)The energy map of the light image in four scale (d) The energy map of the right image in four scale

shown in Fig.3(c). As shown in this figure, the magnitude of DPM increases and the magnitude decay in fine subband becomes slower with the DMOS increases, except for the DPM of the stereo images with WN and blur distortion. The magnitude for the DPM of the stereo image with WN distortion shows slight oscillations due to the uniformity of the distortion. Meanwhile, the introduction of blur distortion makes high-frequency information lost and disparity matching produce great differences compared with original stereo image, and this accelerates the magnitude attenuation. In general, the statistic characteristic of the DPM can reflect different distortion degree of stereo image and the features can be represented as

$$f_{DPM} = [m_1, m_2, ..., m_8, v_1, v_2, ..., v_8, e_1, e_2, ..., e_8] \tag{28}$$

Table 2: The performance of the proposed metric compared with several metrics on LIVE 3D phase

Criteria		LIVE I			LIVE II		
		$PLCC$	$SROCC$	$RMSE$	$PLCC$	$SROCC$	$RMSE$
2D-extended	DIIVINE [18]	0.7601	0.8051	11.2964	0.7257	0.7149	8.0162
	BLIINDS-II [26]	0.9211	0.9067	7.2314	0.8653	0.8467	5.1132
	BRISQUE [19]	0.9328	0.9236	5.1254	0.8631	0.8328	7.0136
3D-FR	Chen's scheme [5]	0.9267	0.9257	7.6931	0.9010	0.8930	10.5800
	STRIOE [17]	0.9275	0.9223	-	0.9019	0.8920	-
	Lin's scheme [11]	0.8645	0.8559	10.9898	0.6584	0.6375	8.4956
	Shao's scheme1 [30]	0.9350	0.9251	5.8155	0.8628	0.8494	5.8155
3D-NR	Shao's scheme2 [27]	0.9071	0.8961	-	0.9071	0.8961	-
	Appina's scheme [1]	0.9170	0.9110	6.5980	0.8450	0.8880	7.2790
	Shao's scheme3 [29]	**0.9565**	**0.9449**	**4.7552**	0.9265	0.9106	4.3300
	Proposed(SVM)	0.9541	0.9459	4.9186	0.9313	0.9152	4.0893
	Proposed(DBN)	0.9556	0.9437	4.9171	**0.9335**	**0.9206**	**4.0053**

and the depth perception score is obtained by a DBN model as mentioned earlier which is expressed as

$$Q_{DPM} = DBN_{DPM}\left(f_{DPM}\right) \tag{29}$$

The stereo image quality is the result of the interaction between the image content quality and depth perception. A linear equation is adopted to calculate the overall score of the stereo image, that is [20], LIVE 3D image database phase II [4] and MCL database [33].

$$Q = Q_B + \alpha \cdot Q_{DPM} \tag{30}$$

where α is used to modify the proportional relationship between the two scores. Experiment shows that $\alpha = 0.3$ is the most effective value which will be discussed in section IV.

4. Experimental Results and Analysis

In this section, in order to verify the effectiveness and robustness of the proposed objective evaluation metric, we analyze its performance on the following three publicly available stereo image databases: LIVE 3D image database phase I [5], LIVE 3D image database phase II [4] and MCL database [33].

4.1. LIVE 3D image database

LIVE 3D image database phase I consists of 20 original stereopairs and 365 corresponding symmetrically distorted stereopairs (80 each for $JP2K$, $JPEG$, WN and FF; 45 for $Blur$) while LIVE 3D image database phase II consists of 8 original stereopairs, and 120 symmetrically and 240 asymmetrically distorted stereopairs which are based on the same distorted types with LIVE 3D image database phase I. These two databases both have co-registered human scores in the form of DMOS.

4.2. Performance measure

To verify the performance of the proposed metric, three evaluation criteria are chosen: Pearson Linear Correlation Coefficient ($PLCC$), Spearman Rank-order Correlation Coefficient ($SROCC$), and Root Mean Square Error ($RMSE$), between the objective scores after nonlinear regression and the subjective scores. The five-parameters logistic mapping function is adopted in the nonlinear regression and expressed as:

$$Q_p = \beta_1 \cdot \left[\frac{1}{2} - \frac{1}{1 + \exp\left(\beta_2 \cdot (x - \beta_3)\right)}\right] + \beta_4 \cdot x + \beta_5 \tag{31}$$

where β_1, β_2, β_3, β_4 and β_5 are determined by using both the subjective scores and the objective scores. In addition, a value approaching 1 for $PLCC$ and

Table 3: The Performance with Proposed Weight Compared with Average Weight

		PLCC	*SROCC*	*RMSE*
LIVE I	Average	0.9543	**0.9482**	**4.7012**
	Proposed weight	**0.9556**	0.9437	4.9171
LIVE II	Average	0.9128	0.8997	4.0543
	Proposed weight	**0.9335**	**0.9206**	**4.0053**

SROCC and a value approaching 0 for *RMSE* indicate good performance in the term of correlation with human opinion.

In the process of score prediction, the image samples in each database were randomly divided into two parts. Specifically, the first part includes 80% image samples which were used for training and the rest 20% image samples were used for testing. In order to ensure the robustness of the proposed approach, 1000 iterations of the training and testing procedure 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 quality score.

4.3. Overall Performance in 3D Image Databases

For better demonstration of the effectiveness of the proposed metric, several existing state-of-art metrics for 3D images are chosen as comparison on LIVE phase I and II, including three 2D-extended metrics (DIIVINE [18], BLIINDS-II [26], BRISQUE [19]), four FR metrics (Chen's scheme [5], STRIOE [17], Lin's scheme [11], Shao's scheme1 [30]) and three NR metrics (Shao's scheme2 [27], Appina's scheme [1], Shao's scheme [29]). It should be noted that for the 2D-extended metrics, feature vectors are extracted separately for the left and right images. The average value of feature vectors is computed considering weighting factors to obtain the final feature vector for training. Similar to the proposed metric, for the three 2D-extended metrics, we also randomly divide a database

Table 4: Cross-database Performance Compared with Other Metrics

Criteria	LIVE I (training)/LIVE II (testing)			LIVE II(training)/LIVE I (testing)		
	PLCC	*SROCC*	*RMSE*	*PLCC*	*SROCC*	*RMSE*
DIIVINE [18]	0.5967	0.5238	8.491	0.513	0.4897	12.3451
BLIINDS-II [26]	0.7927	0.7532	6.9816	0.8435	0.8301	8.6513
BRISQUE [19]	0.792	0.7643	6.8967	0.8958	0.8865	7.4573
Shao's scheme [29]	0.7791	0.7514	-	**0.8936**	**0.8917**	-
Proposed	**0.8521**	**0.8493**	**5.9876**	0.8687	0.8601	8.0012

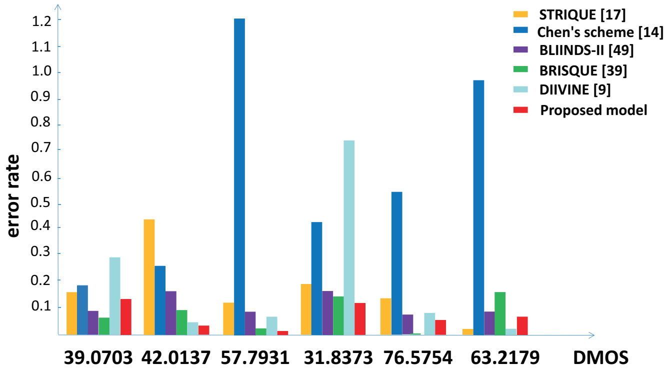


Figure 6: Error rate statistics of different model for distorted stereo pairs with different DMOS

into independent training and testing subsets with 80% for training and 20% for testing. Each partition is randomly conducted 1000 times using SVR on each dataset and the median values are computed. The $PLCC$, $SROCC$ and $RMSE$ results are listed in Table 2 and the best indictors are emphasized by the bold fonts. As shown in Table 2, compared with other metrics, the proposed metric delivers better correlation with the MOS values and performs the best on LIVE 3D image database phase II. Moreover,it also has strong competitiveness compared with Shao's metric [29] which achieves the top performance on LIVE

3D image database phase I. The reason for that the performance of our metric is slightly lower than Shao's is that the parameters of DBN model are not optimal, which is one of the problems we will focus on in the future. More precisely, all the metrics usually perform well on LIVE phase I which simply contains symmetrical distorted stereopairs, but have relatively worse performance on LIVE phase II. Although BRISQUE [19], Chen's metric [5] and Shao's metric [29] have good ability to predict quality on LIVE phase II, the proposed metric is more promising than them. Meanwhile, the performance of the proposed model using SVM are also listed in Table 2. From the comparative results, it is obvious that the index is lower than the model using DBN. Since the HVS is a complex hierarchical system and the high level feature is characterized by the combination of low level features. DBN just simulates the mechanism of the HVS and makes up the disadvantage of SVM that the shallow network can not abstract and optimize the features. Therefore, the model using DBN is consistent with the HVS and can have the better prediction performance.

To further verify the effectiveness and stability of the proposed model, we conducted an error rate statistics experiment to compare the prediction error of the proposed model with other models for different distorted stereo pairs and the result is shown in Fig.6. What needs to be explained here is that the error rate is the relative error with the actual prediction score. It is obvious that most metrics prediction accuracy is not stable and none of the metrics can has most outstanding performance for all distorted stereopairs. For example, BRISQUE has the lowest prediction error rate for the stereoscopic image with DMOS 76.5754, but has the second highest error rate for the stereoscopic image with DMOS 63.2179. And other models have the similar situation. However, the proposed model has the lowest error rate for the stereoscopic images with DMOS 42.0137, 57.7931, 57.7931 and even its prediction effect is not the best for other distorted images, but there are also second or third low error rates. Compared with other models with unstable error rates, our model has relatively stable prediction performance.

The overall performance results elaborate that the proposed metric is more convincing to evaluate the stereoscopic images especially with asymmetrical distortion.

4.4. Cross-Database Performance

We have verified the performance on individual 3D databases. However, the samples of training and testing are selected from the same dataset and thus this approach is not sufficient to support the generality and stability of the proposed evaluation model. In order to exclude the impact of database dependence, in this section, we do the experiment that the proposed metric is trained on one dataset and tested on another dataset using LIVE phase I and phase II. The comparison of the experimental results with several other metrics are listed in Table 4. Due to the difference of the content and the distortion of the stereo images on both datasets, the performance of each metric has a significant decline compared with the results on individual dataset, but the proposed metric still has a relatively better prediction ability among the five metrics. The result in Table 4 demonstrate that when the metrics are trained on LIVE phase I and tested on LIVE phase II, the proposed metric has the best performance. When trained on LIVE phase II and tested on LIVE phase I,the proposed metric also delivers competitive performance and the indictors are very close to the top two, BRISQUE [19] and Shao's metrics [29]. As mentioned earlier, the LIVE phase II contains both symmetrical and asymmetrical distorted stereopairs but LIVE phase I only contains symmetrical distorted stereoscopic images, thus the trained model on LIVE phase II is more complete than on LIVE phase I. The majority metrics perform well on LIVE phase I when the model is trained on LIVE phase II while poor performance on LIVE phase II when the model trained on LIVE phase I is deployed. Since the proposed metric has lower dependence on image content and distortion type, so it can achieve good results in both two cases and deliver higher generalization and robustness capability.

4.5. Performance on Individual Distortion Type

In the previous subsection, we discussed the overall performance of the metric. However, good overall performance does not always mean good performance

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

Criteria		BRISQUE [19]		BLIINDS-II [26]		Proposed	
		PLCC	*SROCC*	*PLCC*	*SROCC*	*PLCC*	*SROCC*
LIVE I	*JPEG*	0.7654	0.7049	0.792	0.7389	**0.8243**	**0.7681**
	JP2K	0.9263	0.8344	0.9233	0.8515	**0.9424**	**0.8971**
	WN	0.9466	0.9074	0.9222	0.8941	**0.9536**	**0.9294**
	FF	**0.864**	**0.801**	0.8023	0.6809	0.7893	0.6853
	Blur	0.9595	0.8667	0.9469	0.8667	**0.9634**	**0.9167**
	All	0.9328	0.9236	0.9211	0.9067	**0.9556**	**0.9437**
LIVE II	*JPEG*	0.8010	0.6905	0.7972	0.6911	**0.8671**	**0.8064**
	JP2K	0.7442	0.6911	0.7960	0.7509	**0.8855**	**0.8593**
	WN	0.8713	0.8252	0.8648	0.8195	**0.8873**	**0.8637**
	FF	**0.9406**	**0.9099**	0.9357	0.9055	0.9162	0.8769
	Blur	0.9337	0.8565	0.9568	0.8198	**0.9877**	**0.8341**
	All	0.8631	0.8328	0.8653	0.8467	**0.9335**	**0.9206**

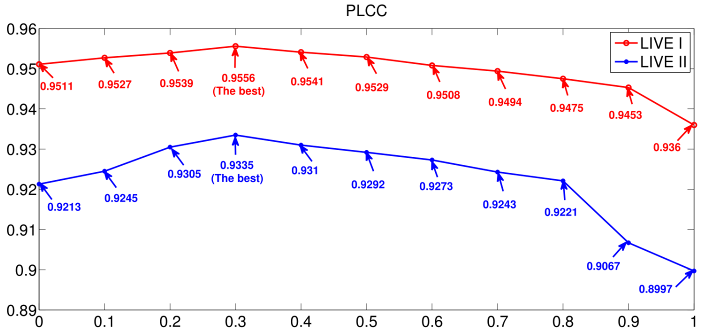


Figure 7: PLCC indicators of the metric with different α

for individual distortion type. To comprehensively investigate the evaluation ability of the proposed metric with different types of distortions, 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. Similarly, the top performance are highlighted with boldface. As shown in this table, the proposed metric has the best performance on most distortion type, whether for symmetric distortion stereopairs or asymmetric distortion stereopairs. For FF distortion, the BRISQUE metric has the best predictive ability. And since our model is based on image statistical properties and gradient characteristics, the prediction performance for FF is slight lower than the other two metrics. 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. Overall speaking, the proposed metric has the best performance on most distortion types and its versatility is verified.

4.6. Impact of The Weight Model

To demonstrate the significance of the proposed weight model, a compared experiment should be conducted where the average score of the left and right images is used as the content quality score, the results are listed in Table 3. For the symmetrical distorted stereopairs, the energy of the two images is almost equal due to the same type and degree of distortion, therefore the performance with the proposed weight is almost equal to the performance with the average weight on LIVE phase I. However, for the asymmetrical distorted stereopairs, the distortion type or degree of the left and right views is much different, and the image content in the local area is mismatching. According to the binocular rivalry principle, the proportion of the two images is not average in the process of stereoscopic perception and as a consequence, the performance with the average weight on LIVE phase II is not ideal. With the asymmetrical distortion, the energy of left and right images is quite different. Moreover, high energy regions contain more visual information and are the predominance in visual perception.

Table 6: The Performance with or without The DPM

		PLCC	*SROCC*	*RMSE*
LIVE I	2D features	0.9487	0.9398	5.1225
	2D features+DPM	**0.9556**	**0.9437**	**4.9171**
LIVE II	2D features	0.9203	0.9123	4.2631
	2D features+DPM	**0.9335**	**0.9206**	**4.0053**

It can be observed that the result on LIVE phase II verifies this hypothesis. The metric using the proposed weight with multiscale energy has better performance with PLCC index improving 0.0207, compared with the metric using the average weight.

4.7. Impact of The DPM

As mentioned earlier, the intuitive perception of human eyes viewing the stereoscopic image is that the scene out of the screen or into the screen. Based on this idea, we quantify this perception and establish the depth perception map. Different from the disparity map which is based on the horizontal parallax of the left and right images to indirectly reflect the stereo image depth information, the depth perception map is obtained by quantifying longitudinal depth information which can be more directly and accurately to reflect the impact of distortion on deep information. To verify the effectiveness of the depth perception map, a comparative experiment is implemented to explain the impact of the DPM and the results are listed on Table 6 which shows the performance of 2D features with DPM and 2D features only on LIVE 3D image database phase I and phase II.

The proposed metric with the DPM has the best performance with the highest performed metrics which are emphasized with bold fonts on both databases and this effectiveness is more prominent on LIVE phase II. The result on LIVE phase II indicates that compared with on symmetric distortion, the impact of the depth perception on the asymmetric distortion is even greater. Since the proposed 2D features are mainly used to reflect the content quality of images, the loss of content caused by asymmetric distortion is more complex and serious than the loss of symmetric distortion, which makes the quality evaluation more difficult and the overall performance on LIVE phase I is superior on LIVE phase II. As mentioned in section 2.1, the DPM can directly reflect the effects of distortion on depth information, and it also shows more details in edge and texture regions, so compared with the performance on LIVE phase I, the performance of 2D features with DPM on LIVE phase II improved more significantly. All the performance indictors verify the effectiveness of depth perception maps for representing depth information and it is significant to evaluate the depth perception. In addition, α is used in Eq.(30) to modify the proportional relationship between the image content quality score and the DPM quality score. The $PLCC$ indictors with different α are shown in Fig.7 and the top indictors are marked with "The best". The value $\alpha = 0.3$ is chosen as the final modification factor with which the metric has the best performance.

4.8. Impact of the HoG features

The HVS tends to focus on areas of high contrast and thus the HoG features of high frequency coefficients are detected as the description of the visual characteristic. In this section, the performance of the proposed metric using HoG features is compared with the metric where HoG features are absent and the results are listed in Table 7. Since NSS features can reflect the structure state of the image but only utilizing them to describe the content information of stereoscopic images is not comprehensive, so the performance of the algorithm is not superior. In this paper, besides the NSS feature, the proposed metric directly extracts the HOG features from the high frequency subband to describe the edge or high contrast region. The experiment results show that the metric with HoG features behaves the better, which is in accord with the phenomenon that high contrast areas attract more attention from people, as shown in $PLCC = 0.9556$ and $PLCC = 0.9335$ on LIVE phase I and on LIVE phase II.

4.9. Testing with other database

MCL database [33] consists of nine image-plus-depth original stereo scenes and corresponding 684 distorted stereopairs. In this database, several distortions are applied to either the texture image or the depth image. This database includes six types of distortion: Gaussian blur, additive white noise, down sampling blur, $JPEG$ and $JP2K$ compression and transmission error, and the human scores are in the form of MOS for all the stereopairs.

Meanwile, the Waterloo-IVC 3D Image Quality database Phase I is created from 6 pristine stereoscopic image pairs and there is 330 distorted stereoscopic images while the Waterloo-IVC 3D Image Quality database Phase II is created from 10 pristine stereoscopic image pairs and there are 460 distorted stereoscopic images. All the images of the two stereo databases are altered by three types of distortions: additive white Gaussian noise contamination, Gaussian blur, and JPEG compression. Each distortion type had four distortion levels.

Table 8 listed the performance of the proposed metric compared with three 2D-extended metrics and other 3D image quality evaluation metrics on MCL database while the top indictors are bolded. From this table, it can be seen that the proposed metric holds the best performance in predicting the quality of the stereopairs on MCL database and shows a stronger robustness for a variety of distorted images. About the performance results on the Waterloo-IVC, we can easily conclude from the Table 9 and the top indictors are bolded. From this table, the performance of the proposed model is better than the most methods which are listed in the Table 9. Although the performance of the proposed model is worse than the Wang's scheme [37] in Waterloo-IVC I, it has competitive result compared with Wang's scheme. At the same time, the proposed model has the best performance in Waterloo-IVC II.

5. Conclusion

In this paper, we have introduced a new stereoscopic image quality assessment framework based on DBN. The contributions of this paper are: (1) HoG

Table 7: The Performance of The Metric with or without HoG features

		$PLCC$	$SROCC$	$RMSE$
LIVE I	without HoG features	0.9423	0.9346	5.1294
	Proposed metric	**0.9556**	**0.9437**	**4.9171**
LIVE II	without HoG features	0.9276	0.9143	4.1221
	Proposed weight	**0.9335**	**0.9206**	**4.0053**

Table 8: The Performance on MCL Database

Model		$PLCC$	$SROCC$	$RMSE$
2D-extended metrics	DIIVINE [18]	0.8432	0.8361	1.0798
	BLIINDS-II [26]	0.8163	0.7994	1.1639
	BRISQUE [19]	0.812	0.8036	1.3213
3D metrics	Chen's scheme [5]	0.8149	0.8056	1.4123
	Lin's scheme [11]	0.7532	0.6726	1.7229
	Shao's scheme [29]	0.9138	0.9040	1.0233
	Proposed	**0.9321**	**0.9234**	**1.0023**

features of the high frequency subband coefficients were extracted as the description of visual properties and used for the first time to evaluate the quality of stereo images; (2) A novel depth perception map is derived to quantify longitudinal depth information of human eye perception; (3) Taking binocular properties into account, a new binocular weighting system is employed based on multi scales and multi orientations sensing characteristics of the human eye. The major advantage of the proposed framework is that it is applicable to both symmetric distortions and asymmetric distortions when predicting the image quality, especially for asymmetrical images, the superiority is more obvious. Meanwhile, the proposed DPM expresses the intuitive longitudinal depth per-

Table 9: The Performance on Waterloo-IVC Phase I and Waterloo-IVC Phase II Database

Criteria	IVC Phase I		IVC phase II	
	$PLCC$	$SROCC$	$PLCC$	$SROCC$
You [44]	0.7125	0.5968	0.6817	0.5873
Benoit [2]	0.6797	0.5852	0.5507	0.4595
Chen [5]	0.7337	0.6815	0.6130	0.5781
Wang [37]	**0.9300**	**0.9177**	0.8918	0.8687
Proposed	0.9116	0.9152	**0.9085**	**0.9093**

ception information and is consistent with subjective perception. In our future work, we continue to focus our research on the stereo perception based on the HVS and propose a more accurate SIQA method.

6. References

[1] B. Appina, S. Khan, S. S. Channappayya, No-reference stereoscopic image quality assessment using natural scene statistics, Signal Processing Image Communication 43 (2016) 1-14.

[2] A. Benoit, P. L. Callet, P. Campisi, R. Cousseau, Quality assessment of stereoscopic images, Eurasip Journal on Image and Video Processing 2008 (1) (2009) 1-13.

[3] R. Bensalma, M. C. Larabi, A perceptual metric for stereoscopic image quality assessment based on the binocular energy, Multidimensional Systems and Signal Processing. 24 (2) (2013) 281-316.

[4] M. J. Chen, L. K. Cormack, A. C. Bovik, No-reference quality assessment of natural stereopairs, IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society. 22 (9) (2013) 3379-3391.

[5] M. J. Chen, C. C. Su, D. K. Kwon, L. K. Cormack, A. C. Bovik, Full-reference quality assessment of stereopairs accounting for rivalry, Signal Processing Image Communication. 28 (9) (2013) 1143-1155.

[6] P. Y. Chen, C. C. Huang, C. Y. Lien, Y. H. Tsai, An efficient hardware implementation of hog feature extraction for human detection, IEEE Transactions on Intelligent Transportation Systems 15 (2) (2014) 656-662.

[7] L. He, D. Tao, X. Li, X. Gao, Sparse representation for blind image quality assessment, IEEE Conference on Computer Vision and Pattern Recognition. 23 (10) (2012) 1146-1153.

[8] J. Kim, S. Lee, Deep learning of human visual sensitivity in image quality assessment framework, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, 1969-1977.

[9] C. Jung, S. Wang, Visual comfort assessment in stereoscopic 3d images using salient object disparity, Electronics Letters 51 (6) (2015) 482-484.

[10] F. Li, F. Shao, Q. Jiang, R. Fu, G. Jiang, M. Yu, Local and global sparse representation for no-reference quality assessment of stereoscopic images, Information Sciences 422 (2017).

[11] Y. H. Lin, J. L. Wu, Quality assessment of stereoscopic 3d image compression by binocular integration behaviors., IEEE Transactions on Image Processing 23 (4) (2014) 1527-1542.

[12] M. Liu, X. Yang, Y. Shang, Image quality assessment based on multi-scale geometric analysis, IEEE Transactions on Image Processing. 18 (7) (2009) 1409-1423.

[13] Z. Li, J. Tang, Weakly supervised deep metric learning for community-contributed image retrieval, IEEE Transactions on Multimedia 17 (11) (2015) 1989-1999.

[14] Z. Li, J. Tang, Weakly Supervised Deep Matrix Factorization for Social Image Understanding, IEEE Transactions on Image Processing, 26 (1) (2017) 276-288.

[15] Y. Lv, M. Yu, G. Jiang, F. Shao, Z. Peng, F. Chen, No-reference stereoscopic image quality assessment using binocular self-similarity and deep neural network, Signal Processing Image Communication 47 (2016) 346-357.

[16] D. F. Mcallister, Stereo and 3-D Display Technologies, Encyclopedia of Imaging Science and Technology. John Wiley and Sons, Inc. 2002.

[17] S. K. Md, B. Appina, S. S. Channappayya, Full-reference stereo image quality assessment using natural stereo scene statistics, IEEE Signal Processing Letters 22 (11) (2015) 1985-1989.

[18] A. K. Moorthy, A. C. Bovik, Blind image quality assessment: From natural scene statistics to perceptual quality, IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society 20 (12) (2011) 3350-64.

[19] A. Mittal, A. K. Moorthy, A. C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society 21 (12) (2012) 4695-708.

[20] A. K. Moorthy, C. C. Su, A. Mittal, A. C. Bovik, Subjective evaluation of stereoscopic image quality, Signal Processing Image Communication 28 (8) (2013) 870-883.

[21] R. Patterson, L. Moe, T. Hewitt, Factors that affect depth perception in stereoscopic displays, Human Factors. 34 (6) (1992) 655-67.

[22] F. Qi, D. Zhao, W. Gao, Reduced reference stereoscopic image quality assessment based on binocular perceptual information, IEEE Transactions on Multimedia 17 (12) (2015) 1-1.

[23] N. Qian, Binocular disparity and the perception of depth, Neuron. 18 (3) (1997) 359-368.

[24] S. Rezazadeh, S. Coulombe, A novel discrete wavelet transform framework for full reference image quality assessment, Signal Image and Video Processing. 7 (3) (2011) 559-573.

[25] S. Ryu, K. Sohn, No-reference quality assessment for stereoscopic images based on binocular quality perception, IEEE Transactions on Circuits and Systems for Video Technology 24 (4) (2014) 591-602.

[26] M. A. Saad, A. C. Bovik, C. Charrier, Blind image quality assessment: a natural scene statistics approach in the dct domain, Image Processing IEEE Transactions on 21 (8) (2012) 3339-3352.

[27] F. Shao, W. Lin, S. Wang, G. Jiang, M. Yu, Q. Dai, Learning receptive fields and quality lookups for blind quality assessment of stereoscopic images,IEEE Transactions on Cybernetics (2015) 1.

[28] F. Shao, W. Lin, S. Gu, G. Jiang, T. Srikanthan, Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics, IEEE Transactions on Image Processing 22 (5) (2013) 1940-1953.

[29] F. Shao, W. Tian, W. Lin, G. Jiang, Toward a blind deep quality evaluator for stereoscopic images based on monocular and binocular interactions, IEEE Transactions on Image Processing 25 (5) (2016) 1-1.

[30] F. Shao, K. Li, W. Lin, G. Jiang, M. Yu, Q. Dai, Full-reference quality assessment of stereoscopic images by learning binocular receptive field properties, IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society 24 (10) (2015) 2971-83.

[31] H. R. Sheikh, A. C. Bovik, Image information and visual quality, IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society 15 (2) (2006) 430-444.

[32] J. Shen, Q. Li, G. Erlebacher, Hybrid no-reference natural image quality assessment of noisy, blurry, jpeg2000, and jpeg images, IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society. 20 (8) (2011) 2089-2098.

[33] R. Song, H. Ko, C. C. J. Kuo, Mcl-3d: a database for stereoscopic image quality assessment using 2d-image-plus-depth source, Journal of Information Science and Engineering 31 (5).

[34] I. Tsirlin, L. M. Wilcox, R. S. Allison, the effect of crosstalk on the perceived depth from disparity and monocular occlusion, IEEE Transactions on Broadcasting 57 (2) (2011) 445-453.

[35] K. Umeda, S. Tanabe, I. Fujita, Representation of stereoscopic depth based on relative disparity in macaque area v4., Journal of Neurophysiology 98 (1) (2007) 241-52.

[36] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600-612.

[37] J. Wang, A. Rehman, K. Zeng, S. Wang, Z. Wang, Quality prediction of asymmetrically distorted stereoscopic 3d images, IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society 24 (11) (2015) 3400-14.

[38] Y. Xie, L. F. Liu, C. H. Li, Y. Y. Qu, Unifying visual saliency with hog feature learning for traffic sign detection, Intelligent Vehicles Symposium IEEE (2009) 24-29.

[39] W. Xue, X. Mou, L. Zhang, A. C. Bovik, X. Feng, Blind image quality assessment using joint statistics of gradient magnitude and laplacian features, IEEE Transactions on Image Processing 23 (11) (2014) 4850-62.

[40] J. Yang, H. Wang, W. Lu, B. Li, A. Badiid, Q. Meng, A no-reference optical flow-based quality evaluator for stereoscopic videos in curvelet domain, Information Sciences 414 (2017).

[41] J. Yang, Y. Liu, Z. Gao, R. Chu, Z. Song, A perceptual stereoscopic image quality assessment model accounting for binocular combination behavior, Journal of Visual Communication and Image Representation 31 (C) (2015) 138-145.

[42] J. Yang, Y. Wang, B. Li, W. Lu, Q. Meng, Z. Lv, D. Zhao, Z. Gao, Quality assessment metric of stereo images considering cyclopean integration and visual saliency, Information Sciences An International Journal 373 (C) (2016) 251-268.

[43] J. Yang, P. An, J. Ma, K. Li, L. Shen, No-reference stereo image quality assessment by learning gradient dictionary-based color visual characteristics, in: IEEE International Symposium on Circuits and Systems, 2018.

[44] J. You, X. Wang, L. Xing, A. Perkis, Perceptual quality assessment for stereoscopic images based on 2d image quality metrics and disparity analysis, in: International Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2010.

[45] L. Z, T. J, Unsupervised feature selection via nonnegative spectral analysis and redundancy control., IEEE Transactions on Image Processing. 24 (12) (2015) 5343-5355.

[46] S. Zhang, C. Bauckhage, A. B. Cremers, Efficient pedestrian detection via rectangular features based on a statistical shape model, IEEE Transactions on Intelligent Transportation Systems 16 (2) (2015) 763-775.

[47] W. Zhang, C. Qu, L. Ma, J. Guan, R. Huang, Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network, Pattern Recognition 59 (2016) 176-187.

[48] W. Zhou, L. Yu, W. Qiu, Y. Zhou, M. Wu, Local gradient patterns (lgp): An effective local-statistical-feature extraction scheme for no-reference image quality assessment, Information Sciences. 397-398 (2017) 1-14.

[49] W. Zhou, L. Yu, Binocular responses for no-reference 3d image quality assessment, IEEE Transactions on Multimedia 18 (6) (2016) 1077-1084.