Pattern Recognition 59 (2016) 176–187

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network

Wei Zhang [a], Chenfei Qu [a], Lin Ma [b,*], Jingwei Guan [c], Rui Huang [d]

[a] School of Control Science and Engineering, Shandong University, China
[b] Huawei Noah's Ark Lab, Hong Kong
[c] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong
[d] NEC Laboratories China, China

ARTICLE INFO

Article history:
Received 4 August 2015
Received in revised form
24 January 2016
Accepted 31 January 2016
Available online 6 February 2016

Keywords:
Stereoscopic image
Quality assessment
Convolutional neural network (CNN)

ABSTRACT

In this paper, we propose to learn the structures of stereoscopic image based on convolutional neural network (CNN) for no-reference quality assessment. Taking image patches from the stereoscopic images as inputs, the proposed CNN can learn the local structures which are sensitive to human perception and representative for perceptual quality evaluation. By stacking multiple convolution and max-pooling layers together, the learned structures in lower convolution layers can be composed and convolved to higher levels to form a fixed-length representation. Multilayer perceptron (MLP) is further employed to summarize the learned representation to a final value to indicate the perceptual quality of the stereo image patch pair. With different inputs, two different CNNs are designed, namely one-column CNN with only the image patch from the difference image as input, and three-column CNN with the image patches from left-view image, right-view image, and difference image as the input. The CNN parameters for stereoscopic images are learned and transferred based on the large number of 2D natural images. With the evaluation on public LIVE phase-I, LIVE phase-II, and IVC stereoscopic image databases, the proposed no-reference metric achieves the state-of-the-art performance for quality assessment of stereoscopic images, and is even competitive to existing full-reference quality metrics.

1. Introduction

Image perceptual quality assessment [5] plays the essential role in the image processing and communication, such as image capturing, compression, storage, transmission, displaying, printing, and sharing. There existed many image quality metrics aiming at guiding the performance optimization during each step of image processing and communication. As human eyes are the ultimate receivers of the images, subjective evaluation [5,7,8] is regarded as the most reliable way to evaluate the perceptual quality of the image. The subjective evaluation requires a group of non-professional subjects to participate and watch the test images in some particular circumstances. Each participant gives a subjective rating for the target image. Finally, the mean opinion score (MOS) is calculated as the quality index. Apparently, subjective evaluation costs too much time and effort in the whole procedure. More importantly, it is impractical to perform subjective image quality assessment (IQA) in real-time. Hence, the objective quality metrics [9–12,15,63] that can automatically evaluate the image perceptual quality and guide the image processing applications are demanded.

With the rapid development of content generation and display technology, three-dimensional (3D) applications and services are becoming more and more popular to improve the visual quality of experiences (QoE) for human viewers. The 3D contents displaying on the 3D devices, such as the 3D films and video games, have now brought more entertainments and vivid experiences to the consumers, which have attracted more and more attentions from not only academia but also the industry. For these applications, the quality of 3D content is the most critical part to guarantee the visual QoE. However, in the 3D processing chain including capturing, processing, coding, transmitting, reconstruction, retrieving, etc., artifacts are inevitably introduced due to the resource shortage in processing. Therefore, how to evaluate the perceptual quality of 3D content becomes an important and challenging task in 3D visual signal processing, which can automatically evaluate the perceptual degradation during each processing stage. Compared to conventional 2D IQAs, IQAs for 3D signals are more challenging. The artifacts of 3D content affect more on human visual system (HVS) than conventional 2D contents. First, watching 3D contents for a long

* Corresponding author.
E-mail address: lma@ee.cuhk.edu.hk (L. Ma).

http://dx.doi.org/10.1016/j.patcog.2016.01.034
0031-3203/

time will bring adverse symptoms such as dizziness, nausea, and vomiting, which may affect the perceptual quality [3,16,19,20]. The visual discomfort and fatigue of watching 3D contents has been studied in [24,27]. Besides, the quality of the perceived depth [21,36] needs to be considered, which is greatly influenced by the stereopair images [22,23]. Also, incorrect usage of stereography may result in negative influences on visual QoE.

Nowadays, more and more images are appearing and shared on the Internet. With such a large amount of images, we can rely on intelligent image understanding [1,4,13,14,17] techniques to automatically process and analyse the images. Deep neural networks, more specifically the convolutional neural networks (CNNs), have been extensively studied for recognition [57,59], localization [42,48], and understanding [55,68,56]. CNN is a biologically inspired learning model. The features are learned end-to-end from raw data for classification or prediction. More specifically, CNN takes the raw images as input, and ensemble the feature learning and the training as a whole process. With a designed deep structure, CNN can effectively learn the complicated mapping relations between the raw image and the labels. Moreover, the spatial structure of images is adequately considered and used in CNN [58] for regularization through restricted connectivity between layers (local filters), parameter sharing (convolutions), and special local invariance-building neurons (max pooling). Furthermore, parameters in local filters and between layers are connected and trained as a whole to encode some characteristics about human visual system (HVS), such as the edges and contours, which are vital for human to perceive and understand an image.

Recently, a number of databases depicting the stereoscopic image quality are constructed [7,30,50]. Based on the large amount of labeled data, we resort to CNNs to construct a no-reference (NR) quality metrics for stereoscopic images, which can not only analyze the image content but also help model the HVS property. It is expected that the structures learned by CNNs are sensitive to the HVS quality perception. With such QA models, we can control and optimize the perceptual quality of the 3D contents, specifically the stereoscopic images, at each processing stage. The best visual QoE can thus be provided to the consumers. In this paper, we propose to construct CNNs to learn the local structures for NR IQA of stereoscopic images. The learned structures are expected to be sensitive to the distortions, and to indicate the perceptual quality of the stereoscopic images. With different inputs, two different CNNs are designed for the stereoscopic IQA. It is found that the performance of quality assessment on stereoscopic images can be boosted by transferring the learned parameters from 2D natural images to stereoscopic ones.

This paper is organized as follows. Section 2 gives a survey of the existing quality metrics for stereoscopic images. In Section 3, NR metrics based on CNN are proposed to learn the structures of stereoscopic images. Section 4 provides the implementation details of the proposed CNN models. Section 5 gives experimental results and comparative analysis. Finally, conclusions are drawn in Section 6.

2. Related work

Based on the volume of accessible information in the images, the current IQA methods for stereoscopic images can be generally divided into three categories: full-reference (FR), reduced-reference (RR), and no-reference (NR).

FR IQAs require the original image accessible for comparison to generate the quality index of the content. Gorley and Holliman [29] proposed a PSNR-based measurement to account for the sensitivity of HVS to contrast and luminance changes at regions of high spatial frequency between the left and right views of a stereoscopic image pair. Benoit et al. [30] presented an IQA to predict the quality by evaluating the left reference image and left distorted one, the right reference image and right distorted one, and the disparity maps in a 2D manner. Some well-known 2D quality metrics such as C4 [31] and SSIM [32] are used to produce the quality score for stereoscopic images. You et al. [33] investigated the capabilities of 2D IQAs for stereoscopic images, and proved that disparity is an important factor in stereopsis. Yang et al. [34] proposed an FR IQA metric which measures not only the average PSNR of the two images in the stereo-pair, but also the absolute difference between the left and right view. Chen et al. [35] synthesized an intermediate 'cyclopean' images, and then applied 2D FR metric on it to predict the perceptual quality. The FR metrics can yield good performances, where the original stereoscopic image pair is required. Therefore, FR IQAs are mostly employed for guiding the image/video compression and watermarking. However, in most practical applications, the original stereoscopic images are unavailable, and FR IQAs cannot help evaluate the perceptual quality of stereoscopic images.

In order to relieve the limitations of FR IQAs, RR [46,47] metrics are presented, which only require partial information of the reference images and are mostly used in applications like real-time visual information communications over wired or wireless networks. The RR IQA is employed to monitor image quality degradations or control the network streaming resources. As edges and contours of the depth map indicate different depth levels of stereoscopic images, Hewage and Martini [28] presented an RR IQA metric by exploiting the edge information of the depth map. Ma et al. [45] proposed an RR IQA metric for stereoscopic images by evaluating the distortion in the reorganized discrete cosine transform (DCT) domain. Wang et al. [52] relied on the natural image statistics in the contourlet domain to design an RR metric for stereoscopic images. However, when using the FR metrics, we still need to extract features from the original images, which is still an obstacle for the real-world applications. Another issue is that we need to transmit or embed the extracted RR features with the distorted images. This will introduce an additional burden for quality assessment.

In practical applications, NR IQAs are more appealing because the distorted images can be assessed without any reference. Existing NR IQA metrics normally assume that the distortion type is known beforehand. Based on the behaviors of the distortions, i.e. blur [6], the perceptual quality of an image is easier to be evaluated. The general NR IQA [44] for evaluating the image perceptual quality without knowing the distortion type is highly demanded. Chen et al. [50] proposed to utilize the 2D cues in cyclopean images [35] and 3D cues in disparity [33], and delivered competitive performance compared to the FR IQA metrics. Sazzad et al. [60] developed an NR method based on spatio-temporal segmentation using the perceptual differences of local features in stereopairs. Akhter et al. [37] employed a logistic regression model to predict the quality, where the features are extracted from stereo-pairs and the disparity maps. Ryu and Sohn [54] presented an NR IQA scheme by modelling the binocular quality perception of the HVS in the context of blurriness and blockiness. Shao et al. [53] constructed binocular guided quality lookup and visual codebook to achieve NR IQA by simply pooling process. These NR quality metrics are mostly based the handcrafted features to represent the characteristics of the stereoscopic images. However, these features may not well reflect the perceptual quality of stereoscopic images. Recently, Kang et al. [44] did a pioneering work on discussing the capability of the CNN for evaluating the quality of an image. Inspired by [44], we propose an end-to-end CNN for the quality assessment of stereoscopic images. To the best of our knowledge, this is the first attempt of employing CNN for NR quality assessment of stereoscopic images. The proposed CNN

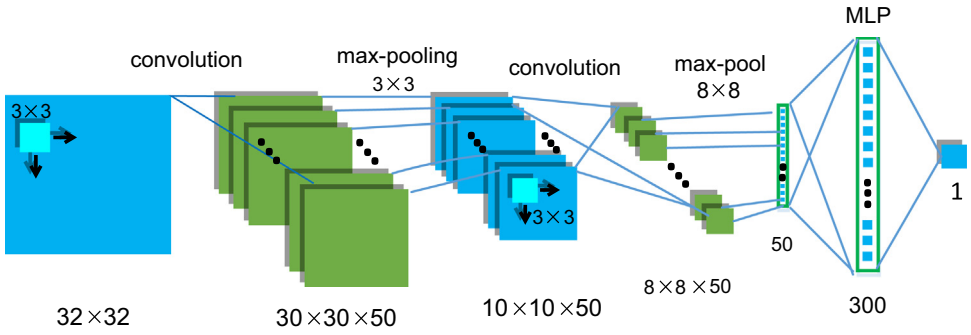

Fig. 1. Framework of the proposed one-column CNN.

Fig. 2. The stereoscopic image sample. (a) The left view image; (b) the right view image; (c) the difference image.

couples the feature extraction and learning process together to produce the perceptual quality from the image pixel. Moreover, we consider both the image contents and the view difference information of the stereoscopic images in the proposed quality metric.

3. Learning structure of stereoscopic image with convolutional neural network

In this paper, we focus on learning the structures of the stereoscopic images for NR IQA. Nowadays, convolutional neural networks (CNNs) have been successfully employed to learn the image representation for various applications, such as image classification [38–41], object detection [42,43], human parsing [18,2,25] and activity recognition [26]. In this paper, we rely on CNN for learning local structures of the stereoscopic images. The structures are learned via multiple layers of convolution and max-pooling, which are expected to be sensitive to the quality perception of the stereoscopic images.

The stereoscopic images differ from the 2D natural images, as the left and right views together can provide depth perception. Therefore, perceptual evaluation of the stereoscopic images needs to consider the information from both the left and right views. We propose two CNNs to fully exploit the structures of the stereoscopic images, which are expected to be sensitive for quality perception. As demonstrated in [45], the difference image between the left view and right view is more important than the left and right views for quality assessment. We firstly introduce a one-column CNN to learn the structures of the difference image for the aim of quality evaluation.

3.1. One-column CNN

The proposed one-column CNN for learning the structures of the difference image is illustrated in Fig. 1. The input of the constructed CNN is an image patch sampled from the difference image. Two layers of convolutions are employed to generate the intermediate representation, each of which is followed by a max-pooling layer, which can further reduce the computation for upper layers. With the convolution and max-pooling processes, the multilayer perception (MLP) with two fully-connected layers is employed to summarize the representation into a final score indicating the perceptual quality of the input image patch.

3.1.1. Preparation of the image patch

As aforementioned, the difference image is more important than the left and right views for stereoscopic IQA. The reason is that the difference image not only considers the image content but also the depth and disparity information of the stereoscopic images [45]. Generally, the difference image can be obtained by:

$$I_d(x,y) = I_l(x,y) - I_r(x + d(x,y), y + d(x,y)) \quad (1)$$

where I_l and I_r are the left and right view images, respectively. $d(x,y)$ denotes the disparity. I_d is the obtained difference image by referring the disparity information. However, for NR quality assessment, since the original (distortion-free) images are missing, the disparity cannot be identified accurately from the distorted ones. Similar to [34,45,51], we calculate the difference image from the left image and right image directly as:

$$I_d(x,y) = I_l(x,y) - I_r(x,y) \quad (2)$$

As shown in Fig. 2, the difference image containing the view differences can also imply the disparity and depth information of the stereoscopic image pair.

3.1.2. Convolution

Convolutional process is the biologically inspired variant of MLPs, which exploits the spatially local correlation by enforcing a local connectivity pattern. In this paper, we employ a small receptive field defined as 3×3 for the convolution process, which is the smallest size to capture the notion of left/right, up/down, and center. This is because that it is easy to obtain an effective receptive field of 5×5 by stacking two 3×3 convolution layers, and an effective receptive field of 7×7 by stacking three 3×3 convolution layers. As stated in [39], compared with one 7×7 convolution layer, three 3×3 convolution layers incorporate more non-linear layers, which could make the decision function more discriminative. Second, the number of convolution parameters can be significantly

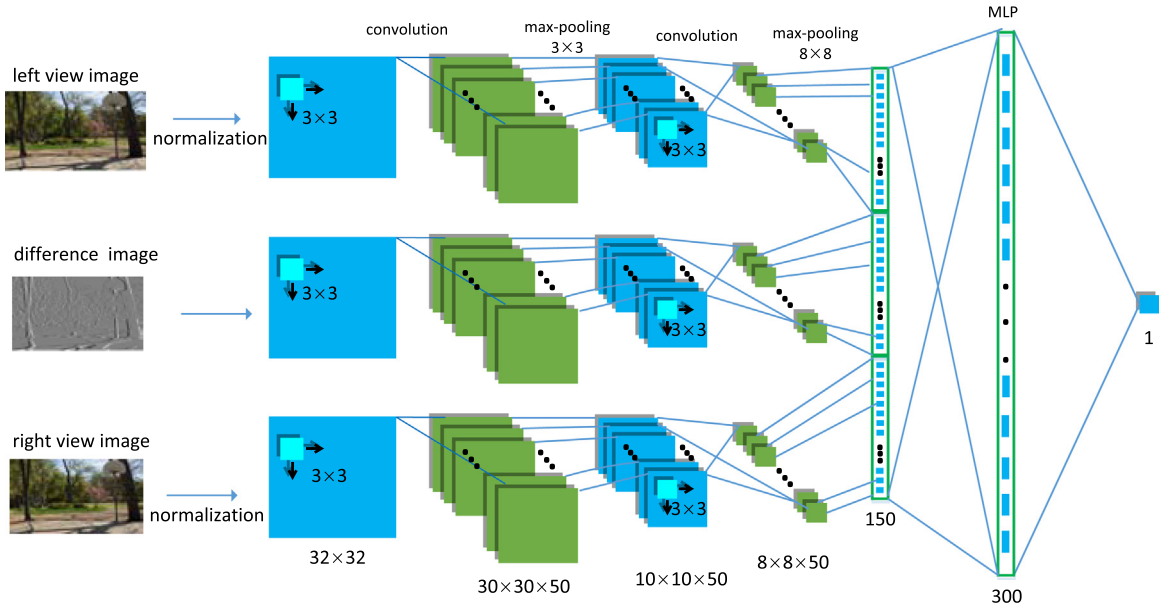

Fig. 3. Framework of the proposed three-column CNN.

reduced. If each layer of 3×3 convolution has K channels for convolution, all three 3×3 convolution layers requires $3 \times (3^2 \times K) = 27K$ parameters. For one single 7×7 convolution layer with K channels, the number of parameters is $7^2 \times K = 49K$, which is more than that of three 3×3 convolution layers.

With such configuration (3×3 convolution), the output of each convolution process only depends on the local 3×3 spatial content, and is unresponsive to the signals outside the receptive field with respect to the retina. Such convolution process ensures that the learnt 'filters' produce the strongest responses to the spatially local input patterns. The convolution process can be defined as the following:

$$h_{ij}^k = \omega^k x_{ij} + b^k \tag{3}$$

where ω^k and b^k are the parameters of the convolution filters for the kth feature map generation. x_{ij} denotes the local image patch lying in the receptive field. h_{ij}^k is the generated kth feature map. Note that the parameters ω^k and b^k are shared for each convolution layer. Twofold benefits are provided. First, shared parameters allow features to be detected regardless of their locations in the image, which can help learn the local structures of the image. Second, learning efficiency can be improved as the number of the parameters to be learned is reduced significantly. Compared to the fully connected layers like autoencoder and restricted Boltzmann machine (RBM), convolutional layer can learn a good representation of the image patch with a small number of parameters. Furthermore, in order to form a richer representation of the input image patches, multiple feature maps $\{h^k, k = 0, \ldots, K\}$ are generated as the output of the convolution layer.

3.1.3. Max-pooling

After convolution, a richer representation with multiple feature maps is obtained. Pooling is performed to improve position invariance of the convolution filters. In this paper, we employ the max-pooling [49,39,40] to partition the input image patch into a set of rectangles and output the maximum value for each sub-region. The max-pooling process is defined as:

$$\mu_{xy}^k = \max_{(x,y) \in \Omega} (h_{xy}^k) \tag{4}$$

where Ω denotes the local window for max-pooling. $h_{xy}{}^k$ denotes the kth feature map after the convolution. $\mu_{xy}{}^k$ represents the feature value obtained after pooling process.

The effects of max-pooling are twofold. First, the max-pooling process eliminates the non-maximal values. Together the stride parameter, max-pooling can reduce the dimension of the feature map as well as the computational complexity of upper layers. Second, max-pooling provides a form of translation-invariant features for the upper layer convolution process. With the feature map in the same window, only the maximum value is pooled out, which will not affect the upper layer process.

3.1.4. Multilayer perception (MLP)

After performing two layers of convolution and pooling processes, the final representation is obtained. MLP with two fully-connected layers are utilized to summarize the representation and generate the final score as follows:

$$S = \omega_s(\sigma(\omega_h(\vartheta_{im}) + b_h)) + b_s \tag{5}$$

where $\sigma(\cdot)$ is the nonlinear activation function. ϑ_{im} denotes the learned representation with two layers of convolution and max-pooling. ω_h and b_h are used to map the obtained image representation ϑ_{im} to the representation in the hidden layer. ω_s and b_s are the parameters to compute the final score of the input image patch. S is the learned score to indicate the perceptual quality of the input image patch.

3.2. Three-column CNN

Besides the difference image, the left view and right view images are also available and important for the stereoscopic image quality assessment. Therefore, we propose a new architecture with three column CNNs to jointly consider the content property (from the left and right view image), and the view difference property (from the difference image). The architecture of the three-column CNN is illustrated in Fig. 3.

For each column in the three-column CNN, the architecture is the same as the one-column CNN and has two layers of convolution and max-pooling. The three identical CNNs are used to learn the structures from the left view, right view, and difference image, respectively. With two layers of convolution and max-pooling processes, three different image patches are represented as three different vectors, which are expected to contain the structures of stereoscopic images from different viewpoints. Finally, these three vectors are concatenated together and fed into the upper MLP

layer as shown in Fig. 3 to learn their compositions and finally output the quality score.

4. Implementation details

In this section, we will first describe the configurations of the proposed two CNN models for NR stereoscopic IQA, and then introduce how the proposed CNNs are trained.

4.1. CNN configurations

Configurations of the proposed two CNNs are outlined in Table 1. For one-column CNN, image patch with size of 32×32 is cropped only from the difference image, while three image patches of the same size extracted from the left view, right view, and difference image are taken as input for three-column CNN. We have two layers of convolution and max-pooling to learn the structures of the stereoscopic image patches and produce the final representation. For the convolution, 3×3 convolution kernels are employed. For max-pooling, 3×3 together with stride of 3 are used in the lower layer. For upper layer, the 8×8 max-pooling is used to represent the image patch as one 50-dimensional feature vector. After convolution and max-pooling, an MLP with two fully-connected layers is employed to map the representation to the hidden layer and produce the perceptual quality of the input patch.

4.2. Training

We train our network on non-overlapping 32×32 image patches generated from the stereoscopic images. Since the training images in our database have homogeneous distortions, each input patch was assigned a score the same as the one of its source image. Given the image patch p and its quality score Q_p, we define the training objective function as follows:

$$L = \frac{1}{N}\sum_{n=1}^{N} \| S^n - Q_p^n \|_2^2 \tag{6}$$

where S^n is the generated quality score via one-column or three-column CNN by Eq. (5) for the nth image patch. ${Q_p}^n$ is the corresponding ground-truth value. By minimizing the objective function defined in Eq. (6), the parameters of the proposed CNNs can be learned via stochastic gradient decent (SGD) and back-propagation. Besides, a validation set and dropout procedure are included to prevent overfitting.

However, since only a small number of stereoscopic images with annotated subjective scores are available in the training set, it may not be able to tune CNNs well for learning structures in our model. To relieve this problem, we turn to pretraining with 2D natural images to give initial estimations of the CNN parameters, and then transfer the learnt parameters from 2D images to the stereoscopic ones.

Table 1
Configurations of the proposed CNNs. (conv denotes the convolution layer; max denotes the max-pooling layer; FC denotes the full-connected layer.)

One-column CNN Input: 32×32 patch	Three-column CNN Input: three 32×32 patches from left, right, and difference images		
conv-50 (3×3)	conv-50 (3×3)	conv-50 (3×3)	conv-50 (3×3)
max (3×3)	max (3×3)	max (3×3)	max (3×3)
conv-50 (3×3)	conv-50 (3×3)	conv-50 (3×3)	conv-50 (3×3)
max (8×8)	max (8×8)	max (8×8)	max (8×8)
FC-300	FC-300		
FC-1	FC-1		

4.3. 2D-to-stereoscopic transfer

As illustrated in Fig. 4, the histograms of left and right images significantly differ from that of the difference image. Also, it can be observed that the distribution of the difference image is approximately of zero mean. Therefore, in order to produce similar image patches as the difference image, we firstly normalize the image patches as follows:

$$\begin{aligned}
\hat{I}(x,y) &= \frac{I(x,y)-\mu(x,y)}{\sigma(x,y)+c} \\
\mu(x,y) &= \frac{1}{M \times N}\sum_{(i,j)\in\Omega} I(x+i,y+j) \\
\sigma(x,y) &= \frac{1}{M \times N}\sqrt{\sum_{(i,j)\in\Omega} ((x+i,y+j)-\mu(x,y))^2}
\end{aligned} \tag{7}$$

where Ω denotes the local region for calculating the mean and variance is defined as a 3×3 window. $M \times N$ denotes the total number of the image pixels in the local window. c is a small positive value in case of the denominator becomes zero. After normalization, the original patch is of zero mean and unit variance. And the effects of luminance and contrast information are also alleviated for further learning structures. The histogram distribution of the normalized image appears to be similar with that of the difference image.

In order to further demonstrate the effectiveness of the normalization process as in Eq. (7), we employ different measurements to evaluate the histogram distance between the difference image and the original/normalized view image. More specifically, the measurements, Kullback–Leibler divergence (KLD), Jenson–Shannon divergence (JSD), Jeffrey divergence (JD), Chi-square (CS), Kolmogorov–Smirnov (KS), quadratic form (QF), and histogram intersection, as stated in [70,71] are used. For these measurements, the smaller the distance, the more similar the two histogram distributions. Table 2 illustrates the histogram distances in terms of different measurements between the difference image and original/normalized view images, shown in Fig. 4. It can be observed that the histogram distance between the difference image and the normalized view image is much smaller than that between the difference image and original view image. As such, we can assume that the normalized view images can well resemble the statistical characteristics and structures of the difference image. Therefore, it is reasonable to import the normalized image patch for the training process. It is expected that we can transfer the parameters learned from 2D natural images to the difference image of the stereoscopic images.

Since the 2D natural images are much more than the stereoscopic images, we can sample the patches from 2D images and perform the normalization process before feeding them into the one-column CNN. After pretraining, the parameters of the convolution and fully-connected layers can be obtained. For one-column CNN, these parameters are used to initialize the network before training on the difference images of the stereoscopic images. For the three-column CNN, only the parameters of convolution layers are employed to initialize the model. For the upper two fully-connected layers, we randomly initialize the parameters for the stereoscopic image patches.

5. Experiments

In this section, we evaluate the effectiveness of our proposed CNNs for learning the structures of stereoscopic images. We begin by

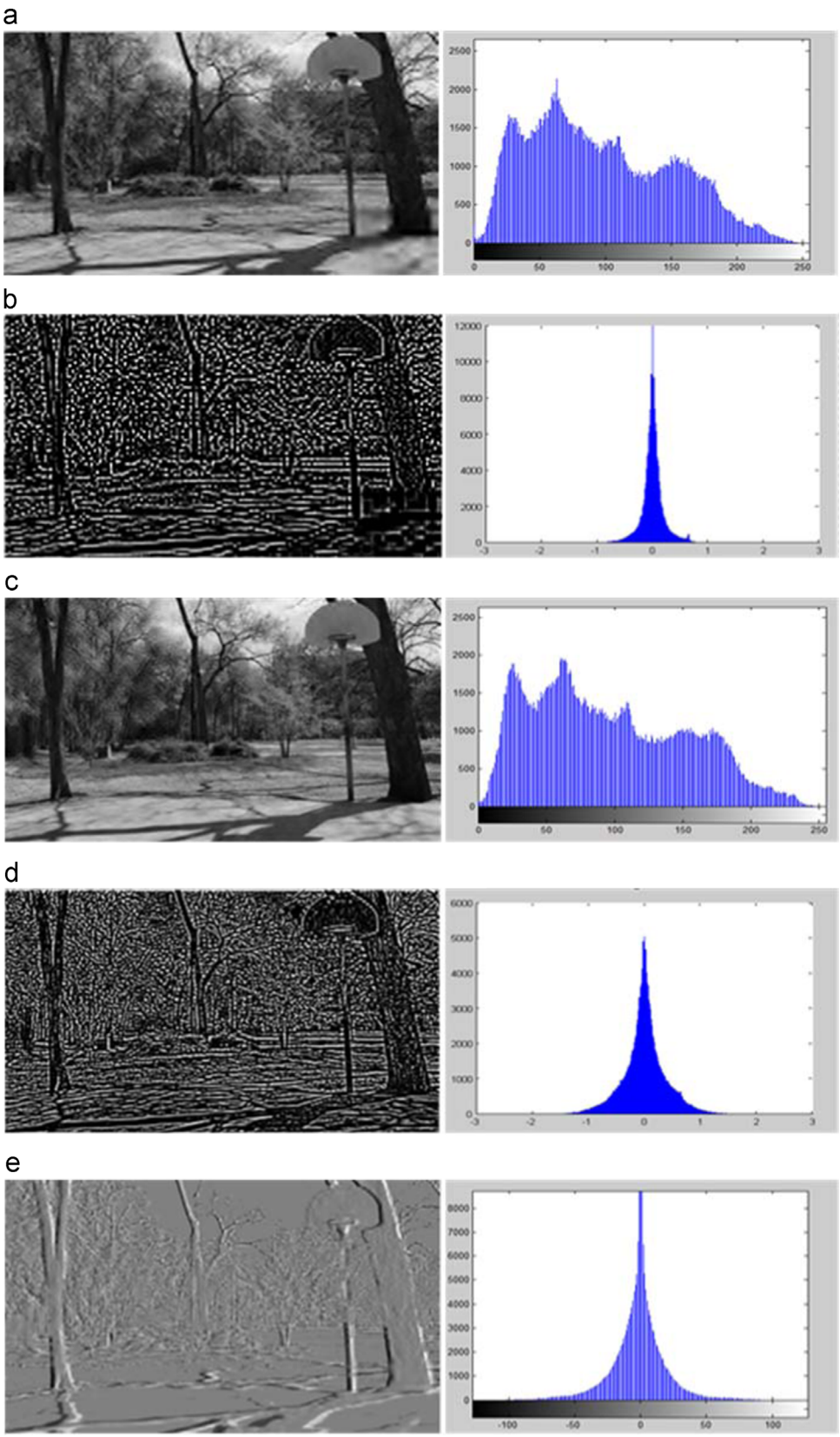

Fig. 4. The histograms of the stereoscopic images. (a) The left image and its histogram; (b) the normalized left image and its histogram; (c) the right image and its histogram; (d) the normalized right image and its histogram; (e) the difference image and its histogram.

Table 2
Relationship between the difference image and the original/normalized view image.

Histogram distance	Left	Normalized left	Right	Normalized right
KLD	1.8009	0.0991	1.6257	0.0620
JSD	0.4021	0.0260	0.3780	0.0162
JD	0.8041	0.0519	0.7559	0.0325
CS	0.6679	0.0511	0.6316	0.0321
KS	0.8013	0.2085	0.7747	0.1593
QF	0.8800	0.2299	0.8527	0.1727
Histogram intersection	0.8080	0.2082	0.7843	0.1593

describing the datasets used for validation, followed by experimental results and comparative analysis on stereoscopic IQA.

5.1. Datasets

- *LIVE 2D natural image dataset* [8]: A total of 779 distorted images with five different distortions, specifically the JPEG2000 compression (JP2K), JPEG compression (JPEG), white noise (WN), Gaussian blur (BLUR), and fast fading (FF) are generated from 29 reference images. Differential mean opinion scores (DMOSes) are provided for each image, roughly in the range [0,100]. Higher DMOS value indicates lower perceptual quality. As mentioned before, the insufficient number of stereoscopic images cannot well train and finetune our proposed CNNs. Therefore, the 2D natural images are used to firstly learn the parameters in CNNs. Afterwards, the learned parameters from 2D natural images can be further transferred to the stereoscopic images. As discussed in Section 4.3, the normalized versions of the 2D natural image present similar characteristics with the difference image of the stereoscopic images. That is also the main reason that the normalized image can be used for learning the parameters.
- *LIVE stereoscopic image dataset phase-I* [7]: The LIVE stereoscopic image dataset phase-I consists of 20 reference images and 365 distorted images (45 images for BLUR and 80 images for each JPEG, JP2K, FF, WN distortion type) with co-registered human scores in the form of DMOS (roughly in the range [0,100]). The LIVE stereoscopic image dataset phase-I contains the same types of distortions with LIVE 2D images, and all distortions are symmetric in nature. The 'symmetric' means that the same level of distortion was created between the image pair. Based on the transferred parameters learned from 2D natural images, the stereoscopic images are further employed to tune the parameters. Specifically, we random select 10 images of BLUR distortion and 20 images of other distortions as the training samples. Based on the training samples of the stereoscopic images, we can further train the transferred parameters from 2D natural images. After obtaining the trained parameters, the rest distorted stereoscopic images are used for testing.
- *LIVE stereoscopic image dataset phase-II* [50]: Though LIVE stereoscopic image dataset phase-I is a valuable resource for the research on 3D IQA, it only has symmetrically distorted stereo images. However, in the real world, a stereo image pair may be distorted symmetrically or asymmetrically. To provide a complete stereoscopic image database, LIVE stereoscopic image dataset phase-II provides both symmetrically and asymmetrically distorted stereo images with human subjective ratings. LIVE phase-II database consists of 8 reference images and 360 distorted images with co-registered human scores in the form of DMOS. For each distortion type (JPEG, JP2K, BLUR, FF, and WN), every reference stereopair was processed to create three symmetric distorted stereopairs and six asymmetric distorted stereopairs.
- *IVC stereoscopic image dataset* [30]: This dataset contains 96 stereoscopic images and their associated subjective scores. The resolution of these images is 512×512, which were displayed on a 1280×1024 monitor with an uniform gray around the image to keep the native resolution of the image. 6 different stereoscopic images are used in this dataset, which is composed of the 6 reference images (undistorted) and 16 distorted versions of each sources generated from 3 different distortion types (JPEG, JP2K, BLUR) symmetrically to the stereopair images.

Table 3
SROCC of IQA metrics in predicting stereoscopic image quality on LIVE phase-I dataset.

Type	Algorithm	JP2K	JPEG	WN	BLUR	FF	ALL
FR	Benoit et al. [30]	0.9103	0.6028	0.9292	0.9308	0.6989	0.8892
	You et al. [33]	0.8598	0.4388	0.9395	0.8822	0.5833	0.8789
	Gorley and Holliman [29]	0.4203	0.0152	0.7408	0.7498	0.3663	0.1419
	MS-SSIM [35]	0.948	0.888	0.53	0.925	0.707	0.916
RR	Hewage and Martini [28]	0.8558	0.5001	0.8963	0.6900	0.5477	0.8140
	Wang et al. [32]	0.8832	0.5420	0.9066	0.9246	0.6548	0.8890
	Ma et al. [45]	0.8866	0.6163	0.9124	0.8791	0.6964	0.9052
NR	Akhter et al. [37]	0.866	0.675	0.914	0.555	0.640	0.383
	Shao et al. [53]	0.9003	0.6073	0.9032	0.9235	–	0.8941
	Ryu and Sohn [54]	–	–	–	–	–	0.86
	Chen et al. [50]	0.863	0.617	0.919	0.878	0.652	0.891
	Proposed one-column CNN	0.889	0.613	0.909	0.877	0.758	0.925
	Proposed three-column CNN	0.931	0.693	0.946	0.909	0.834	0.943

Table 4
LCC of IQA metrics in predicting stereoscopic image quality on LIVE phase-I dataset.

Type	Algorithm	JP2K	JPEG	WN	BLUR	FF	ALL
FR	Benoit et al. [30]	0.9398	0.6405	0.9253	0.9488	0.7472	0.9025
	You et al. [33]	0.8778	0.4874	0.9412	0.9198	0.7300	0.8814
	Gorley and Holliman [29]	0.4853	0.3124	0.7961	0.8527	0.3648	0.4511
	MS-SSIM [35]	0.948	0.888	0.53	0.925	0.707	0.916
RR	Hewage [28]	0.9043	0.5305	0.8955	0.7984	0.6698	0.8303
	Wang et al. [32]	0.9162	0.5697	0.9133	0.9574	0.7833	0.8921
	Ma et al. [45]	0.9182	0.7222	0.9131	0.9247	0.8068	0.9056
NR	Akhter et al. [37]	0.905	0.729	0.904	0.617	0.503	0.626
	Shao et al. [53]	0.8723	0.8975	0.9161	0.9233	–	0.8994
	Ryu and Sohn [54]	0.86	0.63	0.94	0.96	0.78	0.80
	Chen et al. [50]	0.907	0.695	0.917	0.917	0.735	0.895
	Proposed one-column CNN	0.898	0.632	0.923	0.928	0.845	0.926
	Proposed three-column CNN	0.926	0.740	0.944	0.930	0.883	0.947

5.2. Results and analysis

5.2.1. Proposed CNN models for stereoscopic IQA

As usual [8], SROCC, LCC, and RMSE are used to evaluate the performance of different quality metrics. Larger SROCC and LCC values indicate better performance, while smaller RMSE value represents better performance. As shown in Tables 3–5, the FR metric Gorley [29] performs the worst. One reason is that the

Table 5
RMSE of IQA metrics in predicting stereoscopic image quality on LIVE phase-I dataset.

Type	Algorithm	JP2K	JPEG	WN	BLUR	FF	ALL
FR	Benoit et al. [30]	4.4266	5.0220	6.3076	4.5714	8.2578	7.0617
	You et al. [33]	6.2066	5.7097	5.6216	5.6798	8.4923	7.7463
	Gorley and Holliman [29]	11.324	6.212	10.198	7.562	11.569	14.635
	MS-SSIM [35]	5.581	5.320	5.216	4.822	7.837	6.533
RR	Hewage [28]	5.5300	5.5431	7.4056	8.7480	9.2263	9.1393
	Wang et al. [32]	5.1890	5.3741	6.7772	4.1777	7.7245	7.4081
	Ma et al. [45]	5.1294	4.5229	6.7843	5.5105	7.3411	6.9542
NR	Akhter et al. [37]	5.483	4.273	7.092	11.387	9.332	14.827
	Chen et al. [50]	5.402	4.523	6.433	5.898	8.322	7.247
	Proposed one-column CNN	6.000	5.926	6.107	6.839	6.135	6.148
	Proposed three-column CNN	4.986	4.396	5.676	5.539	6.049	5.336

Table 6
LCC of IQA metrics in predicting stereoscopic image (with symmetric distortion) quality on LIVE phase-II dataset.

Type	Algorithm	JP2K	JPEG	WN	BLUR	ALL
2D	SSIM [72]	0.8162	0.6770	0.9749	0.8325	0.7326
	FSIM [65]	0.8183	0.8456	0.9630	0.8638	0.8301
	GSMD [66]	0.8749	0.8443	0.9613	0.9279	0.9245
3D	Bensalma et al. [64]	0.6897	0.5514	0.9359	0.9527	0.8232
	Chen et al. [35]	0.6704	0.6013	0.9463	0.9178	0.8135
	Shao et al. [53]	0.9032	0.8732	0.9170	0.9773	0.9119
	Proposed three-column CNN	0.9212	0.9270	0.9571	0.8992	0.9121

Table 7
SROCC of IQA metrics in predicting stereoscopic image (with symmetric distortion) quality on LIVE phase-II dataset.

Type	Algorithm	JP2K	JPEG	WN	BLUR	ALL
2D	SSIM [72]	0.7261	0.7180	0.9452	0.7704	0.7003
	FSIM [65]	0.8243	0.8406	0.9365	0.8495	0.9086
	GSMD [66]	0.8669	0.8380	0.9269	0.8356	0.9102
3D	Bensalma et al. [64]	0.6078	0.5475	0.9243	0.8461	0.8046
	Chen et al. [35]	0.6617	0.6304	0.9070	0.8450	0.8372
	Shao et al. [53]	0.9043	0.9102	0.9365	0.9113	0.8966
	Proposed three-column CNN	0.8974	0.9424	0.9407	0.4897	0.9145

method treats the left and right view images independently. Another reason is that the method is based on PSNR, which has been demonstrated to be limited in perceptual evaluation. Also, the disparity or depth information is not considered, which is vital in the stereoscopic IQA. For the other FR metrics, Benoit et al. [30] and You et al. [33] incorporated the depth and disparity information for stereoscopic IQA. Chen et al. [35] constructed the cyclopean image and then performed 2D quality assessment. As such, the depth and disparity information is considered, and better performance is achieved compared to the metric in [29]. The MS-SSIM [35] performs the best. The reason can be attributed to that SSIM presents powerful abilities for capturing the structural similarities. For the RR metric of [28], its performance is better than the FR metrics of [33,29], which demonstrates that the edge pixels are very sensitive to quality perception of stereoscopic images. Compared to FR and RR metrics, the NR method is more difficult. Therefore, the NR metrics of [37,53,50,54] present worse performances than FR metrics of [35,30] and RR metrics of [32,45].

The proposed one-column CNN and three-column CNN models outperform the other metrics when evaluating all the distorted images together. Specifically, our models give higher SROCC and LCC values. One reason is that the difference image is well exploited to generate perceptual quality. Another reason is that we employ the CNN to learn the structures of the stereoscopic images rather than using the handcrafted features for quality assessment. This shows the promising abilities of CNN for stereoscopic IQA.

The core idea of three-column CNN is to use the difference image between stereopair and 2D images to predict the quality. Hence, three-column CNN can take fully consideration of the specific structure of stereo image pair by training the three image patches from the left image, right image, and difference image simultaneous. As shown in Tables 3–5, the three-column CNN performs better than the one-column CNN. It can be concluded that the three-column model contains not only the independent information of stereo image pair, but also the joint information between these two views.

As discussed in the introduction, our constructed one-column and three-column CNNs are end-to-end networks, which take the image patch as input and generate the quality index for the provided image patch. Therefore, it is unsurprised that CNN performed better than existing methods, as demonstrated by the experimental results. First, benefiting from the rich image datasets, CNN can learn features from the raw images effectively and automatically for stereoscopic image quality assessment, instead of relying on handcrafted features as previous work. Second, when training CNN, feature extraction and quality assessment are combined together into one optimization process. Moreover, we include the difference image into the proposed one-column and three-column CNNs. As such, the disparity and depth information is explicitly considered, which is proved to be important for the stereoscopic image quality perception.

5.2.2. *Performances on different distortions*

Breaking down the performance by distortion types, a significant improvement over other quality metrics can be observed for all distortion types as shown in Tables 3–5. Moreover, it can be observed that the proposed CNN models achieve the best performances on the distortion type of WN and the worst performance on the type of JPEG. This is due to that the distortions of JPEG images are less perceptually separated, and thus are more challenging to be assessed [50].

5.2.3. *Cross dataset test*

In this section, we examine the generalization capability of our proposed CNN model. We follow the protocol of [44] to investigate cross dataset performance by training our proposed three-column CNN on LIVE phase-I and testing on LIVE phase-II and IVC datasets, respectively.

Cross dataset test on LIVE stereoscopic image dataset phase-II: For fair comparison, only the images with four types of distortions (JPEG, JP2K, BLUR, and WN) from the LIVE phase-II dataset are used for testing and comparison. The LCC and SROCC values of IQAs on LIVE phase-II are given in Tables 6–9. More specifically, Tables 6 and 7 illustrate the LCC and SROCC values on the images with symmetric distortions, while Tables 8 and 9 demonstrate the LCC and SROCC values on the rest images with asymmetric distortions. Three 2D IQAs, specifically SSIM [72], FSIM [65], and GSMD [66], as well as three 3D IQAs, specifically the metrics of Bensalma et al. [64], Chen et al. [35], and Shao et al. [53] are tested and compared. It should be noted that only the metric of Shao et al. [53] works in a NR manner. The other IQAs are all FR quality metrics, which require the original stereoscopic images for quality

Table 8
LCC of IQA metrics in predicting stereoscopic image (with asymmetric distortion) quality on LIVE phase-II dataset.

Type	Algorithm	JP2K	JPEG	WN	BLUR	ALL
2D	SSIM [72]	0.6755	0.6845	0.8230	0.8403	0.7497
	FSIM [65]	0.7846	0.7963	0.9410	0.8879	0.6775
	GSMD [66]	0.8680	0.8690	0.9160	0.7411	0.6533
3D	Bensalma et al. [64]	0.6194	0.6305	0.9325	0.8621	0.7432
	Chen et al. [35]	0.7220	0.5636	0.9449	0.6918	0.6337
	Shao et al. [53]	0.7893	0.7052	0.9235	0.8547	0.5651
	Proposed three-column CNN	0.7821	0.5825	0.7955	0.9236	0.7625

Table 9
SROCC of IQA metrics in predicting stereoscopic image (with asymmetric distortion) quality on LIVE phase-II dataset.

Type	Algorithm	JP2K	JPEG	WN	BLUR	ALL
2D	SSIM [72]	0.7237	0.7144	0.8821	0.8068	0.7193
	FSIM [65]	0.8064	0.8050	0.9521	0.8501	0.6610
	GSMD [66]	0.8536	0.8758	0.9366	0.8877	0.6420
3D	Bensalma et al. [64]	0.6194	0.6779	0.9409	0.8402	0.6968
	Chen et al. [35]	0.7220	0.6359	0.9292	0.6912	0.6108
	Shao et al. [53]	0.7893	0.6961	0.9235	0.8031	0.5244
	Proposed three-column CNN	0.7928	0.5805	0.7797	0.8653	0.7078

Table 10
LCC and SROCC of IQAs on IVC stereoscopic image database.

Type	Algorithms	LCC	SROCC
2D	PSNR	0.5843	0.5554
	Carnec et al. [31]	0.7874	0.7304
	Skeikh et al. [61]	0.7051	0.6135
	MS-SSIM [35]	0.7676	0.6919
	Venkata et al. [62]	0.6816	0.5973
	SSIM [32]	0.6817	0.6478
3D	Ryu et al. [67]	0.7579	0.6869
	Campisi et al. [69]	0.7873	0.7295
	Proposed three-column CNN	0.7917	0.7644

assessment. It can be observed that our proposed three-column CNN performs the best on both symmetric and asymmetric distortions. SSIM [72] performs similarly on the images with symmetric or asymmetric distortions. It is based on a top-down assumption that the HVS is highly adapted for extracting structural information from the scene, and therefore a measure of structural similarity should be a good approximation of perceived image quality. However, the SSIM index is a single-scale approach. The right scale depends on viewing conditions (e.g., display resolution and viewing distance). On the contrary, GMSD only uses the gradient magnitude information, which produced the best results among 2D IQAs on symmetric part, but performs the worst on asymmetric images. The main reason is that it ignores how the difference between left and right views influences the quality. FSIM employs the image gradient magnitude and the phase congruency with a pooling strategy, and provides a good performance on symmetric distortions. For asymmetric distortions, FSIM performs poorer compared with our proposed CNN, even it requires the original stereoscopic images as a reference.

For the 3D quality metrics, Shao et al. [53] perform better than Bensalma et al. [64] and Chen et al. [35] on the images with symmetric distortions, even though it performs on an NR manner.

Table 11
SROCC, LCC and RMSE with different kernel sizes.

Measurement	Kernel size	JP2K	JPEG	WN	BLUR	FF	ALL
SROCC	3×3	0.931	0.693	0.946	0.909	0.834	0.943
	5×5	0.921	0.650	0.932	0.900	0.870	0.932
	7×7	0.911	0.632	0.924	0.889	0.802	0.926
LCC	3×3	0.926	0.740	0.944	0.930	0.883	0.947
	5×5	0.923	0.710	0.932	0.914	0.865	0.933
	7×7	0.909	0.663	0.920	0.909	0.854	0.925
RMSE	3×3	4.986	4.396	5.676	5.539	6.049	5.336
	5×5	4.496	4.792	5.796	5.734	6.293	5.462
	7×7	5.055	5.062	6.087	5.769	6.423	5.569

Shao et al. [53] demonstrate poorer performances on the images with asymmetric distortions. Since this scheme does not need the reference images and human opinion scores, as a result, the objective scores among four distortion types show less correlations. When breaking down to the distortion types, Bensalma et al. [64] perform closely to our proposed metrics on asymmetric images. The reason can be attributed to that it develops a model allowing to reproduce the binocular signal generated by simple and complex cells, and to estimate the associated binocular energy, which has shown a high correlation with the human judgement for different impairments. Additionally, Bensalma et al. [64] and Shao et al. [53] show better on WN and BLUR distortion types, and worse on JP2K and JPEG distortion types on asymmetric images.

Cross dataset test on IVC stereoscopic image dataset: From Table 10, it can be observed that the proposed metric performs the best among 2D and 3D IQAs. Carnec et al. [31], extracting the structure features (orientation, length, width and contrast) of original images in HVS, perform slightly poorer than our proposed three-column CNN. Skeikh et al. [61] present an information fidelity criterion for image quality assessment that relates the image quality with the amount of information shared between a reference and a distorted image. Venkata et al. [62] formulate a nonlinear quasi-local processing model of the HVS by modifying the contrast pyramid, which describes the importance of HVS sensitivity to different visual signals, such as the luminance, the contrast, the frequency content, and the interaction between different signal components. For the 3D IQAs, Campisi et al. [69] conduct a preliminary test on the acuity difference between different eyes, where the metric making use of reliable 2D metrics applied on both the left and the right views is proposed. However, the depth information has not been taken into consideration. Ryu et al. [67] propose an extended version of the SSIM index based on a binocular model, in which luminance, contrast, and structural similarities are computed for original and distorted stereo images. Afterwards, the binocular perception models compute the binocular perceptual luminance, contrast, and structural similarities. However, the two IQAs work on the FR manner, which require the whole original stereoscopic images for quality assessment.

5.2.4. *Effects of convolution kernel size*

In this section, we will examine how the kernel size affects the performance. The kernel size is not only related to the structures learned from the convolution processes but also has great influence on the number of parameters. Therefore, different kernel sizes may produce different IQA performance. The performance with different kernel sizes is illustrated in Table 11. Such testing is conducted on the three-column CNN, where the network structure stays the same and only the kernel size varies. Specifically, the kernel size varies from 3×3 to 7×7. From these experimental

Table 12
SROCC, LCC and RMSE with and without transferring.

Measurement	Algorithm	JP2K	JPEG	WN	BLUR	FF	ALL
SROCC	One-column CNN without transferring	0.702	0.691	0.873	0.577	0.930	0.899
	One-column CNN with transferring	0.889	0.693	0.909	0.877	0.843	0.925
LCC	One-column CNN without transferring	0.692	0.755	0.863	0.624	0.928	0.886
	One-column CNN with transferring	0.898	0.740	0.923	0.928	0.845	0.926
RMSE	One-column CNN without transferring	10.394	9.363	6.524	5.490	6.856	7.673
	One-column CNN with transferring	6.000	5.926	6.107	6.839	6.135	6.148

results, it can be observed that the 3×3 kernel size yields the best performances, while the 7×7 kernel size is the worst. This shows that the convolution process with 3×3 kernel size can better extract the local structures from the stereoscopic images for quality assessment. Another possible reason is that the CNN network with 3×3 kernel size requires a smaller number of parameters, and thus can be well trained with limited training samples.

5.2.5. *Effects of transfer from 2D images to stereoscopic images*

In this section, we examine the effects of transferring parameters learned from 2D natural images to stereoscopic IQA. Specifically, we import the difference image patches into the one-column CNN without any initialization. Another testing is conducted with the initialization of the parameters learned from 2D images before training on the difference image patches. As shown in Table 12, the performance of the IQA with transferring from 2D images is improved considerably. One possible reason is that the number of difference images is limited. So it is hard to train a satisfied CNN model if with no transferring from the 2D natural images. Another possible reason is that the parameters learned from 2D natural image can somewhat capture the image structures, which are also important for stereoscopic perception. Hence, after transferring these parameters to the difference image, not only the disparity/depth information but also the 2D image contents from the left and right views can be learned for stereoscopic IQA.

6. Conclusions

In this paper, we presented an NR IQA metric for stereoscopic images based on convolutional neural networks. With multiple layers of convolution and max-pooling, the local structures can be learned and composed to high level representations. An MLP layer is employed to summarize the representation as a final value to indicate the perceptual quality. Experiments on the public LIVE phase-I, LIVE phase-II, and IVC stereoscopic image datasets demonstrated the superiority of the proposed metric to the state-of-the-art metrics.

Conflict of interest

None declared.

Acknowledgments

This work was supported by the NSFC Grant nos. 61203253, 61573222, 61233014, and 61401167, Open Program of Jiangsu Key Laboratory of 3D Printing Equipment and Manufacturing 3DL201502, Major Research Program of Shandong Province 2015ZDXX0801A02, and Program of Key Lab of ICSP MOE China.

References

[1] X. Liu, W. Yang, L. Lin, Q. Wang, Z. Cai, J. Lai, Data-driven scene understanding with adaptively retrieved exemplars, IEEE Multimed. (2015).
[2] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, L. Lin, S. Yan, Deep human parsing with active template regression, IEEE Trans. Pattern Anal. Mach. Intell. (2015).
[3] P. Seuntiens, L. Meesters, W. Ijsselsteijn, Perceived quality of compressed stereoscopic images: effects of symmetric and asymmetric JPEG coding and camera separation, ACM Trans. Appl. Percept. 3 (April (2)) (2006) 95–109.
[4] L. Lin, R. Zhang, X. Duan, Adaptive scene category discovery with generative learning and compositional sampling, IEEE Trans. Circuits Syst. Video Technol. 25 (2) (2015) 251–260.
[5] W. Lin, C.C.J. Kuo, Perceptual visual quality metrics: a survey, J. Vis. Commun. Image Represent. 22 (4) (2011) 297–312.
[6] J. Guan, W. Zhang, J. Gu, H. Ren, No-reference blur assessment based on edge modeling, J. Vis. Commun. Image Represent. 29 (2015).
[7] A.K. Moorthy, C.C. Su, A. Mittal, A.C. Bovik, Subjective evaluation of stereoscopic image quality, Signal Process.: Image Commun. 28 (8) (2013) 870–883.
[8] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A statistical evaluation of recent full reference quality assessment algorithms, IEEE Trans. Image Process. 15 (November (11)) (2006) 3440–3451.
[9] A. Liu, W. Lin, M. Narwaria, Image quality assessment based on gradient similarity, IEEE Trans. Image Process. 21 (4) (2012) 1500–1512.
[10] M. Narwaria, W. Lin, Objective image quality assessment based on support vector regression, IEEE Trans. Neural Netw. 21 (3) (2010) 515–519.
[11] M. Narwaria, W. Lin, I.V. McLoughlin, S. Emmanuel, L.T. Chia, Fourier transform-based scalable image quality measure, IEEE Trans. Image Process. 21 (8) (2012) 364–3377.
[12] L. Liang, W.S. Wang, J. Chen, S. Ma, D. Zhao, W. Gao, D. Zhao, No-reference perceptual image quality metric using gradient profiles for JPEG2000, Signal Process.: Image Commun. 25 (7) (2010) 502–516.
[13] T. Chen, L. Lin, L. Liu, X. Luo, X. Li, DISC: deep image saliency computing via progressive representation learning, IEEE Trans. Neural Netw. Learn. Syst. (2015).
[14] R. Zhang, L. Lin, R. Zhang, W. Zuo, L. Zhang, Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification, IEEE Trans. Image Process. 24 (December (12)) (2015) 4766–4779.
[15] S. Wang, X. Zhang, S. Ma, W. Gao, Reduced reference image quality assessment using entropy of primitives, in: Picture Coding Symposium, 2013.
[16] D.V. Meegan, L.B. Stelmach, W.J. Tam, Unequal weighting of monocular inputs in binocular combination: implications for the compression of stereoscopic imagery, J. Exp. Psychol.: Appl. 7 (November (2)) (2001) 143–153.
[17] P. Luo, L. Lin, X. Liu, Learning compositional shape models of multiple distance metrics by information projection, IEEE Trans. Neural Netw. Learn. Syst. (2015).
[18] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, S. Yan, Matching-CNN meets KNN: Quasi-parametric human parsing, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[19] W.J. Tam, L.B. Stelmach, P.J. Corriveau, Psychovisual aspects of viewing stereoscopic video sequences, in: Proc. SPIE 3295 (January) (1998) 226–235.
[20] M.J. Chen, A.C. Bovik, L.K. Cormack, Study on distortion conspicuity in stereoscopically viewed 3D images, in: Proceedings of the IEEE IVMSP Workshop, June 2011, pp. 24–29.
[21] M.J. Chen, D.K. Kwon, L.K. Cormack, A.C. Bovik, Optimization the 3D image display using the stereoacuity function, in: IEEE International Conference on Image Processing, October. 2012, pp. 617–620.
[22] M.J. Chen, D.K. Kwon, A.C. Bovik, Study of subject agreement on stereoscopic video quality, in: IEEE Southwest Symposium on Image Analysis and Interpretation, April 2012, pp. 173–176.
[23] W. Chen, J. Fournier, M. Barkowsky, P. Le Callet, Quality of experience model for 3DTV, Proc. SPIE 8288 (March) (2012) 1–9.
[24] M.T.M. Lambooij, W.A. Ijsselsteijn, I. Heynderickx, Visual discomfort in stereoscopic displays: a review, Proc. SPIE 6490 (April) (2007) 17.
[25] S. Liu, X. Liang, L. Liu, K. Lu, L. Lin, X. Cao, S. Yan, Transferred human parsing with video context, IEEE Trans. Multimed. 17 (8) (2015) 1347–1358.
[26] K. Wang, X. Wang, L. Lin, M. Wang, W. Zuo, 3D activity recognition with reconfigurable convolutional neural networks, in: ACM International Conference on Multimedia, 2014.

[27] W.J. Tam, F. Speranza, S. Yano, K. Shimono, H. Ono, Stereoscopic 3D-TV: visual comfort, IEEE Trans. Broadcast. 57 (April (2)) (2011) 335–346.
[28] C. Hewage, S.T. Worrall, S. Dogan, A.M. Kondoz, Prediction of stereoscopic video quality using objective quality models of 2-D video, Electron. Lett. 44 (July (16)) (2008) 963–965.
[29] P. Gorley, N. Holliman, Stereoscopic image quality metrics and compression, Proc. SPIE 6803 (February) (2008) 5.
[30] A. Benoit, P. Le Callet, P. Campisi, R. Cousseau, Quality assessment of stereoscopic images, EURASIP J. Image Video Process. 2008 (2009) 1–13.
[31] M. Carnec, P. Le Callet, D. Barba, An image quality assessment method based on perception of structural information, in: IEEE International Conference on Image Processing, vol. 3, September 2003, pp. 185–193.
[32] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: Proceedings of the Asilomar Conference on Signals, Systems, and Computers, vol. 2, November 2003, pp. 1398–1402.
[33] J. You, L. Xing, A. Perkis, X. Wang, 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, pp. 1–6.
[34] J. Yang, C. Hou, Y. Zhou, Z. Zhang, J. Guo, Objective quality assessment method of stereo images, in: 3DTV Conference, May 2009, pp. 1–4.
[35] M.J. Chen, D.K. Su, C.C. Kwon, L.K. Cormack, A.C. Bovik, Full-reference quality assessment of stereopairs accounting for rivalry, in: Proceedings of the Asilomar Conference on Signals, Systems and Computers, November 2012, pp. 1–5.
[36] C. Hewage, M. Martini, Reduced-reference quality metric for 3D depth map transmission, in: 3DTV Conference, June 2010, pp. 1–4.
[37] R. Akhter, J. Baltes, Z.M. Parvez Sazzad, Y. Horita, No reference stereoscopic image quality assessment, Proc. SPIE 7524 (February) (2010).
[38] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, in: European Conference on Computer Vision, 2014, pp. 346–361.
[39] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv:1409.1556arXiv:1409.1556, 2014.
[40] C. Szegedy, W. Liu, Y. Jia, P. Sermannet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going Deeper with Convolutions, arXiv:1409.4842arXiv:1409.4842, 2014.
[41] K. He, X. Zhang, S. Ren, J. Sun, Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification, arXiv:1502.01852arXiv:1502.01852, 2015.
[42] W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Y. Xiong, C. Qian, Z. Zhu, R. Wang, C.C. Loy, X. Wang, X. Tang, DeepID-net: Multi-stage and Deformable Deep Convolutional Neural Networks for Object Detection, arXiv:1409.3505arXiv:1409.3505, 2014.
[43] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[44] L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional neural networks for no-reference image quality assessment, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1733–1740.
[45] L. Ma, X. Wang, Q. Liu, K.N. Ngan, Reorganized DCT-based image representation for reduced reference stereoscopic image quality assessment, Neurocomputing, Accepted.
[46] L. Ma, S. Li, F. Zhang, K.N. Ngan, Reduced-reference image quality assessment using reorganized dct-based image representation, IEEE Trans. Multimed. 13 (August (4)) (2011) 824–829.
[47] L. Ma, S. Li, K.N. Ngan, Reduced-reference video quality assessment of compressed video sequence, IEEE Trans. Circuits Syst. Video Technol. 22 (October (10)) (2012) 1441–1456.
[48] W. Zhang, K. Liu, W. Zhang, Y. Zhang, J. Gu, Wi-Fi positioning based on deep learning, in: International Conference on Information and Automation, 2014, pp. 1176–1179.
[49] Y. LeCun, Y. Bengio, Convolutional networks for images, speech and time series, The Handbook of Brain Theory and Neural Networks (1995).
[50] M.J. Chen, L.K. Cormack, A.C. Bovik, No-reference quality assessment of natural stereopairs, IEEE Trans. Image Process. (2013) 3379–3391.
[51] W. Zhou, G. Jiang, M. Yu, Z. Wang, Z. Peng, F. Shao, Reduced reference stereoscopic image quality assessment using digital watermarking, Comput. Electr. Eng. (2014) 104–116.
[52] X. Wang, Q. Liu, R. Wang, Z. Chen, Natural image statistics based 3D reduced reference image quality assessment in contourlet domain, Neurocomputing (2015) 683–691.
[53] F. Shao, W. Lin, S. Wang, G. Jiang, M. Yu, Blind image quality assessment for stereoscopic images using binocular guided quality lookup and visual codebook, IEEE Trans. Broadcast. (2015) 154–165.
[54] S. Ryu, K. Sohn, No-reference quality assessment for stereoscopic images based on binocular quality perception, IEEE Trans. Circuits Syst. Video Technol. (2014) 591–602.
[55] L. Ma, Z. Lu, L. Shang, H. Li, Multimodal convolutional neural networks for matching image and sentence, in: International Conference on Computer Vision, 2015.
[56] L. Ma, Z. Lu, H. Li, Learning to answer questions from image using convolutional neural network, in: The 30th AAAI Conference on Artificial Intelligence, 2016.
[57] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.
[58] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[59] H. Zhang, F. Zhou, W. Zhang, X. Yuan, Z. Chen, Real-time action recognition based on a modified deep belief network model, in: International Conference on Information and Automation, 2014, pp. 225–228.
[60] Z.M.P. Sazzad, S. Yamanaka, Y. Horita, Spatio-temporal segmentation based continuous no-reference stereoscopic video quality prediction, in: International Workshop on Quality of Multimedia Experience, 2010, pp. 106–111.
[61] H.R. Sheikh, A.C. Bovik, G. de Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Trans. Image Process. (2005) 2117–2128.
[62] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, A.C. Bovik, Image quality assessment based on a degradation model, IEEE Trans. Image Process. (2000) 636–650.
[63] Z. Wang, A.C. Bovik, A universal image quality index, IEEE Signal Process. Lett. (2002) 81–84.
[64] R. Bensalma, M.-C. Larabi, A perceptual metric for stereoscopic image quality assessment based on the binocular energy, Multidimens. Syst. Signal Process. (2013) 281–316.
[65] L. Zhang, D. Zhang, X. Mou, D. Zhang, FSIM: a feature similarity index for image quality assessment, IEEE Trans. Image Process. (2011) 2378–2386.
[66] W. Xue, L. Zhang, X. Mou, A.C. Bovik, Gradient magnitude similarity deviation: a highly efficient perceptual image quality index, IEEE Trans. Image Process. (2014) 684–695.
[67] S. Ryu, D.H. Kim, K. Sohn, Stereoscopic image quality metric based on binocular perception model, in: IEEE International Conference on Image Processing, 2012, pp. 609–612.
[68] W. Zhang, Y. Zhang, L. Ma, J. Guan, S. Gong, Multimodal learning for facial expression recognition, Pattern Recognit. (2015) 3191–3202.
[69] P. Campisi, P. Le Callet, E. Marini, Stereoscopic images quality assessment, in: European Signal Processing Conference, 2007, pp. 2110–2114.
[70] B. Schauerte, G. A. Fink, Web-based learning of naturalized color models for human–machine interaction, in: International Conference on Digital Image Computing: Techniques and Applications, 2010, pp. 1–3.
[71] B. Schauerte, R. Stiefelhagen, Learning robust color name models from web images, in: International Conference on Pattern Recognition, 2012, pp. 11–15.
[72] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. (2004) 600–612.

Wei Zhang is an associate professor with the School of Control Science and Engineering, Shandong University. He received his Ph.D. at The Chinese University of Hong Kong (CUHK). He previously worked as a postdoc scholar at the University of California, Berkeley (UC Berkeley). He is a member of IEEE and CSIG, and was appointed as Distinguished Expert by Shandong government. He is the finalist of Hong Kong Young Scientist 2010 and received several international awards from IEEE. He served as the chair or steering committee member of many famous international conferences such as CVPR, ICCV, ECCV, ICIP, ROBIO, ICIA, and ICAL.

Chenfei Qu is currently an M.Phi. candidate with the School of Control Science and Engineering, Shandong University.

Lin Ma is now a researcher at Huawei Noah's Ark Lab, Hong Kong. He received his Ph.D. degree in the Department of Electronic Engineering at the Chinese University of Hong Kong (CUHK) in 2013. He received the B.E., and M.E. degrees from Harbin Institute of Technology, Harbin, China, in 2006 and 2008, respectively, both in computer science. He was a research intern in Microsoft Research Asia from October 2007 to March 2008. He was a research assistant with the Department of Electronic Engineering, CUHK, from November 2008 to July 2009. He was a visiting student with the School of Computer Engineering, Nanyang Technological University (NTU), from July 2011 to September 2011. His research interests lie in the areas of deep learning and multimodal learning, specifically for image and language, image/video processing and quality assessment. He got the best paper award in Pacific-Rim Conference on Multimedia (PCM) 2008. He was awarded the Microsoft Research Asia fellowship in 2011. He was a finalist to HKIS young scientist award in engineering science in 2012.

Jingwei Guan is currently a Ph.D. candidate in the Department of Electronic Engineering of The Chinese University of Hong Kong (CUHK). She received the Bachelor of Engineering degree from Shandong University, China, in 2013.

Rui Huang is currently a research staff member at NEC Laboratories China. Before Joining NEC, he was an assistant professor at Huazhong University of Science and Technology when he first came back to China after living in the US for 8 years, during which period he got his Ph.D. degree (2008) and worked as a research associate in the Department of Computer Science, Rutgers University. He received his B.Sc. degree at Peking University (1999), and M.Eng. degree at Chinese Academy of Sciences (2002).