https://doi.org/10.2352/ISSN.2470-1173.2021.9.IQSP-297

No-reference Stereoscopic Image Quality Predictor using Deep Features from Cyclopean Image

Oussama Messai +, Aladine Chetouani •, Fella Hachouf +, Zianou Ahmed Seghir *
+ Laboratoire ARC, Université des Frères Mentouri Constantine 1, Algérie.
• PRISME Laboratory, University of Orleans, France.
**** Computing Department, University of Abbes laghrour Khenchela, Algeria.***

Abstract

With the expanding use of stereoscopic imaging for 3D applications, no-reference perceptual quality evaluation has become important to provide good viewing experience. The effect of the quality distortion is related to the scene's spatial details. Taking this into account, this paper introduces a blind stereoscopic image quality measurement using synthesized cyclopean image and deep feature extraction. The proposed method is based on Human Visual System (HVS) modeling and quality-aware indicators. First, the cyclopean image is formed, taking on the existence of binocular rivalry / suppression that includes the asymmetric distortion case. Second, the cyclopean image is decomposed into four equivalent parts. Then, four Convolutional Neural Network (CNN) models are deployed to automatically extract quality feature sets. Finally, a feature bank is then created from the four patches and mapped to quality score using a Support Vector Regression (SVR) model. The best known 3D LIVE phase I and phase II databases were used to evaluate the efficiency of our technique. Compared with the state-of-the-art stereoscopic image quality measurement metrics, the proposed method has shown competitive outcomes and achieved good performance.

Introduction

In order to display 3D content, stereo imaging technique is usually used and has attracted considerable interest in the last few years. This stereoscopic visualization technique can be used in many applications (e.g., robotic navigation, medical surgery and entertainments.) [1, 2]. However, based on the MPAA statistics report, the amount of stereo content will continue to increase for the next couple of years [3].

Along with the 3D content growth, measurement of quality or visual discomfort is very important to guarantee a good experience for the viewer/user. However, during acquisition and processing, the stereo image quality could be affected by various distortions such as Blur, JPEG, fast fading, etc. Hence, methods that assess the quality/discomfort of stereoscopic content is needed. However, Stereo Image Quality Assessment (SIQA) methods are classified into two types; subjective SIQA and objective SIQA. The former methods are expensive and time consuming since they involve human score opinion for judging the quality, while the latter are cheap and fast because they rely on machine algorithmic score. On the other hand, objective SIQA can be divided into three classes; Full-Reference (FR), Reduced-Reference (RR) and No-Reference (NR) metrics. However, because human is the final recipient of the 3D content, it is necessary to verify the metric output with the subjective evaluation which is the human visual quality assessment. It is expressed mainly in terms of Mean Opinion Score (MOS) or Difference Mean Opinion Score (DMOS).

In this study, we focus on objective NR Stereo Image Quality methods and the aim is to design a deep learning-based method that can well simulate the human judgment. Recently, many researchers have used cyclopean image hypothesis for SIQA. Therefore, perception mechanism of our brain is simulated with cyclopean view hypothesis as done in previous works [4, 5, 6]. The adopted cyclopean image model considers the binocular rivalry which often makes the observers exhaustion and visual discomfort. Through the usage of this latter, we propose a new objective SIQA metric based on the combination of Deep Learning-based feature sets, extracted from a four CNN models. The overall quality index is computed using a SVR.

The rest of this paper is organized as follows. Section presents related work. The overall framework of the proposed model is described in Section . Section discusses experimental results. Finally, section concludes the paper.

Related Work

Over the past, a lot of effort has been conducted into how the Human Visual System (HVS) handles the signals that the eyes receive. The SIQA designs can be divided into two categories. The first concept uses the 2D IQA methods to measure stereo IQ by computing the mean predicted quality scores of the left and right image [7, 8, 9]. The second category takes assumption that stereoscopic image quality may not be accurate when computing the mean of the two images scores [10]. This concept focuses on human visual system modeling where depth data is used. Indeed, the research has shown that the quality of visual stereoscopic contents is linked to the quality of depth information [11]. Authors in [12] have proposed a full-reference SIQA model, they measure the difference between the left and right reference images and the distorted ones, then compute the difference between the disparity map of the reference stereo image and the distorted ones. Similar FR SIQA has been developed in [13]. Authors deployed multiple 2D IQA metrics to predict quality scores from disparity map and stereoscopic images . Then fuse the scores to get final prediction. Moreover, IQA metric called Binocular Energy Quality Metric (BEQM) has been proposed in [14]. They predicted stereoscopic image quality by calculating the binocular energy variation between the reference and distorted stereo images. PSNR-based

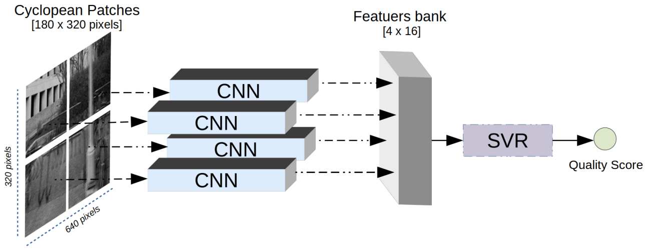

Figure 1: **Flowchart of the proposed method.**

metrics for SIQA have been proposed by [15, 16]. Hewage *et al.* [16] computed the PSNR between the reference and distorted disparity maps to assess the quality. Meanwhile, Gorley *et al.* [15] did not deploy the disparity information but rather they calculated quality scores on corresponded feature points provided by SIFT [17] and RANSAC [18], in-which These feature points are derived from the left and right views.

One of the issues with stereoscopic images is binocular rivalry. It mostly causes discomfort and visual frustration to observers [19]. FR SIQA method proposed in [20] that used the linear expression of cyclopean view [21] influenced by binocular suppression/rivalry between left and right images. Zhou *et al.* [22] also simulated binocular visual system and proposed a NR SIQA. Fang et al. [23] suggested an unsupervised references-less metric for stereo images. Authors extracted quality indicators in spatial and frequency domains from both monocular and cyclopean view patches. Then used statistical distance method named Bhattacharyya to get quality score. Meanwhile, other methods that depend on the depth information have been introduced. For instance, Akhter *et al.* [24] have designed blind SIQA algorithm, they mapped characteristics derived from disparity map and stereo image to quality ratings. More recently, an advanced features and complex combination were used to develop modern NR SIQA metrics. For example, Karimi *et al.* [25] combined statistical features derived from a synthesized phase/shift and contrast images. While in [26], monocular super-pixel and Natural Scene Statistics (NSS) features were concatenated with other binocular features. These features are derived from the stereoscopic image and fed to a regression model for quality prediction. Another work that relies on binocular features in [27], where visual saliency, local magnitude, and local phase are extracted from the stereo image as basic feature vectors.

Most of the above SIQA approaches use handcrafted quality features that are derived manually from the stereoscopic picture. With the use of Deep Learning, the suggested approach allows learning quality features from the input data automatically. However, a work has been done in [28] that explore this concept and propose NR SIQA metric. The authors pursued a two steps of training. They first trained the CNN model to extract features from small stereo-pair image patches. The model is then followed by feature concatenation layer and regression layer for second training to predict the quality.

Proposed Method

The proposed metric as shown in Fig. 1 involves three simple steps: in the first step, the cyclopean image is computed. Second, divide the cyclopean image into four equivalent parts and train four CNN models that generate a feature bank. Third, the quality score is predicted from the extracted features using a SVR.

Cyclopean Image

The HVS receives visual stimuli from both eyes and integrates it into one 3D vision [19]. Studies on human binocular visual system is needed to develop SIQA model. Therefore, considering the binocular suppression/rivalry, the cyclopean view hypothesis is utilized. Binocular rivalry occurs when monocular left and right stimuli are different. Therefore, the quality of stereo image being viewed may not be determined easily by the mean of left and right views quality scores. A remarkable explanation for this phenomenon has been established by Levelt [21]. The authors have carried out several tests that demonstrate how high binocular suppression is regulated by low level sensory factors. They found that visual contents with more contours or high contrast tend to dominate the rivalry. With regard to this result, a Gabor filter is appropriate for extracting specific contours using frequency and orientation parameters. The response of this filter mimics the suppression selection of the cyclopean image while it is computed. Fig. 2 shows a cyclopean image computed from the undistorted left image and right image that is distorted. The red boxes in figure zoom into the same location of each view. It can be noted from the figure that the asymmetric distortion is clearly stated in the formed cyclopean image. However, a study performed in [29] investigates how basic cells in the visual cortex can be represented by Gabor functions. Thus, the Gabor filters are utilized to simulate vision of the human visual system. The Gabor filter design used in the suggested approach is the same as used in previous work [4].

To construct the cyclopean image, we have utilized the mathematical model used in [20], which is :

$$C(x,y) = w_l(x,y) \times I_l(x,y) + w_r(x+d,y) \times I_r(x+d,y) \tag{1}$$

where I_l and I_r are the left and right views, respectively. w_l and w_r are the weighting coefficients for the left and right eyes, respectively. The weights are computed from the Gabor filter bank responses. d is the disparity index that matches pixels from left image I_l with those in the right image I_r.

Deep Feature Extraction

In recent years, deep learning has been deployed to solve difficult problems such as image classification and speech recognition. The end-to-end network allows to extract automatically relative features from the raw data showing significant accuracy improvement in the IQA domain.

Generally, at each region corner of the cyclopean image, the structure differs e.g. textures, color and pixel intensities. As we want to derive various quality features, we simply divide the input cyclopean image into four equivalent patches. This partition covers the four corners and deals with different structures individually. Four CNNs are then needed to extract quality feature sets from each structure. In the case of using just one CNN, the model will tend to extract the general characteristics since the network weights remain the same. The four trained CNNs have similar architecture but different weights, thus they enrich the features bank as shown in Fig. 1. Meanwhile, we assume that using two models will provide fewer quality indicators than four.

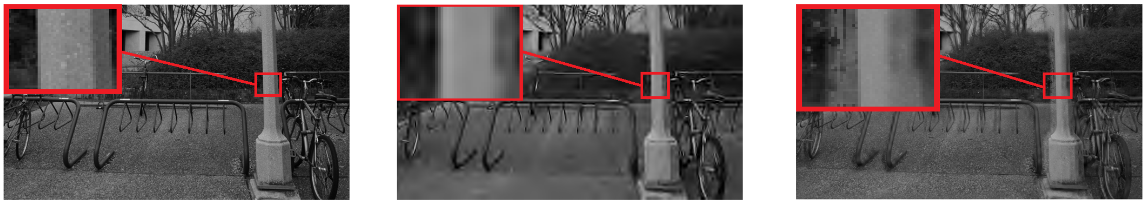

Figure 2: **(a) Left image without distortion, (b) Right image JPEG distortion, (c) cyclopean image of both images. For each view, red box is zoomed to the left for better visualization.**

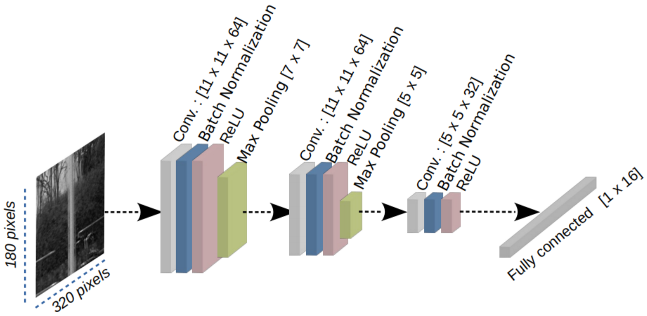

Figure 3: **The proposed Convolutional Neural Network architecture for feature extraction.**

For the feature extraction, we design a light-weight CNN model from scratch and compare its performance with most common pre-trained models: AlexNet [30], VGG-16 and VGG-19 [31], Resnet18 [32], Inception-v1 [33].

The cyclopean image is thus fed to the CNN models to to extract quality-aware indicators. Each CNN expects an input of size 180 × 320 pixels. For each patch, one CNN model is trained and used to extract a vector of size 1×16. The suggested CNN architecture consists of 12 layers as shown on Fig. 3, after each convolution layer, a batch normalization layer is applied to speed up the learning [34], followed by a ReLU layer as activation function and Max-pooling layer to reduce dimensionality. The network includes three convolutional layers. The first and second convolution layers produce 64 filters of size [11 × 11]. While the third convolution layer has 32 filters of size [5 × 5]. All convolution layers have 4 pixels stride in both horizontal and vertical directions. The first used Max-pooling layer has a size of [7 × 7] and a stride of [2×2] pixels. The second Max-pooling layer has a size of [5 × 5] and a stride of [1×1] pixel. After all, a fully connected 1 ×16 layer is used to provide 16 elements quality indicators. The four networks are trained for 150 epochs with a learning rate of 0.01. Stochastic Gradient Descent (SGD) with momentum has been applied to update the network weights. The extracted feature vectors [4×16] is then fed to a SVR model with a Gaussian kernel function to predict the quality scores.

Table 1: **PLCC correlation results of our CNN regression model vs. CNN + SVR combination over the four patches from LIVE 3D II database.**

Number of patch	1	2	3	4	All
CNN + FC regressor	0.918	0.920	0.922	0.905	0.910
CNN + SVR regressor	0.927	0.928	0.932	0.907	0.932

Table 2: **Performance of different pre-trained feature extractors on LIVE-II database.**

Model	PLCC	SROCC	RMSE
AlexNet	0.922	0.921	4.355
VGG-16	**0.948**	**0.941**	**3.817**
VGG-19	0.946	0.938	3.888
Resnet18	0.930	0.930	4.122
Resnet50	0.940	0.939	3.894
Inception-v1	0.938	0.939	3.897

Experimental Results

Datasets and training protocol

Two databases were used to test the efficiency of our metric, namely the LIVE 3D phase I and phase II databases. The former contains symmetrically distorted stimuli, while the latter includes both symmetrically and asymmetrically distorted stimuli. The LIVE 3D phase I [35] consists of 365 distorted stereo images of size 360 x 640 pixels. Five degradation types have been considered (White Noise: WN, JPEG2000: JP2K, JPEG, and Fast Fading: FF and Blur). All distortions are performed symmetrically. The LIVE 3D phase II [20] contains 360 distorted stereoscopic images have the same size as LIVE 3D phase I. It contains asymmetric and symmetrical stereoscopic distortions. Together the two databases form 725 distorted stereoscopic images. We normalize the train set outputs (DMOS) to min-max normalization [0 to 1], where the closest to zero the better quality is. The 5-fold cross validation technique has been adopted. The dataset is split into 5 folds, where each fold is divided to 80% train set and 20% test set chosen randomly. The protocol has been repeated for 10 iterations to show the generalization ability of our method and the mean performance values are reported.

The performance has been measured across three metrics: The *RMSE*, the Pearson linear correlation coefficient (*PLCC*), the Spearman's rank order correlation coefficient (*SROCC*) between the machine quality judgments (objective scores) and the human ratings (subjective scores). (*DMOS*). *RMSE* and *PLCC* measure the assessment accuracy, while *SROCC* evaluates the prediction notability. High values for *PLCC* and *SROCC* (close to 1) and low values for *RMSE* (close to 0). LCC indicates the linear order similarity between human quality opinions and the metrics predictions. While SROCC reveals the accuracy of the methods.

Quality evaluation

We first evaluated the relevance to use SVR as regressor instead of FC layers, usually done. To this end, the SVR has been replaced by a FC regression layer. Table 1 indicates the comparison PLCC correlation results of the designed network model on LIVE 3D II database. The combination of CNN and SVR has increased

Table 3: **SROCC results on the 3D LIVE Phase-I and Phase-II databases.**

		LIVE I						LIVE II					
Method	Type	WN	JP2K	JPEG	Blur	FF	All	WN	JP2K	JPEG	Blur	FF	All
Benoit [12]	FR	0.923	0.751	0.867	0.455	0.773	0.728	0.923	0.751	0.867	0.455	0.773	0.728
You [13]		0.909	0.894	0.795	0.813	0.891	0.786	0.909	0.894	0.795	0.813	0.891	0.786
Gorley [15]		0.875	0.110	0.027	0.770	0.601	0.146	0.875	0.110	0.027	0.770	0.601	0.146
Chen [20]		0.940	0.814	0.843	0.908	0.884	0.889	0.940	0.814	0.843	0.908	0.884	0.889
Hewage [16]		0.880	0.598	0.736	0.028	0.684	0.501	0.880	0.598	0.736	0.028	0.684	0.501
Bensalma [14]		0.905	0.817	0.328	0.915	0.915	0.874	0.938	0.803	0.846	0.846	0.846	0.751
Akhter [24]	NR	0.714	0.724	0.649	0.682	0.559	0.543	0.714	0.724	0.649	0.682	0.559	0.543
Zhou [22]		0.921	0.856	0.562	0.897	0.771	0.901	0.936	0.647	0.737	0.911	0.798	0.819
Fang [23]		0.883	0.880	0.523	0.523	0.650	0.877	0.955	0.714	0.709	0.807	0.872	0.838
Chen [27]		0.926	0.839	0.832	0.951	0.918	0.920	0.910	0.825	0.843	0.929	0.896	0.852
Kim [28]		-	-	-	-	-	-	0.922	0.885	0.763	0.932	**0.945**	0.938
Karimi [25]		0.945	0.917	0.750	0.919	**0.837**	0.947	0.953	0.875	0.832	0.874	0.907	0.913
Liu [26]		0.951	0.888	0.785	0.917	0.821	0.928	0.946	**0.909**	0.825	**0.936**	0.938	0.901
Proposed		0.925	0.921	0.666	0.924	0.799	0.928	0.928	0.897	0.809	0.900	0.880	0.909
Proposed vgg-16		**0.964**	**0.943**	**0.834**	**0.953**	0.803	**0.956**	**0.959**	0.888	**0.875**	0.935	**0.945**	**0.948**

Table 4: **PLCC results on the 3D LIVE Phase-I and Phase-II databases.**

		LIVE I						LIVE II					
Method	Type	WN	JP2K	JPEG	Blur	FF	All	WN	JP2K	JPEG	Blur	FF	All
Benoit [12]	FR	0.926	0.784	0.853	0.535	0.807	0.784	0.926	0.784	0.853	0.535	0.807	0.784
You [13]		0.912	0.905	0.830	0.784	0.915	0.800	0.912	0.905	0.830	0.784	0.915	0.800
Gorley [15]		0.796	0.485	0.312	0.852	0.364	0.451	0.322	0.372	0.874	0.934	0.706	0.515
Chen [20]		0.957	0.834	0.862	0.963	0.901	0.907	0.957	0.834	0.862	0.963	0.901	0.907
Hewage [16]		0.891	0.664	0.734	0.450	0.746	0.558	0.891	0.664	0.734	0.450	0.746	0.558
Bensalma [14]		0.914	0.838	0.838	0.838	0.733	0.887	0.943	0.666	0.857	0.907	0.909	0.769
Akhter [24]	NR	0.772	0.776	0.786	0.795	0.674	0.568	0.929	0.772	0.776	0.786	0.795	0.674
Zhou [22]		-	-	-	-	-	0.929	-	-	-	-	-	0.856
Fang [23]		0.900	0.911	0.547	0.903	0.718	0.880	0.961	0.740	0.764	0.968	0.867	0.860
Chen [27]		-	-	-	-	-	0.937	-	-	-	-	-	0.937
Kim [28]		-	-	-	-	-	-	0.910	0.910	0.768	0.951	0.957	**0.941**
Karimi [25]		0.955	0.939	0.771	0.959	0.882	**0.956**	0.966	0.897	0.866	0.957	0.918	0.923
Liu [26]		0.966	0.938	0.810	0.956	0.855	0.945	**0.969**	0.936	0.867	**0.987**	**0.959**	0.913
Proposed		0.936	0.905	0.811	**0.967**	**0.887**	0.911	0.931	**0.944**	0.689	0.951	0.851	0.932
Proposed vgg-16		**0.970**	**0.960**	**0.845**	0.962	0.865	0.955	0.959	0.887	**0.888**	0.981	0.931	**0.941**

Table 5: **RMSE results on the 3D LIVE Phase-I and Phase-II databases.**

		LIVE I						LIVE II					
Method	Type	WN	JP2K	JPEG	Blur	FF	All	WN	JP2K	JPEG	Blur	FF	All
Benoit [12]	FR	4.028	6.096	3.787	11.763	6.894	7.490	4.028	6.096	3.787	11.763	6.894	7.490
You [13]		4.396	4.186	4.086	8.649	4.649	6.772	4.396	4.186	4.086	8.649	4.649	6.772
Gorley [15]		5.202	9.113	6.940	4.988	8.155	9.675	5.202	9.113	6.940	4.988	8.155	9.675
Chen [20]		3.368	5.562	3.865	3.747	4.966	4.987	3.368	5.562	3.865	3.747	4.966	4.987
Hewage [16]		10.713	7.343	4.976	12.436	7.667	9.364	10.713	7.343	4.976	12.436	7.667	9.364
Bensalma [14]		-	-	-	-	-	7.558	-	-	-	-	-	7.203
Akhter [24]	NR	7.416	6.189	4.535	8.450	8.505	9.294	7.416	6.189	4.535	8.450	8.505	9.294
Zhou [22]		-	-	-	-	-	6.010	-	-	-	-	-	6.041
Fang [23]		-	-	-	-	-	7.191	-	-	-	-	-	5.767
Karimi [25]		5.017	4.644	4.290	4.458	**5.997**	4.998	**2.936**	5.083	4.071	4.581	4.974	4.436
Liu [26]		-	-	-	-	-	5.268	-	-	-	-	-	7.658
Proposed		6.046	4.246	4.725	4.419	6.502	5.905	3.770	**4.160**	4.280	3.506	5.288	4.629
Proposed vgg-16		**4.013**	**3.597**	**3.488**	**3.943**	6.223	**4.865**	3.002	4.526	**3.366**	**2.685**	**4.176**	**3.817**

Table 6: **Asymmetric versus Symmetric SROCC results on 3D LIVE Phase II database.**

Distortion Type	Benoit [12]	You [13]	Gorley [15]	Chen [20]	Hewage [16]	Bensalma [14]	*Akhter* [24]	*Proposed*	*Proposed vgg-16*
Symmetric	0.860	0.914	0.383	0.923	0.656	0.841	0.420	0.921	**0.936**
Asymmetric	0.671	0.701	0.056	0.842	0.496	0.721	0.517	0.909	**0.953**

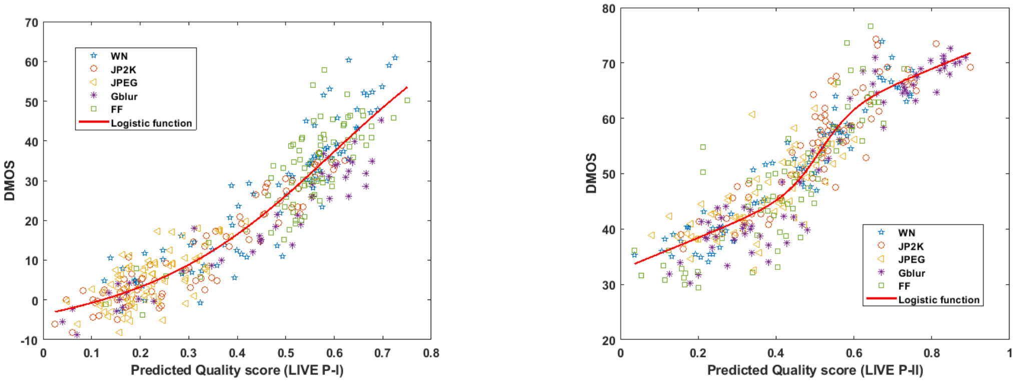

Figure 4: **Scatter plot of subjective scores DMOS against scores from the proposed metric using the designed CNN on LIVE 3D Phase I and Phase II databases.**

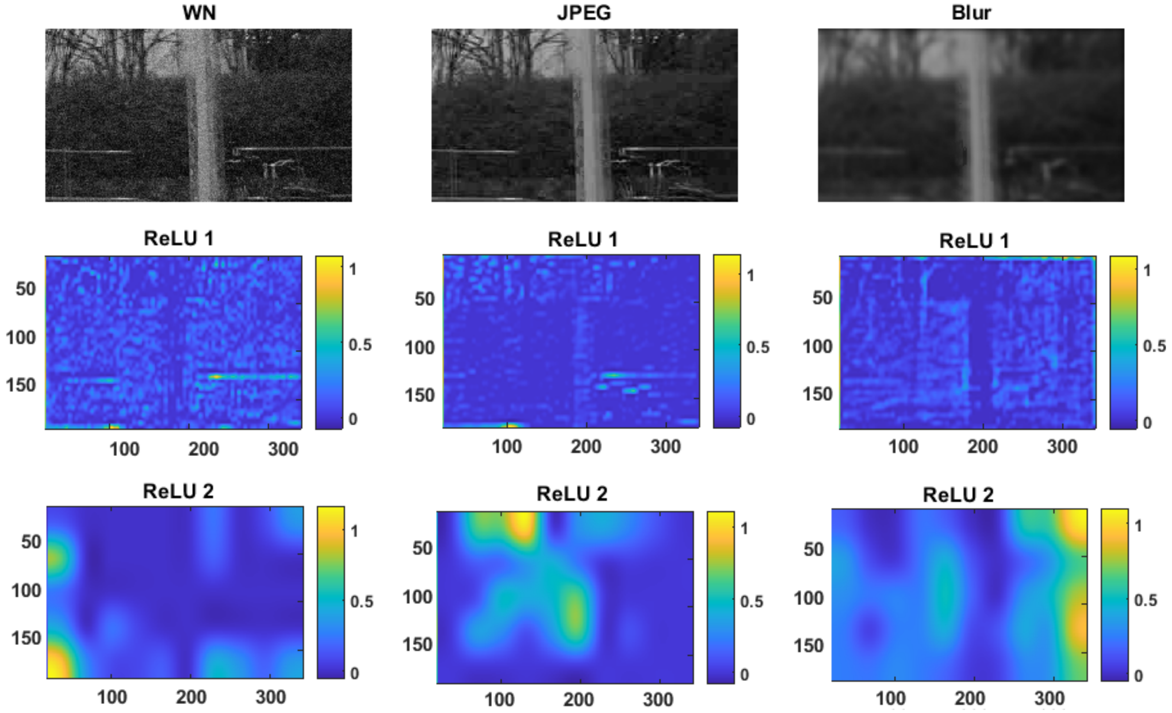

Figure 5: **The first and second ReLU activation layer outputs from a test cyclopean patch for three degradation types.**

the quality prediction accuracy compared to CNN model alone. In addition, the use of four different CNN models enrich the quality features bank to improve the overall quality prediction. Although the proposed CNN architecture ends up with PLCC of 0.932 over LIVE-II dataset, we furthermore test and investigate the performance of six common pre-trained CNN models. Where the same training protocol and configurations of our designed model were used. Each pre-trained model has been adjusted and then used to extract [1 x 16] feature vector from each patch. Table 2 presents these experiments using the LIVE-II database and the best ranked extractor was found to be vgg-16. Overall, the pre-trained models except alexnet outperform our CNN design which was expected since the pre-trained CNNs are large and deeper networks. For instance, vgg-16 has about 138 million (approx) parameters while alexnet has around 62 millions. Alexnet gives the lowest correlation performance among all models. The vgg-16 and vgg-19 yield similar correlation performance with little differences since they have nearly the same architecture. These models contain more series of convolutional layers than our architecture and thus extract higher and better quality indicators for prediction. In the meantime, going deeper than vgg-16 model, resnet and inception extractors appear to slightly diverge from the path toward the best indicators. However, our built CNN is almost two times faster than vgg-16. The run-time indicates 108 ms (milliseconds) for vgg-16, and 56 ms for our CNN using the same hardware and stereoscopic image. With the provided competitive performance, this will be beneficial in case of limited resources. Otherwise, the implementation of vgg-16 would be better choice.

Our method has been then compared with several FR and NR SIQA metrics, including six FR and seven NR SIQA metrics. Tables 3, 4 and 5 show the results of all SIQA algorithms on LIVE 3D phase I and phase II databases. The best outcome of NR category is highlighted in bold. We reported the outcomes of using the scratched CNN and the pre-trained vgg-16. The results

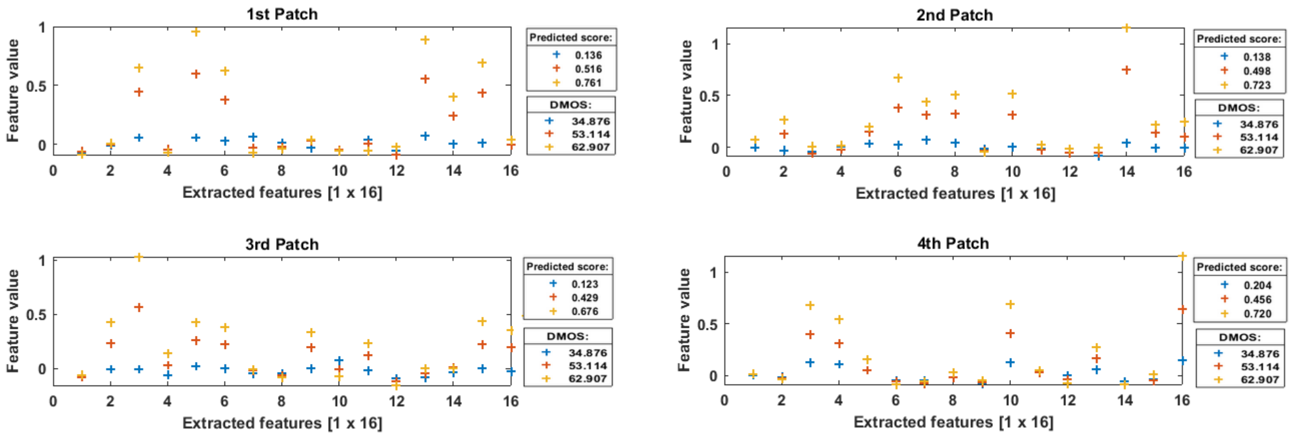

Figure 6: **Extracted features bank from three cyclopean images of the same scene. Each plot represents the sixteen extracted quality indicators from a different patch. The first to the fourth patch from above to below, respectively.**

obtained on LIVE 3D Phase I show the efficiency of our method, since it outperforms all the compared metrics in terms of SROCC and RMSE, including FR ones. For FF distortions, Karimi *et al.* [25] metric obtained better results, but our method remains the best on the rest type of distortions. Meanwhile in term of PLCC, our method has the best correlation on the five distortions.

For LIVE 3D phase II, the same behaviour has been noticed with the best overall performance and competitive results on the degradation types. Compared to the results on LIVE 3D Phase I, we obtained higher Spearman's correlations for WN, JPEG and FF, but still not high as other degradation types.

Table 6 shows the performance on symmetric and asymmetric distorted stimuli. As can be seen, performances of all method are often higher for symmetric distribution. Our method outperforms most of the compared FR and NR metrics with 0.936 and 0.953 as SROCC for symmetric and asymmetric distributions, respectively. Hence, the proposed scheme has good correlation with human subjective evaluation across four types of distortion as well as symmetric/asymmetric distributions. Scatter plots that exhibit the prediction responses against human score (DMOS) on LIVE 3D phase I and phase II are given in Fig. 4. As can be seen, the distribution of the predicted scores well fit the DMOS with low dispersion. According to the different degrees of deformations/distortions. Each distortion type scores are well spread according to human predictions. This can show a consistency performance for all types distortion.

Quality indicators visualization

In this section, we investigate the extracted features by the designed networks (shown in Fig. 3), and examine which parts of the cyclopean image are most important for our CNN models. A patch was chosen from the cyclopean which formed using distorted stereo images. These latter are fed to a trained CNN model as test patches and then inspect the outputs of activation functions (ReLU) after the first and second convolutional layers. The convolution layer produces 64 channels. Among the 64 channels output from ReLU layer, their mean values are computed and the strongest channel has been selected by indexing the maximum. Fig. 5 despite the first and second ReLU layer responses for the input cyclopean patch that were constructed under three types of distortions: WN, JPEG, and Blur. As can be seen, where the warmer (closer to 1) regions activate the ReLU function and thus influence the decision of the network. It is remarkable that the first activation function reflect the presence of pixel deformation. The JPEG compression is well known artifact that causes undesirable blocks in the image due to the quantization. This issue is stated in ReLU 1 activation map of JPEG patch that shows the selection of these blocks as a highly important information to pass through the network. As well as for WN and Blur cyclopean patches, the ReLU 1 activation function have succeeded to focus on noise and blur artifacts. However, additionally, with the help of this activation function, we can form a distortion map. The latter can then be used by enhancement algorithms to concentrate on the most damaged regions instead of analyzing the while scene.

While the second activation function (ReLU layer) is controlled by a deeper representation that makes it harder to fully comprehend the outputs. However, for JPEG cyclopean patch, most deformed regions are placed above and by the edge of a pillar in the scene. Meanwhile for Blur, the deformed regions are located around everywhere the pillar. From the second ReLU output maps, the warmer regions are somewhat distributed according to the most infected regions in the scene. For further analysis, Fig. 6 provides visualisation of the extracted features from each patch. For comparison, three of the same scene cyclopean images of different distortion types and degrees were used. A quality score has been computed via the proposed scheme for each patch. As can be seen, the feature values are within range of 0 to 1 appear diversity as the degree of degradation varies. Note that the blue dots refer to features from non distorted stereoscopic image input. The distribution of blue dots are similar in all patches. The orange and red feature distributions refer to distorted stereoscopic image inputs. Here we notice non similar distribution at each patch because the approach tends to extract quality features relevant to the spatial information at each corner of the scene (as discussed earlier in section deep feature extraction). Consequently, each model derives distinct features and enrich the feature bank which is utilized to measure the quality. With regard to these observations, we can conclude that the trained networks focus on the pixel deformations to extract a complex quality indicators. The decision that defines these indicators is then guided by the type and degree of distortions.

Conclusion

In this paper, a new deep feature extraction approach has been explored for NR SIQA. The simplicity of proposed scheme is an advantage for implementation in the multimedia software. The proposed metric uses cyclopean image hypothesis that considers binocular rivalry phenomenon. Then, four CNN models are used to extract bank of features from the cyclopean image. This bank is then mapped to a quality score using a SVR. The obtained results have corroborated the correspondence between the proposed metric and the subjective DMOS over asymmetric and symmetric distributions. Based on the performance achieved, the followed workflow that combines multi-extractors with SVR could be useful for future works. The proposed method still has room for improvement. For that, in future we plan to optimize the quality feature sets to achieve even better accuracy.

Acknowledgments

The computation in this work was funded by TIC-ART (Traitement des Images au serviCe de l'ART) project, Centre-Val de Loire department, France.

References

[1] Jacky Baltes, Sancho McCann, and John Anderson, "Humanoid robots: Abarenbou and daodan," *RoboCup-Humanoid League Team Description*, 2006.

[2] Albert William and Darrell Bailey, "Stereoscopic visualization of scientific and medical content," 2006.

[3] Motion Picture Association of America, "2016 theatrical market statistics report," 2016.

[4] Oussama Messai, Fella Hachouf, and Zianou Ahmed Seghir, "Adaboost neural network and cyclopean view for no-reference stereoscopic image quality assessment," *Signal Processing: Image Communication*, p. 115772, 2020.

[5] O. Messai, F. Hachouf, and Z. A. Seghir, "Deep learning and cyclopean view for no-reference stereoscopic image quality assessment," in *2018 International Conference on Signal, Image, Vision and their Applications (SIVA)*, 2018, pp. 1–6.

[6] Aladine Chetouani, "Full reference image quality metric for stereo images based on cyclopean image computation and neural fusion," in *2014 IEEE Visual Communications and Image Processing Conference*. IEEE, 2014, pp. 109–112.

[7] Sid Ahmed Fezza, Aladine Chetouani, and Mohamed-Chaker Larabi, "Using distortion and asymmetry determination for blind stereoscopic image quality assessment strategy," *J. Visual Communication and Image Representation*, vol. 49, pp. 115–128, 2017.

[8] A. Chetouani, "A fusion-based blind image quality metric for blurred stereoscopic images," in *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 2017, pp. 1–5.

[9] Aladine Chetouani, "Toward a universal stereoscopic image quality metric without reference," in *Advanced Concepts for Intelligent Vision Systems*, Cham, 2015, pp. 604–612, Springer International Publishing.

[10] Aladine Chetouani, "A Neural-Based Stereoscopic Image Quality Assessment with Reference," in *Electronic Imaging*, San Francisco, United States, 2018.

[11] Pieter Seuntiens, "Visual experience of 3d tv," *doctor doctoral thesis, Eindhoven University of Technology*, 2006.

[12] A. Benoit, Patrick Le Callet, Patrizio Campisi, and Romain Cousseau, "Quality assessment of stereoscopic images," *EURASIP journal on image and video processing*, vol. 2008, no. 1, pp. 1–13, 2009.

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

[14] R. Bensalma and Mohamed-Chaker Larabi, "A perceptual metric for stereoscopic image quality assessment based on the binocular energy," *Multidimensional Systems and Signal Processing*, vol. 24, no. 2, pp. 281–316, 2013.

[15] P. Gorley and Nick Holliman, "Stereoscopic image quality metrics and compression," in *Electronic Imaging 2008*. International Society for Optics and Photonics, 2008, pp. 680305–680305.

[16] CTER Hewage, Stewart T Worrall, Safak Dogan, and AM Kondoz, "Prediction of stereoscopic video quality using objective quality models of 2-d video," *Electronics letters*, vol. 44, no. 16, pp. 963–965, 2008.

[17] David G Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Ieee, 1999, vol. 2, pp. 1150–1157.

[18] Martin A Fischler and Robert C Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," in *Readings in computer vision*, pp. 726–740. Elsevier, 1987.

[19] R. Blake, David H Westendorf, and Randall Overton, "What is suppressed during binocular rivalry?," *Perception*, vol. 9, no. 2, pp. 223–231, 1980.

[20] M. Chen, Che-Chun Su, Do-Kyoung Kwon, Lawrence K Cormack, and Alan C Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," *Signal Processing: Image Communication*, vol. 28, no. 9, pp. 1143–1155, 2013.

[21] WJM Levelt, "On binocular rivalry (p. 107)," *The Hague-Paris: Mouton*, 1968.

[22] Wujie Zhou, Weiwei Qiu, and Ming-Wei Wu, "Utilizing dictionary learning and machine learning for blind quality assessment of 3-d images," *IEEE Transactions on Broadcasting*, vol. 63, no. 2, pp. 404–415, 2017.

[23] Meixin Fang and Wujie Zhou, "Toward an unsupervised blind stereoscopic 3d image quality assessment using joint spatial and frequency representations," *AEU-International Journal of Electronics and Communications*, vol. 94, pp. 303–310, 2018.

[24] R. Akhter, ZM Parvez Sazzad, Yuukou Horita, and Jacky Baltes, "No-reference stereoscopic image quality assessment," in *IS SPIE Electronic Imaging*. International Society for Optics and Photonics, 2010, pp. 75240T–75240T.

[25] Maryam Karimi, Najmeh Soltanian, Shadrokh Samavi, Kayvan Najarian, Nader Karimi, and SM Reza Soroushmehr, "Blind stereo image quality assessment inspired by brain sensory-motor fusion," *Digital Signal Processing*, vol. 91, pp. 91–104, 2019.

[26] Yun Liu, Chang Tang, Zhi Zheng, and Liyuan Lin, "No-

reference stereoscopic image quality evaluator with segmented monocular features and perceptual binocular features," *Neurocomputing*, 2020.

[27] Lei Chen and Jiying Zhao, "No-reference perceptual quality assessment of stereoscopic images based on binocular visual characteristics," *Signal Processing: Image Communication*, vol. 76, pp. 1–10, 2019.

[28] Jinwoo Kim, Sewoong Ahn, Heeseok Oh, and Sanghoon Lee, "Cnn-based blind quality prediction on stereoscopic images via patch to image feature pooling," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1745–1749.

[29] D. J Field, "Relations between the statistics of natural images and the response properties of cortical cells," *JOSA A*, vol. 4, no. 12, pp. 2379–2394, 1987.

[30] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *CoRR*, vol. abs/1404.5997, 2014.

[31] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[33] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[34] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[35] A.K. Moorthy, Che-Chun Su, Anish Mittal, and Alan Conrad Bovik, "Subjective evaluation of stereoscopic image quality," *Signal Processing: Image Communication*, vol. 28, no. 8, pp. 870–883, 2013.