MC360IQA: THE MULTI-CHANNEL CNN FOR BLIND 360-DEGREE IMAGE QUALITY ASSESSMENT

Wei Sun[†], Ke Gu[‡], Weike Luo[†], Xiongkuo Min[†], Guangtao Zhai[†] , Siwei Ma[♯] and Xiaokang Yang[†]

[†]Institue of Image Commu. and Infor. Proce., Shanghai Jiao Tong University, China
[‡]Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[♯]School of Electronic Engineering and Computer Science, Peking University, China
Email: sunguwei@sjtu.edu.cn, guke.doctor@gmail.com,swma@pku.edu.cn
{lwk9419, minxiongkuo, zhaiguangtao, xkyang}@sjtu.edu.cn

ABSTRACT

In this paper, we present a multi-channel convolution neural network (CNN) for blind 360-degree image quality assessment (MC360IQA). To be consistent with the visual content of 360-degree images seen in the VR device, our model adopts the viewport images as the input. Specifically, we project each 360-degree image into six viewport images to cover omnidirectional visual content. By rotating the longitude of the front view, we can project one omnidirectional image onto lots of different groups of viewport images, which is an efficient way to avoid overfitting. MC360IQA consists of two parts, multi-channel CNN and image quality regressor. Multi-channel CNN includes six parallel ResNet34 networks, which are used to extract the features of the corresponding six viewport images. Image quality regressor fuses the features and regresses them to final scores. The results show that our model achieves the best performance among the state-of-art full-reference (FR) and no-reference (NR) image quality assessment (IQA) models on the available 360-degree IQA database.

Index Terms— Blind Image Quality Assessment, 360-degree image, Virtual Reality, multi-channel CNN

1. INTRODUCTION

360-degree images/videos, also known as panoramic, omnidirectional or VR images/videos, have been accessed by more people with the rapid development of Virtual Reality (VR) technology. As a new type of multimedia, 360-degree images/videos record views in every direction at the same time. By rotating the head orientation, users can see the content of images/videos from any directions through VR devices. The immersive experience of real-world scenes makes the 360-degree images/video popular in social media, live concert events or sports events, and VR movies.

Due to the omnidirectional view recording, 360-degree images/videos often have high resolution and are often compressed heavily for easy transmission and storage. However, 360-degree images/videos at low resolution or with serious compression artifacts usually make people feel uncomfortable, sometimes even produce motion sickness, which dramatically degrades the quality of experience (QoE) [1]. Therefore, it is crucial to study the quality assessment of 360-degree images, especially for compressed 360-degree images, which has significant implications in leading the development of 360-degree image compression.

Image quality assessment (IQA) has been thoroughly studied in the past twenty years. However, as far as we know, limited work has been done on the quality assessment of 360-degree images. IQA algorithms can be generally classified into full-reference IQA (FR IQA), reduced-reference IQA (RR IQA) and no-reference IQA (NR IQA). FR IQA and RR IQA models need full and part reference image information respectively while NR IQA takes only the distortion image as input.

For FR IQA, several PSNR-based IQA models have been proposed for 360-degree images. These models mainly consider the geometric distortion occurring in the projection. 360-degree images are usually mapped to the rectangular plane for easy storage and visualization. The equirectangular projection is the simplest and most widely used projection for 360-degree images. Yu et al. [2] proposed a sphere based PSNR (S-PSNR), which computes PSNR for the set of points uniformly distributed on a spherical surface instead of on the rectangular domain. Sun et al. [3] proposed the Weighted Spherical PSNR (WS-PSNR), of which the weight is determined by how much the sampled area is stretched in the representation. Zakharchenko et al. [4] proposed Craster Parabolic Projection PSNR (CPP-PSNR). They remapped both the distorted and reference images to the Craster parabolic projection and computed the PSNR in that domain. However, due to the inconsistency between PSNR and the experience of the human vision system (HSV), the performance of these models is significantly inferior to the traditional successful IQA models for 2D natural images according to the studies of [5] [6] [7]. For NR IQA, as far as we know, there is no NR

978-1-7281-0397-6/19/$31.00 ©2019 IEEE

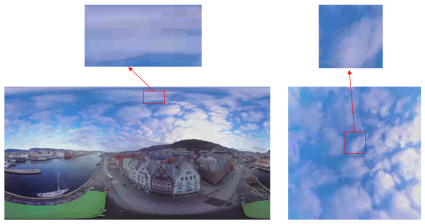

Fig. 1: Distortion comparison between the viewport image seen in VR devices and its corresponding omnidirectional image in the equirectangular format.

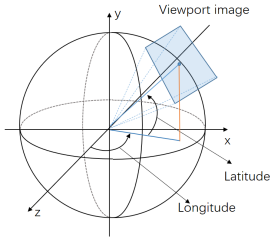

Fig. 2: Illustration for viewport images when the user sees the omnidirectional image in VR devices at a certain head pose.

IQA model specially designed for 360-degree images. The study in [5] also revealed that the current general purposed NR IQA models perform poorly on omnidirectional images. So, it is essential to develop NR IQA models for 360-degree images.

In this paper, we propose a multi-channel convolution neural network (CNN) model for NR 360-degree image quality assessment (MC360IQA). Different from the methods mentioned above, we use the viewport-based images projected by equirectangular images instead of equirectangular images. We argue that equirectangular images suffer great structure distortion which seriously affects human's perception of omnidirectional image quality. Figure 1 shows the distortion comparison between the viewport image seen in the VR device and its corresponding omnidirectional image in the equirectangular format. It shows that the compression artifacts in the equirectangular projection are block-based while the distortion seen in the viewport is totally different. This illustrates why many NR IQA algorithms work poorly in this field. Therefore, we use the viewport-based images as the input of MC360IQA. In the pre-processing stage, we project the equirectangular image into six equally-sized viewport images, represented each cube face with a field of view of 90 degree. We also alter the longitude of view angle to project many different groups of viewport images from one omnidirectional to avoid overfitting. MC360IQA model consists of two part. The first part includes six parallel CNN channels used to extract features of the six viewport images. The second part is the image quality regressor, which concatenates the features of the six viewport images and regresses them to the final quality scores. The experimental results show that

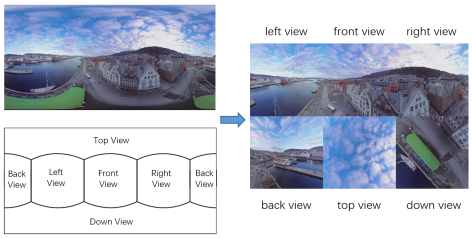

Fig. 3: The viewport images and their corresponding part in the omnidirectional image

the proposed model achieves the best performance among the state-of-art NR and FR IQA models.

The rest of this paper is organized as follows. In section 2, we describe the implementation of the MC360IQA in detail. In section 3, we give the results of MC360IQA and compare the performance of MC360IQA with other popular IQA models on the available 360-degree IQA database. Section 4 gives the concluding remarks.

2. PROPOSED METHOD

In this section, we detail the pipeline of MC360IQA for evaluating the 360-degree image quality. The model takes a 360-degree image as input and firstly projects it into six viewport images using the method described in Section 2.1. Then six viewport images are sent to the multi-channel CNN, the details of which we depict in Section 2.2. The features extracted by multi-channel CNN are fused and finally regressed to the objective quality score.

2.1. Our projection method

When users see the visual content of the 360-degree image in the VR device, the equirectangular image is first represented by a sphere in 3D spherical coordinates and then the visual content is rendered as a plane segment tangential to the sphere decided by the view angle and the FoV of the VR device. We show this process in Fig. 2. Users can view all the contents of the 360-degree image by rotating the head to change the viewing angle. When assessing the quality of a 360-degree image, the viewer should look around the 360-degree image from several view angles to cover the entire 360-degree image.

Inspired by this behavior, we propose using the viewport-based images to evaluate omnidirectional image quality. The pixel in the viewport image can be calculated by mapping it backwards to find the best estimate pixel in the spherical image. The detailed procedure can be found in [2]. We set the field of view (FoV) as 90 degree, which is consistent with the FoV of most popular VR devices such as HTC VIVE, Oculus, Gear VR, etc. To cover the full visual content of the omnidirectional image, six viewport images are rendered by one omnidirectional image. Two of these views are oriented towards the nadir and zenith, and

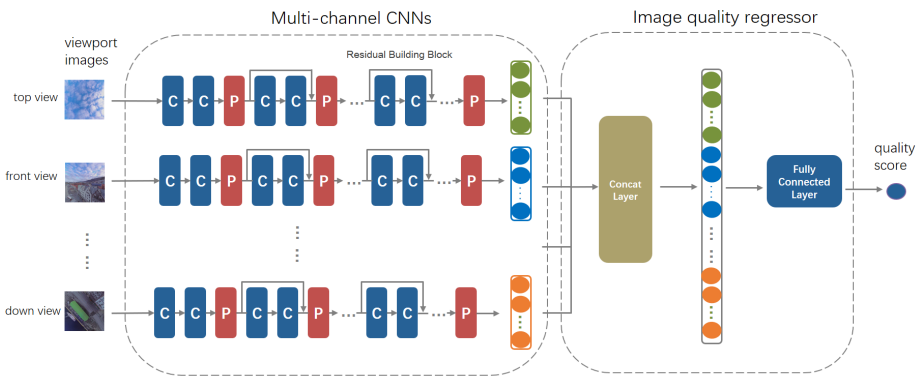

Fig. 4: The network architecture of MC360IQA. The multi-channel CNN includes six parallel ResNet34. We omit the three of them which are sent left, right and back view images for saving the layout.

the other four are pointed towards the horizon but rotated horizontally to cover the entire band at the sphere's equator, which is shown in Fig. 3. We use the symbols VP_{front},VP_{back},VP_{right},VP_{left},VP_{top},VP_{down} to represent the six viewport images in the front, back, right, left, top and down view, respectively. To avoid overfitting, we rotate the longitude of the view angle of the front view from 0 to 360 degree with an interval of φ degree and then project omnidirectional images to six viewport images at each front view angle respectively. Finally, we can get N groups of viewport images derived from one omnidirectional image. We denote them as VP^i_{view}, where $view \in \{front, back, right, left, top, down\}$ and $i \in [1, 2, ..., N]$, $N = 360/\varphi$.

2.2. MC360IQA

Convolution Neural Networks have shown great performance in solving visual signal problems in recent years. Many successful CNN models such as VGG [8], GoogleNet [9], ResNet [10] have been proposed for solving image recognition, detection, segmentation problems, etc. These models usually have strong ability in extracting high-level semantic features. We adopt ResNet as the base CNN-channel since Resnet has an excellent generalization ability in lots of visual tasks and has a relatively small memory consumption. We detail the architecture of MC360IQA as follows.

The MC360IQA model consists of two parts, multi-channel CNN and image quality regressor. We illustrate the framework in Fig. 4. The multi-channel CNN includes six parallel ResNet34s which are used to extract features of corresponding six viewport images. ResNet utilizes residual learning to further deepen the CNN network, which can be generally represented by several deeper bottleneck architectures. Each bottleneck includes three layers convolutions where the dimension of kernels is 1×1, 3×3, 1×1, respectively. The identity shortcut connection is inserted from the input of bottleneck to the output of the bottleneck. The complete network structure can be found in [10]. The six ResNet34 channels share the same weights and are trained to extract the unified features for different compression artifacts. We replace the last layer of each baseline ResNet34 with 10 output features by average pooling. The image quality regressor first fuses the features by concatenating the outputs of multi-channel CNN. According to [11], users focus more on the equator area and seldom view the nadir and zenith area, which indicates the importance of each viewport image is different for the final quality score. Therefore, another function of the image quality regressor is to assign weights for different viewport images. Finally, the quality score can be calculated by using a full connected layer in the image quality regressor.

For the end-to-end training, the loss function is set as:

$$L = (q_{predict} - q_{label})^2 \quad (1)$$

where $q_{predict}$ is the predicted score calculated by MC360IQA and q_{label} is the mean opinion score (MOS) derived from subjective experiments.

3. EXPERIMENTS AND RESULTS

3.1. Dataset

Compressed VR Image Quality Database (CVIQD2018) [5] is the only available compressed 360-degree image quality database so far. CVIQD2018 consists of 16 source 360-degree images and 528 compressed ones from three codecs. Each source image was compressed with quality factors ranging from 50 to 0 with an interval of -5 by JPEG codec and was compressed with factors from 30 to 50 with an interval of 2 by H.264/AVC and H.265/HEVC codecs. The database includes diverse scenes such as landscapes, towns, objects, and persons, etc. All the images have the same resolution of 4096×2048. The Single-Stimulus (SS) was adopted to gather quality ratings. The quality ratings lie in the range of [1,10], where a higher score indicates better visual quality.

3.2. Experiment setup

The proposed MC360IQA is implemented on pytorch [12]. We use the Resent34 as the baseline CNN model. The base-

Table 1: Performance comparison between 11 state-of-art FR and NR IQA models and two proposed metrics. We highlight the three best performing models in each column.

	Metrics	SRCC	PLCC	RMSE
FR	PSNR	0.7320	0.7662	9.0397
	S-PSNR	0.7574	0.7819	8.7695
	WS-PSNR	0.7467	0.7741	8.9066
	CPP-PSNR	0.7498	0.7755	8.8816
	SSIM	**0.8857**	**0.8972**	**6.214**
	MS-SSIM	0.8762	0.8875	6.4836
NR	QAC	0.8299	0.8681	6.9820
	GMLF	-0.2246	0.6134	11.1101
	NIQE	-0.5126	0.5329	11.9038
	BRISQUE	-0.7448	0.7641	9.0751
	SISBLIM	-0.6554	0.7439	9.4014
	$MC360IQA_{origin}$	**0.9069**	**0.9271**	**6.3924**
	$MC360IQA_{mean}$	**0.9153**	**0.9391**	**5.6728**

line CNN weights are initialized by training on ImageNet [13] and the weights of fully connected layers are randomly initialized. The interval angle φ was set as 2 degree, which means 180 groups of six viewport images can be rendered from one omnidirectional image. We trained and tested our model on a server with Intel Xeon Silver 4114 CPU @ 2.20GHz, 64 GB RAM and NVIDIA GTX 1080Ti. The batch size was set as 50. We chose the RMSprop algorithm [14] for speeding up mini-batch learning. The learning rate and alpha were set as 0.0001 and 0.9, respectively. We stop the training at 20 iterations. For the fair evaluation, we used 5-fold cross validation.

3.3. Results

To calculate the performance of the proposed model, three statistical indices are applied for consistency performance comparison with predicted scores obtained from the proposed model and subjective MOSs, including Spearman rank correction coefficient (SRCC), Pearsons liner Correlation Coefficient (PLCC) and Root mean square error (RMSE).

We compare our model with 6 popular FR IQA models, which are PSNR, WS-PSNR [3], CPP-PSNR [4], S-PSNR [2], SSIM [15], MS-SSIM [16]. Among these, WS-PSNR, CPP-PSNR and S-PSNR are IQA models specially designed for 360-degree images. Five popular general-purposed NR IQA models are compared with our model, which are BRISQUE [17], GMLF [18], NIQE [19], QAC [20] and SISBLIM [21]. For the MC360IQA, two metrics are proposed to measure the quality of 360-degree images. The first metric uses the score calculated by MTC360IQA using the viewport images without longitude rotating, denoted by $MC360IQA_{origin}$. The second metric uses the mean score of N groups of viewport images calculated by MC360IQA, denoted by $MC360IQA_{mean}$. We list the performance of these models in Table 1. The best three performance are highlighted in each column in Table 1.

$$
\begin{aligned}
MC360IQA_{origin} &= MC360IQA(VP^{1}_{view}) \\
MC360IQA_{mean} &= \sum_{i=1}^{N} MC360IQA(VP^{i}_{view})
\end{aligned}
\tag{2}
$$

where $view \in \{front, back, right, left, top, down\}$ and $i \in [1, 2, ..., N]$, $N = 360/\varphi$.

From the performance listed on Table 1, we have several observations. First, our proposed model achieves the best performance among all the state-of-art FR and NR IQA models on the database, which demonstrates the effectiveness of the proposed model. Second, most general purposed NR IQA models perform poorly. This indicates that the existing NR IQA models are not suitable for 360-degree image quality evaluation. Fortunately, the proposed model makes up for this gap. Third, we observed that the current PSNR-based IQAs which are specially designed for 360-degree images also perform poorly and do not improve much performance compared to PSNR. Therefore, FR IQA for 360-degree images also needs to be studied. In the future study, we hope to extend our model to the FR IQA and improve the performance of the model further. Forth, the metric $MC360IQA_{mean}$ is slightly better than the metric $MC360IQA_{origin}$. The reason is that the mean score of the N groups of viewport images is more stable and less susceptible to abnormal predictive scores. But $MC360IQA_{mean}$ needs to consume N times of computing resources than that of $MC360IQA_{origin}$. In comparison, $MC360IQA_{origin}$ is more efficient.

4. CONCLUSION

In this paper, we propose the first blind IQA model for 360-degree image, named MC360IQA. The MC360IQA uses the multi-channel CNN architecture to extract the features of viewport images projected by omnidirectional images and then regresses features to objective scores. According to using different groups of viewport images, we propose the two IQA metrics. The first one uses the scores calculated by viewport images without data augmentation and the second one uses the mean scores of lots of groups of viewport images with data augmentation via altering the longitude of view direction. The experimental results show that the proposed two metrics achieve the best performance among the state-of-art NR and FR IQA models.

5. ACKNOWLEDGEMENT

This work was supported by the National Science Foundation of China (661831015,61521062,61527804) and Equipment Pre-research Joint Research Program of Ministry of Education 6141A020223.

6. REFERENCES

[1] JJ-W Lin, Henry Been-Lirn Duh, Donald E Parker, Habib Abi-Rached, and Thomas A Furness, "Effects of field of view on presence, enjoyment, memory, and simulator sickness in a virtual environment," in *Virtual Reality, 2002. Proceedings. IEEE*. IEEE, 2002, pp. 164–171.

[2] Matt Yu, Haricharan Lakshman, and Bernd Girod, "A framework to evaluate omnidirectional video coding schemes," in *Mixed and Augmented Reality (ISMAR), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 31–36.

[3] S Yule, A Lu, and Y Lu, "Ws-psnr for 360 video objective quality evaluation," *MPEG Joint Video Exploration Team*, vol. 116, 2016.

[4] Vladyslav Zakharchenko, Kwang Pyo Choi, and Jeong Hoon Park, "Quality metric for spherical panoramic video," in *Optics and Photonics for Information Processing X*. International Society for Optics and Photonics, 2016, vol. 9970, p. 99700C.

[5] Wei Sun, Ke Gu, Siwei Ma, Wenhan Zhu, Ning Liu, and Guangtao Zhai, "A large-scale compressed 360-degree spherical image database: From subjective quality evaluation to objective model comparison," in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2018, pp. 1–6.

[6] Evgeniy Upenik, Martin Rerabek, and Touradj Ebrahimi, "On the performance of objective metrics for omnidirectional visual content," in *Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on*. IEEE, 2017, pp. 1–6.

[7] Wei Sun, Ke Gu, Guangtao Zhai, Siwei Ma, Weisi Lin, and Patrick Le Callet, "Cviqd: Subjective quality evaluation of compressed virtual reality images," *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3450–3454, 2017.

[8] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[9] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[11] Yucheng Zhu, Guangtao Zhai, and Xiongkuo Min, "The prediction of head and eye movement for 360 degree images," *Signal Processing: Image Communication*, 2018.

[12] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," 2017.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[14] Alex Graves, "Generating sequences with recurrent neural networks," *CoRR*, vol. abs/1308.0850, 2013.

[15] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

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

[17] Anish Mittal, Anush K. Moorthy, and Alan C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, pp. 4695–4708, 2012.

[18] Wufeng Xue, Xuanqin Mou, Lei Zhang, Alan C. Bovik, and Xiangchu Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Transactions on Image Processing*, vol. 23, pp. 4850–4862, 2014.

[19] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, pp. 209–212, 2013.

[20] Wufeng Xue, Lei Zhang, and Xuanqin Mou, "Learning without human scores for blind image quality assessment," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 995–1002, 2013.

[21] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang, "Hybrid no-reference quality metric for singly and multiply distorted images," *IEEE Transactions on Broadcasting*, vol. 60, pp. 555–567, 2014.