Visual Scan-Path based Data-Augmentation for CNN-based 360-degree Image Quality Assessment

Abderrezzaq Sendjasni, Mohamed-Chaker Larabi, Faouzi Alaya Cheikh

Visual Scan-Path based Data-Augmentation for CNN-based 360-degree Image Quality Assessment

Abderrezzaq Sendjasni [1,2]***, Mohamed-Chaker Larabi*** [1] ***and Faouzi Alaya Cheikh*** [2]
[1] ***CNRS, Xlim UMR 7252, Université de Poitiers, France***
[2] ***NTNU, Norwegian Colour and Visual Computing Lab, Gjøvik, Norway***

Abstract

360-degree Image quality assessment (IQA) is facing the major challenge of lack of ground-truth databases. This problem is accentuated for deep learning based approaches where the performances are as good as the available data. In this context, only two databases are used to train and validate deep learning-based IQA models. To compensate this lack, a data-augmentation technique is investigated in this paper. We use visual scan-path to increase the learning examples from existing training data. Multiple scan-paths are predicted to account for the diversity of human observers. These scan-paths are then used to select viewports from the spherical representation. The results of the data-augmentation training scheme showed an improvement over not using it. We also try to answer the question of using the MOS obtained for the 360-degree image as the quality anchor for the whole set of extracted viewports in comparison to 2D blind quality metrics. The comparison showed the superiority of using the MOS when adopting a patch-based learning.

Introduction

The assessment of image quality (IQA) is an important topic in image processing and computer vision. It is the process of measuring the weighted combination of visual attributes that represent the perceptual quality of a given image. This is performed in a way that should be consistent with human subjective opinions [1] (ground truth obtained by the mean of psycho-visual experiments). Depending on the existence of the pristine images, IQA can be considered as full reference (FR) if the pristine image is available or no reference (NR) if not. It can also be as reduced reference (RR) if only partial information of the source image is used. The NR methods are widely adopted as it reflects real world scenarios where the original images are most likely unavailable. At present, IQA has been widely studied for 2D content [2, 3]. However, for omnidirectional scenes (*a.k.a.* 360-degree), it is still in its infancy and not fully investigated.

360-degree images represent an important part of virtual reality (VR) content, in which the users are provided with real world scenes to live an immersive experience. With commercial head mount displays (HMDs), the viewer is allowed to freely focus on the desired content thanks to his head movements (HM) making the interactive and the immersive experience more interesting. Accordingly, to achieve good quality of experience (QoE), immersive contents with high visual quality should be provided. Based on this, quality assessment of 360-degree images becomes crucial to control QoE, and therefore, adequate IQA tools are of major importance.

As 360-degree content is generally processed, encoded, and transmitted using a 2D plane representation, a straightforward solution is to use the large literature on 2D quality metrics directly on the 2D representation. Still, these metrics do not account for the non-uniform sampling density at pixel locations from the sphere to plane projection [4]. And, their performances are lacking in terms of correlation with subjective quality scores as shown in [5]. Furthermore, the most projection format used is the equirectangular (ERP) one. This projection suffer from geometric distortion due to the projection and therefore, do not represent the viewed content by the users. Thus, having metrics dedicated to 360-degree images accounting for its characteristics (spherical- or projections-based) becomes of major importance in order to meet the challenges related to this type of content.

With the introduction of 360-degree images, a few IQA models have been proposed by extending traditional 2D models such as PSNR or MSE. For example, PSNR-based methods like Spherical PSNR (S-PSNR) [4] which computes the PSNR on a spherical surface instead of the 2D representation. The weighted spherical PSNR (WS-PSNR) [6] uses the scaling factor from a 2D plane to the sphere as a weighting factor for PSNR computation. CPP-PSNR [7] computes PSNR on the craster parabolic projection (CPP) after re-mapping pixels of the original and distorted images from the spherical domain to CPP. As these models do not account for perceptual aspects, they fail in predicting the visual quality accurately. Among the possible ways to reach reliable and accurate solutions for 360-degree IQA, is the use of deep-learning techniques.

The use of convolutional neural networks (CNNs) for quality assessment tasks is fastly growing. This is mainly due to their architecture, capable to extract discriminating features at various levels of abstraction [8, 9], *i.e.* the ability to learn multi-level features. CNNs are involved in various image processing tasks, such as image segmentation, object detection and, image classification. The inherited models are often exploited to regress the quality scores by means of transfer learning and/or by learning human visual system (HVS) based features [10, 11]. Several models have shown greater performances in other tasks as they have been trained on large databases such as ImageNet [12]. Seeking their efficiency, these models are used as backbones in several IQA models [13–16].

CNN-based models dedicated to 360-degree IQA are rather few. For instance, a pre-trained model (MC360IQA) is used in [13] to predict the quality on viewports extracted from the cube-map projection (CMP) of the 360-degree image. Six viewports representing the 6 faces of the CMP are extracted and used as inputs of a pre-trained ResNet-34 [9] forming a multi ResNet-34 (*i.e* a multi-channel paradigm). The outputs of all channels are weighted and concatenated to predict the quality score. In addition, the authors used a data augmentation techniques by rotating the longitude of the viewing angle of the front view from 0 to 360 degree with a 2 degree interval then project to CMP at each front viewing angle. This method results in a lot of redundant content. In [15], a viewports-based approach is also proposed. Here, the authors take benefit from the spatial mutual dependencies among the extracted viewports by using a graph CNN. In

addition, they use the DB-CNN proposed in [17] to compute the global quality with a down-sampled ERP image as an input. Both outputs are combined to predict the quality score. A deep learning framework is proposed in [14] where the quality scores are predicted on weighted patches from the equirectangularly projected (ERP) image. Here, the ResNet-50 [9] model is used to predict the quality score of each patch. However, the ERP suffers from geometric distortions and do not represent the viewed content. In response to these limitations, a CNN based model, proposed in [16], predicts visual quality based on the spherical content of selected viewports rather than projected content. Viewports are selected using visual scan-path predictions. Furthermore, the just-noticeable difference map is used to account for perceptual characteristics of the HVS along with features produced from scan-paths in order to estimate the weights of each viewport. Differently from the previous mentioned works, in [18] a patch-based approach is adopted rather than a mutli-channel one. The patches are 64×64 and extracted from the ERP with a focus on the equatorial region. Each patch inherits the mean opinion score (MOS) of its 360-degree image, which may be questionable as it is too small to represent a 2K+ scene.

Inspired by the previously mentioned works, and motivated by the lack of databases for 360-degree IQA, we propose a data-augmentation technique based on visual scan-paths for a CNN-based IQA model. First, we extract viewports on the spherical content of 360-degree images according to visual scan-path predictions rather than a projected format. This way, we reproduce the actual viewed content and avoid distortions due to the projection process. Then, as we use a patch-based learning scheme, each viewport is labeled with either the MOS of its image or a quality score obtained using NR 2D metric. We compared the efficiency of using the MOS with the use of local quality scores obtained by 2D models. To do so, two widely used NR metrics, BRISQUE [19] and NIQE [20] are used. Finally, a weighted pooling based on local quality is then proposed to compute the final quality scores.

The proposed method

Scan-path based data-augmentation

Fig. 1 depicts a 360-degree image viewing experience. We only consider chosen viewports to predict the quality as it represents the way 360-degree images are generally viewed. The assumption that a person can only see the actual rendered field of view (FoV) from the spherical representation justifies this procedure. The next viewport is determined by his head movement around the x, y, and z axes. This way, the quality prediction scenario seeks to agree with the viewing experience of 360-degree images, and geometric distortions created by the previously described sphere to plane projection are avoided.

It is now widely admitted that when an image is viewed, the HVS gazes on salient details, which translates into eye fixations [21]. In our case, these regions are considered as relevant viewports and are detected using visual scan-path predictions model proposed in [22]. With this model, a visual trajectory including eight pertinent fixation points are predicted. In our model, ten trajectories are extracted representing ten virtual observers and used to account for the diversity of human scan-paths. These scan-paths are then considered for data augmentation, which results in a total of $N = 8 \times 10$ extracted viewports. This will help with the training of the model and avoid overfitting caused by the lack of data. In fact, the efficiency of deep neural networks often increases as more data is available. Unfortunately, we still lack reliable and representative databases for 360-degree IQA that would allow deep learning models to assert their full capabilities. The construction of such databases require important efforts in terms of scenes acquisition, device calibration, paradigm definition, subjective testing and data analysis [5]. Only two 360-degree image databases are been used to train and validate IQA models, namely CVIQD [23] and OIQA [24]. Consequently, the application of strategies to acquire more data with the existing one, is largely encouraged. The use of IQA-based data-augmentation is one option to accomplish this task.

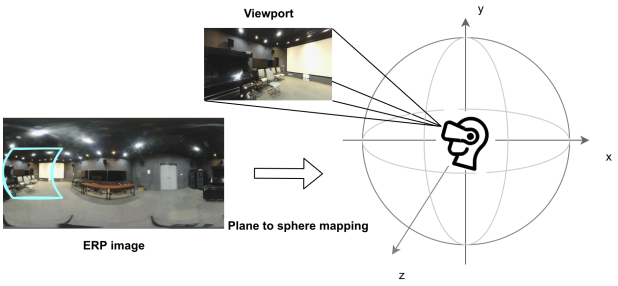

Figure 1: 360-degree images viewed using head-mounted devices. Blue area in the ERP represents the viewport extracted from the sphere.

Data augmentation is a method of creating new training data from existing one. This is accomplished by applying domain-specific (in our case IQA) strategies to elements from the training data in order to generate new and distinct training examples [25]. Since IQA is more sensitive than other image processing tasks such as object detection and classification, conventional approaches including shifting, rotating, flipping and brightness changing of an image, are counterproductive in our context. The particular reason for this, being that the images are labelled (rated) by human observers (MOS), and altering any visual attribute will make the actual rating incompatible. As a result, the use of data-augmentation techniques must be appropriate and concur with IQA. In our model, we adopted visual scan-paths to augment the training data as explained above. The motivation behind such an approach is that each virtual observer (VO) will explore the same scene but will probably provide a different rating as in real subjective experiments. So, from each image in the database, we extract eight viewports for each VO where fixation points are taken as the center of the viewports. This way, we generate ten different instances of the database. During the end-to-end training, each extracted viewport is taken as an individual input to the model.

Architecture of the Model

Fig.2 depicts the architecture of the proposed method. To begin, the scan-path model is used to predict the ten VO potential trajectories and their gaze fixation positions. Then, rather than the projected format, each fixation point is located on the sphere, and the surrounding content is extracted and projected to a 2D plane. Following that, since we are using a patch-based learning scheme, each extracted region is fed to the model as a separate input. Previously mentioned, 360-degree images are rated based on multiply viewed regions. Giving the extracted viewports the same MOS as their 360-degree image as firstly introduce in [26] for 2D content seems inefficient. Therefore, we applied well known and widely used 2D NR quality metrics to predict the quality of extracted viewports, referred to in the following as the local quality. This is motivated by the fact that the extracted viewports have a 2D representation and the model will consider them as separate images.

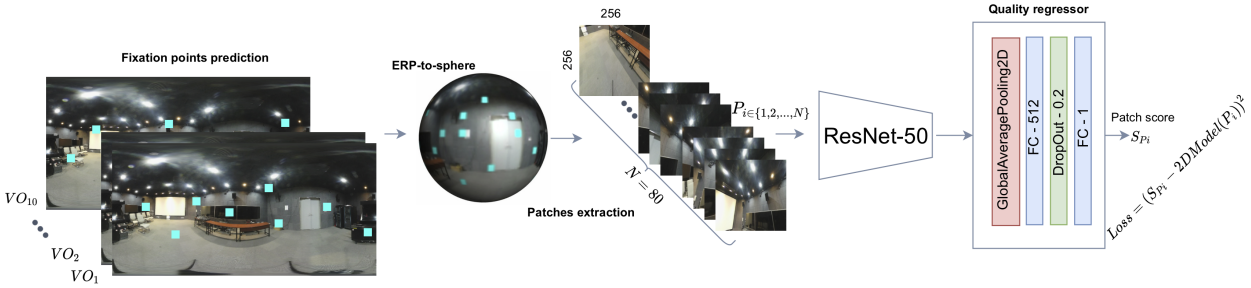

Figure 2: Architecture of the proposed model. Features are only extracted from individual viewports by ResNet-50

In this paper, ResNet-50 [9] is used as the base model to extract visual features from selected viewports. We replaced the top layers with a regression block in order to regress the learned features into a single quality score. ResNet employs residual learning to further deepen the CNN network, which can be interpreted by a number of deeper bottleneck architectures. Each bottleneck has three convolutional layers with kernel dimensions of 1×1, 3×3 and 1×1 respectively. A shortcut connection is then added from the input of the bottleneck to its output. Furthermore, since the shortcuts accelerate deep network convergence, the ResNet has the ability to avoid the problems of vanishing gradient [9]. Several versions of this model were developed based on the number of layers including ResNet-18, ResNet-34, ResNet-50 and ResNet-101. We chose the ResNet-50 with its pre-trained weights obtained on the ImageNet [12] database as it is the most common and widely used. Furthermore, this choice is also motivated based on conclusions of a previous comparative study [27] for which it ranked the best compared to VGG-16/19, ResNet-18/34, and Inception-V3 models.

The output of ResNet-50 is fed to a quality regressor which is composed of a global average pooling so to reduce the spatial dimensions of the extracted feature maps and help to minimize overfilling. Finally, two fully connected (FC) layers are then used to calculate the quality score. The weights for the quality regressor are initialized according to the method provided in [28].

For the end-to-end training, we used the L_2 loss function to compute the error between predicted and target scores. For this latter, we used MOS and the local quality scores obtained by NIQE and BRISQUE.

To compute the quality score of the entire 360-degree images, an average pooling is then calculated. For each VO, eights scores are averaged to a single score for each image. For the data-augmentation based configuration, the final score is obtained by averaging the score of eighty extracted regions.

Reduced Content Biases Splitting

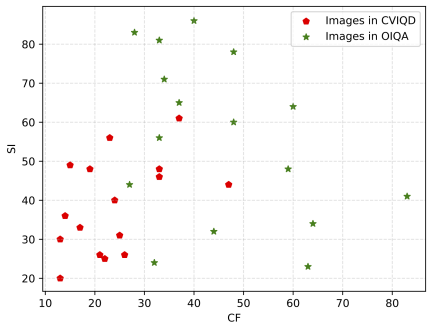

Figure 3: Spatial information (SI)/colorfulness information (CFI) plot of the source images in both CVIQD [23] and OIQA [24] databases.

For deep learning models performance analysis, the results are reported only on the testing set in which the selection may induce biases related to the content. A popular and straightforward approach is to divide the training and testing sets by reference images. This means that the model is evaluated on unseen content independent of database distortions. However, the obtained sets may lack diversity in terms of various visual aspects and may induce representativeness biases, resulting in a test set that is not well representative of the used database. Biases are mostly present, whether the data is divided arbitrarily or based on more qualified criterion. However, minimizing those biases guarantees a validation on representative sets of the trained model.

To minimize content induced biases, we use spatial information (SI) and colorfulness information (CFI) as criteria for the splitting scheme to make sure that, the performance of the models are reported on limited biased set of images. The SI accounts for spatial complexity and CFI accounts for the variety and intensity of colors in the images. SI and CFI are computed according to the ITU-T P.910 [29] recommendations and the metric described in [30] respectively. Fig. 3 shows the SI/CFI plots of the source images of CVIQD and OIQA databases. As it can be seen, the variability of SI is higher in OIQA than in CVIQD, indicating that CVIQD lacks diversity of content in terms of spatial complexity in comparison to OIQA. A similar conclusion holds in the case of CFI. To select the training/testing sets, we used the Euclidean distance between each two pristine images as described in Eq. 1. So, for a couple of pristine images I_1 and I_2 characterized by (CFI_{I_1}, SI_{I_1}) and (CFI_{I_2}, SI_{I_2}) respectively, the distance $D(I_1, I_2)$ is expressed as follows:

$$D(I_1, I_2) = \sqrt{(CFI_{I_1} - CFI_{I_2})^2 + (SI_{I_1} - SI_{I_2})^2} \quad (1)$$

Experimental results

This study is carried out using the CVIQD [23] and OIQA [24] databases containing ERP images. CVIQD includes 16 pristine 360-degree images and 528 impaired ones. The distortion used to create this database are only compression related, *i.e.* are JPEG compression with quality factors ranging from 50 to 0 in addition to H.264/AVC and H.265/HEVC with quality parameter from 30 to 50. The authors used the single stimulus with a rating scale of 10 levels from the lowest to the highest quality to gather the MOS. 20 subjects participated in the creation of this database. OIQA includes 320 distorted 360-degree images created from 16 pristine ones, using 4 distortion types including JPEG compression (JPEG), JPEG2000 compression (JP2K), Gaussian blur (BLUR) and Gaussian white noise (WN). Subjective scores, obtained from 20 subjects, range from 1 (bad) to 10 (excellent).

The databases are split using the well known Pareto principle and the criterion discussed previously, 80% for training and 20% for testing. For the sake of a fair comparison, all config-

Table 1: Performance evaluation of the model. The Best performance is highlighted in **bold**. The mean of 5 folds is provided

	MOS				NIQE				BRISQUE			
	CVIQD		OIQA		CVIQD		OIQA		CVIQD		OIQA	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
VO_1	0.829	0.773	0.898	0.884	0.756	**0.791**	0.445	0.382	0.771	0.673	0.732	0.714
VO_2	0.815	0.753	0.877	0.860	0.791	0.685	0.426	0.398	0.759	0.707	0.747	0.713
VO_3	0.836	0.762	0.907	0.892	0.793	0.677	0.452	0.406	0.743	0.693	0.751	0.731
VO_4	0.835	0.759	0.911	0.895	0.792	0.686	0.419	0.372	0.772	0.716	0.690	0.661
VO_5	0.830	0.765	0.879	0.868	0.752	0.620	0.432	0.415	0.673	0.626	0.779	0.740
VO_6	0.820	0.748	0.916	0.898	0.781	0.653	0.498	0.457	0.792	0.723	0.710	0.662
VO_7	0.838	0.759	0.888	0.872	0.738	0.605	0.450	0.423	0.729	0.656	0.777	0.749
VO_8	0.845	0.783	0.898	0.880	0.801	0.700	0.462	0.399	0.768	0.711	0.767	0.736
VO_9	0.817	0.760	0.902	0.884	0.743	0.616	0.446	0.398	0.735	0.683	0.758	0.712
VO_{10}	0.835	0.754	0.893	0.872	0.722	0.591	0.479	0.412	0.722	0.662	0.743	0.714
Avg	0.830	0.762	0.897	0.881	0.767	0.662	0.451	0.406	0.746	0.685	0.745	0.713
STD	0.010	0.010	0.013	0.013	0.028	0.059	0.024	0.023	0.034	0.031	0.028	0.030
All	**0.871**	**0.801**	**0.920**	**0.904**	**0.827**	0.704	**0.519**	**0.478**	**0.799**	**0.738**	**0.811**	**0.788**

urations were trained/tested using the same splitting scheme. A five-fold cross validation is used for a complete evaluation within the selected databases.

The model is implemented using TensorFlow [31] and trained on a server with Intel Xeon Silver 4208 2.1GHz, 192G RAM and a GPU Nvidia Telsa V100S 32G. The batch size was set to 32 and the Adam optimizer [32] is used with a learning rate of $1e-4$, first parameter $\beta_1 = 0.9$ and second parameter $\beta_2 = 0.999$. We used the early stopping to stop the training once no improvement is observed.

Performance evaluation

To assess the performance of the proposed data-augmentation technique, we used the Pearson linear correlation coefficient (PLCC) and the Spearman rank order correlation coefficient (SRCC).

Table 1 summarizes the performance of individual VO-based training in terms of accuracy of prediction (PLCC) and monotonicity (SRCC), as well as the application of combined VOs on both databases. The latter refers to the data-augmentation based training. Regarding the performances of individual VOs, we can observe that the range of performance is not significantly different. It is confirmed by the standard deviation given in the table. This actively demonstrates that, the various predicted scan-paths are almost of similar importance, and none of them can be considered as non-valid or outlying.

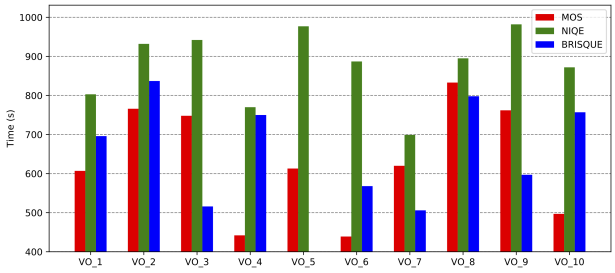

Figure 4: Computational time for VOs individually on CVIQD.

Between the MOS, NIQE and BRISQUE as local quality for the model training, the MOS based one outperformed the others. In terms of difference, the range of correlation for the MOS is the smallest among the studied cases. The obtained results contradict our expectations. Assigning the same MOS value to small regions from the same 360-degree image looks at a first sight as not appropriate. Applying 2D models are adopted to account

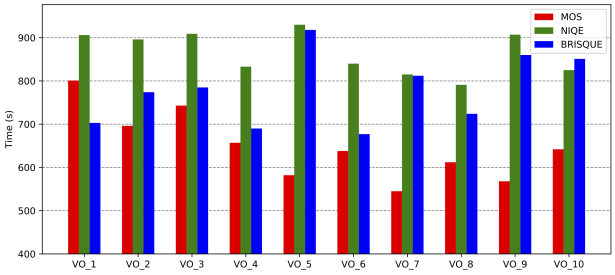

Figure 5: Computational time for VOs individually on OIQA.

for local quality related to extracted regions. However, the used blind metrics did not improve the prediction accuracy globally in terms of PLCC and SRCC.

The proposed data-augmentation by the use of all VOs combined improved the performances for all the three cases, regardless of the used database, as can be seen in Table 1. The PLCC (resp. SRCC) value shifted from an average of 0.830 (resp. 0.762) to 0.871 (resp. 0.801) for the MOS based training on CVIQD. A similar behaviour is observed on OIQA where an improvement is achieved over the performance of individual observers. As for the use with NIQE and BRISQUE, an improvement is also observed both for PLCC and SRCC. The level of improvement should be put into perspective over the range of performances, which is higher for the blind metrics. Based on the previous correlation results, NIQE and BRISQUE do not appear as the best alternative to replace the MOS for data-augmentation. However, other performance data should be analyzed before drawing final conclusions.

When comparing between databases, one can observe a higher performance on OIQA compared to CVIQD, except with NIQE. The difference is obvious with the MOS-based training supporting the previously discussed observation regarding the variety and diversity of the content present on OIQA. This led to a significant performance *i.e.* PLCC (resp. SRCC) value of 0.920 (resp. 0.904) compared to 0.871 (resp. 0.801) on CVIQD.

Table 2: Computational complexity in terms of training time for data-augmentation on CVIQD and OIQA databases. The mean of 5 folds is provided.

	Database	MOS	NIQE	BRISQUE
Time (s)	CVIQD	6285	3093	2577
	OIQA	3212	2360	3320

With the intent to compare the computational complexity of the proposed data-augmentation, we compute the training time for individual VOs as well as their combination (data-

augmentation). It is given on Fig. 4 for CVIQD and Fig. 5 for OIQA where one can notice that, the MOS-based training has the lowest training time for all VOs except for $VO_{3,5,7}$ on CVIQD and VO_1 on OIQA. This could be explained by the lack of scores diversity during the learning process, leading to a faster convergence. At the contrary, BRISQUE generates a considerably higher training time, which is even more extensive for NIQE on both databases. This observation becomes invalid when it comes to the proposed data-augmentation, as shown by Table 2. Hence, the MOS-based case requires more than twice the NIQE/BRISQUE training time on CVIQD, and on OIQA the BRISQUE-based training took the longest time. This can be explained by the fact that, learning from a considerable amount of data that is associated with the same quality score (*i.e.* MOS) tends to make the model converge slowly. More data implies more diversity for the model to learn from. However, associating this diverse data with the same labels has a negative effect by increasing the computational cost. With NIQE and BRISQUE, the model is able to converge quickly as more data is available with distinct quality scores.

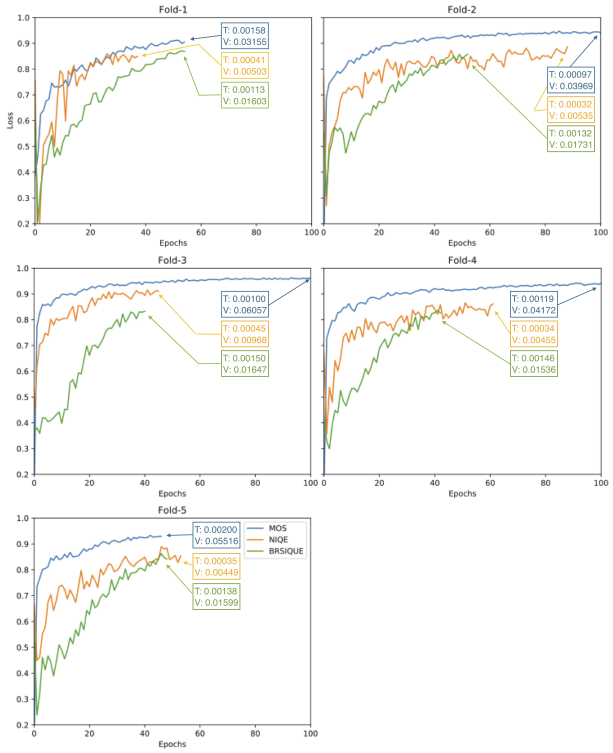

Figure 6: Contrast (max-min/max+min) between training and validation losses for the five folds (0 → equal loss between training and validation and 1→ important gap between both losses) on CVIQD. T and V represent the reached loss values for training and validation, respectively.

In addition to the computational time, we analyzed the evolution of the loss for the data-augmentation case. Figs. 6 and 7 plot the contrast (max-min/max+min) between training and validation losses for the five folds. A contrast equal to 0 depicts an equal loss between training and validation. On the contrary, a contrast equal or close to 1 indicates an important gap between both losses. In addition to the contrast, Fig. 6 and 7 provide the final loss values for both training (T) and validation (V) for each fold and each studied case on CVIQD and OIQA respectively. We can see that the MOS-based learning has more difficulties to generalize either with OIQA or CVIQD. In fact, the gap between T and V for MOS is much higher than those of NIQE- and BRISQUE-based cases. This is also demonstrated by the provided curves for both databases.

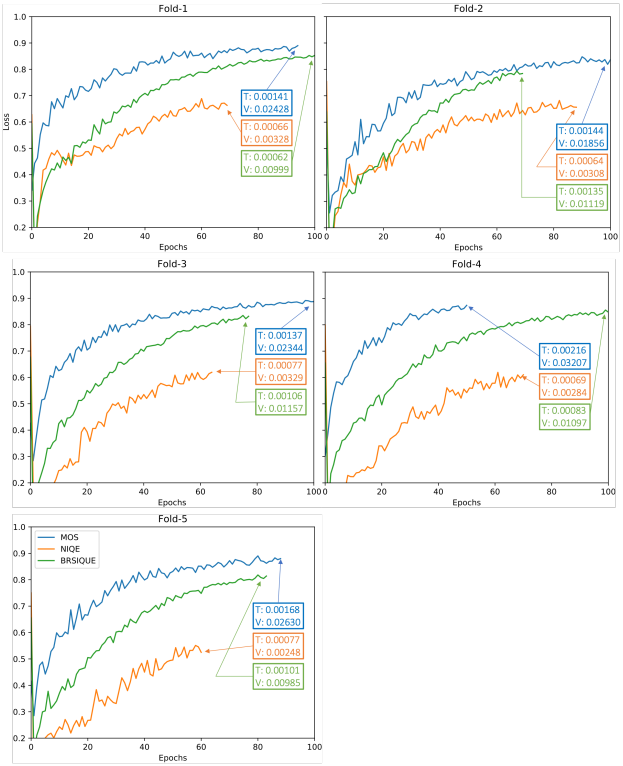

Figure 7: Contrast (max-min/max+min) between training and validation losses for the five folds (0 → equal loss between training and validation and 1→ important gap between both losses) on OIQA. T and V represent the reached loss values for training and validation, respectively.

Conclusion

In this paper, we analyzed the use of visual scan-paths as data-augmentation for 360-degree IQA, mainly for reducing over-fitting and improving the prediction performances. To do so, ten different scan-paths (simulating 10 virtual observers) were generated with eight possible fixation points each, are used as centers of the generated viewports. We used two benchmark IQA 360-degree image databases. Additionally, a comparison is made between the use of blind metrics (BRISQUE and NIQE) and MOS for local quality (quality of viewports) for training the model. The obtained results demonstrated an improvement when using data-augmentation compared to individual virtual observers. The lack of diversity of the MOS values associated with the same 360-degree image, does not allow to reach sufficient generalization of the model. In addition, it requires more computational time than the cases using blind metrics. It is important to increase the number of datasets in order to validate the previous conclusions.

Acknowledgments

This work is funded by the Region "Nouvelle Aquitaine" under project SIMOREVA360 2018-1R50112.

References

[1] DM. Chandler. Seven challenges in image quality assessment: past, present, and future research. *International Scholarly Research Notices*, 2013, 2013.

[2] Y. Niu, Y. Zhong, W. Guo, Y. Shi, and P. Chen. 2d and 3d image quality assessment: A survey of metrics and challenges. *IEEE Access*, 7:782–801, 2018.

[3] G. Zhai and X. Min. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63:1–52, 2020.

[4] M. Yu, H. Lakshman, and B. Girod. A framework to evaluate omnidirectional video coding schemes. In *IEEE International Symposium on Mixed and Augmented Reality*, pages 31–36, Fukuoka, Japan, 2015.

[5] A. Sendjasni, MC. Larabi, and F. Alaya Cheikh. On the improvement of 2D quality assessment metrics for omnidirectional images. In *Electronic Imaging*, pages 287–1, Burlingame, California USA, 2020.

[6] Y. Sun, A. Lu, and L. Yu. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE Signal Processing Letters*, 24:1408–1412, 2017.

[7] V. Zakharchenko, PC. Kwang, and HP. Jeong. Quality metric for spherical panoramic video. In *Optics and Photonics for Information Processing X*, volume 9970, pages 57 – 65, 2016.

[8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, Honolulu, HI, USA, 2017.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, Las Vegas, NV, USA, 2016.

[10] J. Kim and S. Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1969–1977, Honolulu, HI, USA, 2017.

[11] S. Seo, S. Ki, and M. Kim. A novel just-noticeable-difference-based saliency-channel attention residual network for full-reference image quality predictions (early access). *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2020.

[12] O. Russakovsky, J. Deng, H. Su, J Krause, and S. Satheesh et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[13] W. Sun, X. Min, G. Zhai, K. Gu, H. Duan, and S. Ma. MC360IQA: A multi-channel cnn for blind 360-degree image quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):64–77, 2020.

[14] H. G. Kim, H. Lim, and Y. M. Ro. Deep virtual reality image quality assessment with human perception guider for omnidirectional image. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):917–928, 2020.

[15] J. Xu, W. Zhou, and Z. Chen. Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks. *IEEE Transactions on Circuits and Systems for Video Technology (Early Access)*, pages 1–1, 2020.

[16] A. Sendjasni, MC. Larabi, and F. Alaya Cheikh. Perceptually-weighted CNN for 360-degree image quality assessment using visual scan-path and JND. In *IEEE International Conference on Image Processing (ICIP). (To appear)*, Anchorage, Alaska, USA, 2021.

[17] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2020.

[18] T. Truong, H. Tran, and T. Thang. Non-reference quality assessment model using deep learning for omnidirectional images. In *IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, pages 1–5, Morioka, Japan, 2019. IEEE.

[19] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[20] A. Mittal, R. Soundararajan, and A. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[21] D. Noton and L. Stark. Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision research*, 11(9):929–IN8, 1971.

[22] W. Sun, Z. Chen, and F. Wu. Visual scanpath prediction using ior-roi recurrent mixture density network (early access). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[23] W. Sun, K. Gu, S. Ma, W. Zhu, N. Liu, and G. Zhai. A large-scale compressed 360-degree spherical image database: From subjective quality evaluation to objective model comparison. In *IEEE 20th international workshop on multimedia signal processing (MMSP)*, pages 1–6, Vancouver, BC, Canada, 2018.

[24] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang. Perceptual Quality Assessment of Omnidirectional Images. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, Florence, Italy, 2018.

[25] J. Brownlee. *Deep learning for computer vision: image classification, object detection, and face recognition in python*. Machine Learning Mastery, 2019.

[26] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, Columbus, OH, USA, 2014.

[27] A. Sendjasni, MC. Larabi, and F. Alaya Cheikh. Convolutional neural networks for omnidirectional image quality assessment: pre-trained or re-trained? In *IEEE International Conference on Image Processing (ICIP). (To appear)*, Anchorage, Alaska, USA, 2021.

[28] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, Santiago, Chile, 2015.

[29] Siti. https://vqeg.github.io/software-tools/quality%20analysis/siti/.

[30] D. Hasler and S. Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, volume 5007, pages 87–95. International Society for Optics and Photonics, 2003.

[31] A. Martín, P. Barham, J. Chen, Z. Chen, and A. Davis et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation*, pages 265–283, Savannah, GA, USA, 2016.

[32] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.