Quality of Experience using Deep Convolutional Neural Networks and future trends

Woojae Kim, Jaekyung Kim and Sanghoon Lee
Yonsei University, Seoul, Republic of Korea
E-mail: {wooyoa, jkkproject, slee}@yonsei.ac.kr Tel/Fax: +82-2-2123-7734

***Abstract*—The development of immersive display technology enables to represent the details of contents more naturally by providing a more realistic viewing environment while increasing immersion. In parallel, quality of experience (QoE) has been dealt with and discussed from both academy and industry to grade consumer products from the quality perspective. However, for quantification of QoE, it is very challengeable to analyze the human perception more accurately, even if it has been studied in many decades. Currently, there is no solid methodology to verify human perception as a closed-form objectively due to the limitation of human perception analysis. Recently, the deep convolutional neural network (CNN) has emerged as a core technology while breaking most performance records in the area of artificial intelligence via intensive training in accordance with the massive dataset. The main motivation of this paper lies in finding new insight into human perception analysis for QoE evaluation through visualization of intermediate node values. This new QoE assessment approach enables us to figure out the human visual sensitivity without using any prior knowledge. Toward the end, we provide a novel clue of how to obtain visual sensitivity, which is expected to be essentially applied for future QoE applications. In addition, we discuss future applications in QoE assessment with respect to the display types.**

I. Introduction

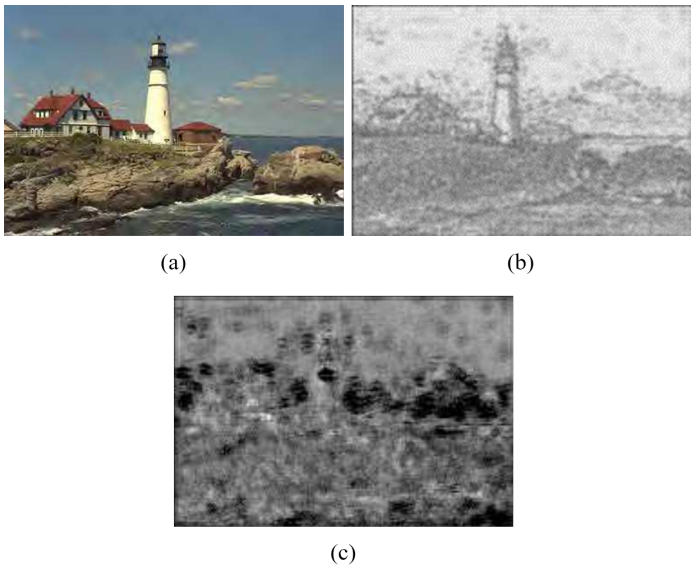
With the development of digital imaging technology, a number of display types have emerged, while offering a variety of viewing environments and accommodating users to enjoy a versatile user experience. With the rapid development of these new technologies, people have easily acquired or even edited contents with imaging devices such as digital cameras, smartphones, multi-cameras and 3D modeling tools. In addition, the contents can be easily visualized in real-life through various devices of 2D display, stereoscopic 3D (S3D) display and head-mounted display (HMD) [1], which even enables users to interact with new spaces and objects. For this reason, the quality of experience (QoE) that people perceive in each display has become much more diverse and personalized than before, while being adaptive to different service scenarios.

Thereby, the study of predicting and evaluating this has been actively carried out not only for engineering inquiry but also for understanding the consumer-centered market and trend. Accordingly, the analysis of human perception and content is becoming a more important problem, for this reason, the emergence of new displays or platforms that requires more sophisticated and novel quality assessment techniques.

For conventional 2D display, image quality assessment (IQA) and video quality assessment (VQA) have been actively studied to solve the deterioration of visual quality with the development of compression or transmission technology in order to provide a higher QoE environment. To achieve this, many researchers have attempted to verify the out-performance of their metrics by demonstrating that the errors obtained by using their metrics are highly correlated with human perception errors [2]. Recently, contrast IQA and sharpness IQA [3] have been developed for the visual preference for post-processing reflecting the aesthetic view of the image. In addition, visual saliency detection, which is to find the local area of the content visually focused by the user has also been widely used as a factor for predicting the target QoE more accurately [4], [5], [6]. In addition, foveation, which has been dealt with as a prominent visual property due to uneven distribution of photoreceptors on the retina, has also been extensively studied in QoE [7], [8]. At the same time, viewing geometry analysis, which estimates the perceived resolution in consideration of user's viewing distance, display resolution and resolution of human vision, has also been utilized in many fields.

In the case of S3D display, it enables to provide a 3D experience using stereoscopic image/video (left/right paired content) to maximize the 2D display experience. Since this is still based on 2D images, numerous I/VQA studies have been performed similarly to what has been done before [9]. However, the S3D display causes discomfort due to the vergence-accommodation mismatch of human vision, and thereby visual discomfort prediction (VDP) studies have been actively investigated. As a result, the binocular fusion principle was used to produce a synthesized cognitive image called cyclopean image [10], [11], [12]. For this research, modeling has been performed from various angles to analyze stereoscopic recognition of human brain processes [10]. In keeping with this trend, virtual reality (VR) using HMD has been actively studied. However, in the VR content, VR sickness caused by VR experience hinders the viewing, and acts as an obstacle to market activation. Therefore, studies on VR sickness assessment (VRSA) have been actively conducted to solve this problem while reflecting the visual-vestibular sensory conflict [13], [14].

However, it is still not easy to design an accurate QoE assessment model since the perceptual features are obtained based on different prior knowledge depending on the types of display and content. In addition, since the user performs QoE evaluation nonlinearly according to the content (or display type), it is essential to predict the appropriate HVS-related prior knowledge in conjunction with the prediction task. In



(a) (b)

(c)

Fig. 1. Examples of distorted image and its visual sensitivity: (a) is a distorted image; (b) is an objective error map and (c) is a perceptual error map inferred by the HVS embedded deep model. Darker regions indicate more pixel-wise distorted pixels.

recent years, deep-learning technology has shown significant performance improvement in the computer-vision and signal-processing even if the mechanism inside is unknown as a black-box method. Beyond the classification framework, the deep-learning method has been successfully applied to regression problems in image/video QoE assessment problems. Thanks to the benefit of massive data information, QoE metrics enable to demonstrate state-of-the-art performance. However, it is still not easy to analyze the model from the synthetic information of an image. Therefore, except for the performance enhancement, there has been a drawback that visual analysis looks impossible from the HVS perspective. In this respect, we first present a new approach of a CNN model of embedding human visual sensitivity, which enables us to analyze the human perception while being accompanied by a high predictive performance. Then, we secondly discuss future applications in QoE assessment according to the deep-learning techniques and display types.

II. Deep Learning based Human Visual Sensitivity Learning

Since the ultimate observers of commercial contents are end-users, the primary goal of QoE assessment is to quantify the perceptual behavior of the HVS for each QoE task. Nevertheless, there is no solid work to quantify the QoE more generally due to the deep involvement of human perception, which has been actively discussed in the area of neuroscience, psychology, physiology and signal processing. One of the core HVS factors is a human visual sensitivity that explains which part of an image is more sensitive to the HVS, i.e., indicates the spatial strength of visual response to the visual information.

As aforementioned, the human visual sensitivity means the perception change of a user according to the characteristics of given spatial signal [15], [16], [17]. When the user views the visual contents, certain local spatial signals of the image are emphasized or masked according to the spatial characteristics of the image. Fig. 1 depicts an example of how much difference the user perceives between a distorted image and the associated error signal. Fig. 1 (a) is a distorted image, (b) is the error map from the original where the darker region is more erroneous than the brighter one, and (c) is the visual sensitivity of being obtained from the deep-CNN where the darker region is more sensitive to the HVS [16]. In (b), it is shown that the rock region contains more errors than the sky region. However, in (a), it is apparent that the distortions of monotonous local areas such as the sky regions look more errors than the rock regions from our observation. With a help of the visual sensitivity map in (c), it is obvious that the observed visual errors and sensitivity are highly correlated, which is quite different from what we observe from the objective errors in (b).

From the visual science perspective, this phenomenon is caused by the visual masking effect by human visual sensitivity, which has been analyzed by the change of QoE awareness according to various characteristics of contents. The visual masking effect means that the HVS exhibits different contrast sensitivity according to spatial pixel distribution. The contrast sensitivity function is a representative model for the description of this phenomenon [18]. Indeed, since the visual cortex is more complicatedly responsive when a human perceives the presence of texture, sub-band decomposition techniques such as Gabor filters or steerable pyramids have also been used for preprocessing to quantify the visual sensitivity numerically. Based on these findings, to accomplish the QoE task, researchers have attempted to embed human visual sensitivity in the metrics of performing the QA, visual discomfort prediction, *etc*.

For clarity, Fig. 2 depicts a simple comparison on how to apply the visual sensitivity for the accomplishment of IQA according to the methodology; conventional full-reference (FR)-IQA metric, deep no-reference (NR)-IQA model, HVS embedded deep IQA model and the future concept of a deep QoE model. Conventional IQA metrics extract feature maps that mimic the visual sensitivity of the HVS from content and predict quality through channel decomposition that reflects visual perception using Gabor filters or steerable pyramids as mentioned previously. Nonetheless, because the human visual behavior of content perception is very complicated to infer through those simple hand-crafted features, this generic process has followed performance limitations.

Recently, numerous attempts have been made to adopt a deep-learning approach for the IQA problem. Even though CNN has demonstrated significant performance improvement in QoE assessment field [15], it is still difficult to figure out the physical meaning of the model by simply regressing subjective scores. Therefore, there has been a lack of visual analysis from the HVS perspective. In order to overcome such a drawback

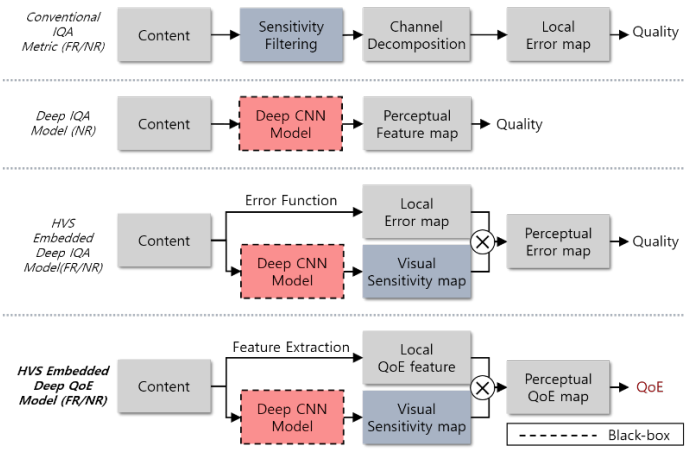


Fig. 2. Flowchart comparison of the conventional FR-IQA metric, deep NR-IQA model, HVS embedded deep IQA model and the concept of a future deep QoE model. FR and NR indicate full-reference and no-reference manner, respectively.

of the CNN based model, we have investigated a new type of a CNN model of embedding human visual sensitivity, which enables to visualize the human perception while being accompanied by high predictive performance. The third row of Fig. 2 demonstrates a flow of the HVS embedded deep IQA model. The biggest difference from the existing CNN based model is that the visual sensitivity map is visualized as an intermediate result of the model and the visual weight of the given local error map can be analyzed. This model obtains the perceptual error map through an elementwise product of the weight and error maps, which are generated by inference based on characteristics of the input image. In addition, the local error map can be obtained in an FR manner using the reference image [16], [17], or in an NR manner inferring the error map from the distorted image itself [19], [20]. Then, through a regression procedure of the perceptual error map to subjective quality scores made by users, the deep IQA model is trained.

Fig. 1 (c) shows the perceptual error map weighted by using the HVS embedded deep IQA where the dark areas represent more distorted pixels perceptually. In contrast to the objective error map in (b), since the area around the sky house is more monotonous, it can be seen that it has become relatively more emphasized in the perceptual error map. The main advantage of this approach is that it learns the visual sensitivity characteristics of the HVS without prior knowledge. The deep CNN can greatly improve the performance of the IQA model as well.

III. Future Applications

1) Deep learning approach on QoE Assessent: For the IQA work, human visual sensitivity has been successfully visualized, which enables to provide deep insight into how human perception is responsive to an input image. This approach is more advanced in the sense that human perception is obtained without prior knowledge, which is different from conventional methods of obtaining handcraft features. So, how can this technology be applied to future QoE metrics? As we have introduced, display technology evolves to a larger, sharper, more immersive environment. The various QoE applications for this could fall in the area of S3D-visual discomfort prediction, sharpness and contrast assessment and VRSA. Due to the intricately involved visual factors, currently, no solid numerical definition has not been published yet, but it is expected that new QoE metrics will be developed by modeling the HVS similarly to the mechanism used for IQA works. In general, the initial input of most content-oriented QoE is visual information, the visual sensitivity of visual content can be applied regardless of the types of tasks such as QA and visual discomfort prediction. In the fourth row of Fig. 2 shows an important clue on how to evolve the HVS embedded deep model to obtain the perceptual quality map of QoE. As shown in the figure, the visual sensitivity induced by the content itself can be mixed with its local QoE features (e.g. motion, depth, contrast, sharpness) and used as an element to predict QoE.

For example, in VRSA work which is one of the recent issues, visually induced motion sickness is caused by sensory mismatches between the motion perceived by the vestibular organ and the motion perceived by the HVS [13]. For this phenomenon, the distribution of the spatial texture has a great effect on the motion perception of the HVS. For example, in the monotonous background of Fig. 1 (a), the human is not aware of the motion that occurs in the content because there are relatively few temporal variations in which the motion is perceived. In contrast, for the rock and lighthouse, there are various spatial frequency components, so temporal variation is large, which makes the user more aware of motion. Thereby, it is expected that the motion component of an image and the weighting process of the visual sensitivity map extracted from the HVS embedded deep model can be effectively applied to calculate the visually perceived QoE.

Moreover, it is expected that for the S3D-visual discomfort prediction works [11], [10], such a primitive approach of inferring visual sensitivity will lead to a higher performance improvement where the depth information induces the visual discomfort on the spatial domain. Furthermore, in other HVS based QoE studies such as 2D/3D visual presence assessment to quantify visual scene satisfaction of viewers [21], perceptual contrast/sharpness assessment model [3], [22] and visual information measurement works [23], [24], the visual sensitivity also has utilized as a pre-processing method reflecting the users' visual perception. Based on these prospects, the HVS embedded deep model is expected to play a key role not only in improving the QoE field performance but also in exploring the visual perception.

2) QoE on Future Displays: On the basis of the recent trend, display technology has been growing for the user to entertain high-quality content. Advances in these technologies continue to raise user satisfaction with larger screens and even sophisticated user interaction. From this point of view, it can be easily inferred that the resolution of the 2D display is getting higher, the S3D display combines a sense of depth

on 2D space, and the HMD device is expanded for a more interactive experience. As if the sprout comes out of the ground, this has been continuing to evolve toward areas where the user is deeply immersive to media just like reality rather than experience. Therefore, it is expected that future displays such as augmented reality, holographic display and light-field display, will allow more realistic stereoscopic regardless of the viewing position. In this respect, it is going to be vital to quantify QoE based on the human-factor accompanied by the display. Currently, the display technology mentioned above is ongoing, and contents are produced by providers and producers with all their best efforts. Thus, the QoE issues keep being brought up at both industry and academic sides.

Recently, for more elaborate QoE control, displays perform scene understanding to augment visual content. Therefore, contextual QoE for the visualized space on the display is expected to play an important role in the future market. In addition, considering the perceptual factors of the device is expected to elevate technology to guarantee the viewing safety and satisfaction of the user.

IV. Conclusion

In this paper, we have introduced a new paradigm for the provision of how to resolve the human perception for future QoE tasks where the CNN model learns the human visual sensitivity. By using the visual sensitivity map learned by a deep model, we expect that each visual QoE feature map can be adopted directly to QoE metrics and it will lead to huge improvement in performance and perceptual analysis. In the constantly evolving immersive display technology, visual sensitivity studies are getting essential in the QoE field. In this regard, HVS embedded deep models will mark a new era in the QoE manner with the help of the infinite possibilities of artificial intelligence.

V. Acknowledgments

This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-IT1702-08

References

[1] Kjell Brunnström, Sergio Ariel Beker, Katrien De Moor, Ann Dooms, Sebastian Egger, Marie-Neige Garcia, Tobias Hossfeld, Satu Jumisko-Pyykkö, Christian Keimel, Mohamed-Chaker Larabi, et al., "Qualinet white paper on definitions of quality of experience," 2013.

[2] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[3] Haksub Kim, Sewoong Ahn, Woojae Kim, and Sanghoon Lee, "Visual preference assessment on ultra-high-definition images," *IEEE Transactions on Broadcasting*, vol. 62, no. 4, pp. 757–769, 2016.

[4] Haksub Kim, Sanghoon Lee, and Alan Conrad Bovik, "Saliency prediction on stereoscopic videos," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1476–1490, 2014.

[5] Haksub Kim and Sanghoon Lee, "Transition of visual attention assessment in stereoscopic images with evaluation of subjective visual quality and discomfort," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2198–2209, 2015.

[6] Anh-Duc Nguyen et al., "Deep visual saliency on stereoscopic images," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1939–1953, 2018.

[7] Sanghoon Lee, Marios S Pattichis, and Alan C Bovik, "Foveated video compression with optimal rate control," *IEEE Transactions on Image Processing*, vol. 10, no. 7, pp. 977–992, 2001.

[8] Sanghoon Lee, Marios S Pattichis, and Alan C Bovik, "Foveated video quality assessment," *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 129–132, 2002.

[9] Heeseok Oh, Sewoong Ahn, Jongyoo Kim, and Sanghoon Lee, "Blind deep s3d image quality evaluation via local to global feature aggregation," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4923–4936, 2017.

[10] Jincheol Park, Heeseok Oh, Sanghoon Lee, and Alan Conrad Bovik, "3d visual discomfort predictor: Analysis of disparity and neural activity statistics," *IEEE Transactions on Image Processing*, vol. 24, no. 3, pp. 1101–1114, 2014.

[11] Heeseok Oh, Sanghoon Lee, and Alan Conrad Bovik, "Stereoscopic 3d visual discomfort prediction: A dynamic accommodation and vergence interaction model," *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 615–629, 2015.

[12] Heeseok Oh, Jongyoo Kim, Jinwoo Kim, Taewan Kim, Sanghoon Lee, and Alan Conrad Bovik, "Enhancement of visual comfort and sense of presence on stereoscopic 3d images," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3789–3801, 2017.

[13] Jaekyung Kim, Woojae Kim, Sewoong Ahn, Jinwoo Kim, and Sanghoon Lee, "Virtual reality sickness predictor: Analysis of visual-vestibular conflict and vr contents," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018, pp. 1–6.

[14] Jinwoo Kim, Woojae Kim, Heesoek Oh, and Sanghoon Lee, "A deep cybersickness predictor based on brain signal analysis for virtual reality contents," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2019, pp. 1–6.

[15] Jongyoo Kim and Sanghoon Lee, "Fully deep blind image quality predictor," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 206–220, 2017.

[16] Jongyoo Kim and Sanghoon Lee, "Deep learning of human visual sensitivity in image quality assessment framework," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[17] Woojae Kim, Jongyoo Kim, Sewoong Ahn, Jinwoo Kim, and Sanghoon Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," *Proc. Eur. Conf. Comput. Vis.(ECCV)*, 2018.

[18] Scott J Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," in *Human Vision, Visual Processing, and Digital Display III*. International Society for Optics and Photonics, 1992, vol. 1666, pp. 2–15.

[19] Jongyoo Kim, Anh-Duc Nguyen, and Sanghoon Lee, "Deep cnn-based blind image quality predictor," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 1, pp. 11–24, 2018.

[20] Jongyoo Kim, Woojae Kim, and Sanghoon Lee, "Deep blind image quality assessment by learning sensitivity map," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6727–6731.

[21] Heeseok Oh and Sanghoon Lee, "Visual presence: Viewing geometry visual information of uhd s3d entertainment," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3358–3371, 2016.

[22] Woojae Kim, Haksub Kim, Heeseok Oh, Jongyoo Kim, and Sanghoon Lee, "No-reference perceptual sharpness assessment for ultra-high-definition images," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 86–90.

[23] Kwanghyun Lee and Sanghoon Lee, "A new framework for measuring 2d and 3d visual information in terms of entropy," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 11, 2015.

[24] Kwanghyun Lee and Sanghoon Lee, "3d perception based quality pooling: Stereopsis, binocular rivalry, and binocular suppression," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 3, pp. 533–545, 2015.