SIP (2019), vol. 8, e23, page 1 of 19 © The Authors, 2019.

doi:10.1017/ATSIP.2019.16

ORIGINAL PAPER

Modern trends on quality of experience assessment and future work

WOOJAE KIM,[1] SEWOONG AHN,[1] ANH-DUC NGUYEN,[1] JINWOO KIM,[1] JAEKYUNG KIM,[1] HEESEOK OH[2] AND SANGHOON LEE[1]

Over the past 20 years, research on quality of experience (QoE) has been actively expanded even to cover aesthetic, emotional and psychological experiences. QoE has been an important research topic in determining the perceptual factors that are essential to users in keeping with the emergence of new display technologies. In this paper, we provide in-depth reviews of recent assessment studies in this field. Compared to previous reviews, our research examines the human factors observed over various recent displays and their associated assessment methods. In this study, we first provide a comprehensive QoE analysis on 2D display including image/video quality assessment (I/VQA), visual preference, and human visual system-related studies. Second, we analyze stereoscopic 3D (S3D) QoE research on the topics of I/VQA and visual discomfort from the human perception point of view on S3D display. Third, we investigate QoE in a head-mounted display-based virtual reality (VR) environment, and deal with VR sickness and 360 I/VQA with their individual approach. All of our reviews are analyzed through comparison of benchmark models. Furthermore, we layout QoE works on future display and modern deep-learning applications.

Keywords: Quality of Experience (QoE), Human Visual System (HVS), Image Quality Assessment (IQA), Video Quality Assessment (VQA), Visual Preference Assessment, Visual Discomfort Assessment (VDP), VR Sickness Assessment (VRSA)

Received 29 July 2019; Revised 10 September 2019

I. INTRODUCTION

Along 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 (UX). 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. Above all, the development of social networks and mobile devices makes sharing of imaged information even more massive than before. 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.

In [1,2], QoE is defined as "a measure of the overall level of customer satisfaction with a vendor". Understanding this definition intuitively, QoE may seem to be similarly categorized as an extended version of quality of service. However, it does not just mean visual quality delivered over network, but needs to be described with additional perspectives over new domains. Currently, QoE is abstractly stated in literature while including a different level of emotional experience for humans. Naturally, the interpretation as a form of numerical formula is also different so that many metrics, feature values, and subjective evaluations have been published in various research fields. Nevertheless, the QoE paradigm can be widely applied to all consumer-related content business or spell out services from both sides of service provider and customer. Indeed, user satisfaction is directly linked to corporate profits so that QoE leads to gain momentum as an important criterion not only in multimedia services and systems but also in other areas including content design, human–computer interaction, and aesthetics. This trend is also related to the rapidly increasing demand and market size in the UX sector, which has recently exploited multi-dimensional visualization [1]. With the opposite pay for it, the analysis of human perception and content is becoming a more complicated

[1]Department of Electrical and Electronics Engineering, Yonsei University, Seoul 03722, Korea

[2]Electronics and Telecommunications Research Institute (ETRI), Daejeon 34129, Korea

Corresponding author:
Sanghoon Lee,
E-mail: slee@yonsei.ac.kr

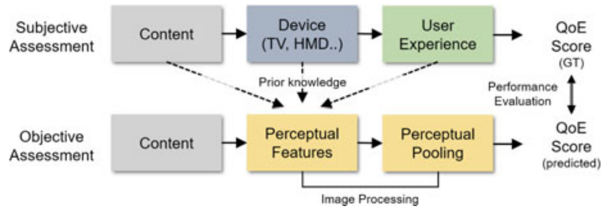

Fig. 1. Framework of subjective and objective evaluation for QoE.

problem, and for this reason, the emergence of new displays or platforms that provide more versatile UX requires more sophisticated and novel quality assessment techniques.

QoE can be quantified by various methods depending on the type of signal to be processed. In this paper, we use a generic QoE measurement process depicted in Fig. 1. As shown in the figure, traditionally, the QoE assessment can be divided into two categories. One is subjective assessment and the other is objective assessment. The former expresses feedback on the most accurate QoE of a user who has experienced digitized contents through the device, and this method has been regarded as the most reliable one among all possible means currently. So far, numerous methodologies have been proposed to perform the subjective assessment of various QoEs [3,4]. The double-stimulus continuous quality scale (DSCQS) measures the difference in QoE of a target content and that of the reference content. Then, the differential mean opinion score (DMOS) can be obtained from the statistical scores of all subjects performed therefrom. However, as one of the drawbacks, this method needs two stimuli simultaneously to the user. If no reference content is available (e.g., visual discomfort), it is difficult to employ DSCQS. For this reason, the single-stimulus continuous quality evaluation (SSCQE) has been widely used to overcome such drawbacks. Nevertheless, SSCQE has a disadvantage in that it causes the user to be fatigued and to spend a lot of time in the evaluation process of the massive contents.

Therefore, most studies have focused on the objective assessment of QoE mentioned in the latter. Unlike the subjective assessment, this predicts the QoE score through feature extraction and regression via image processing of the content information. The main flow of objective assessment lies in analyzing the behavior of the human visual system (HVS) as a function. It then uses the output values of the function as prior knowledge to extract perceptual features from the content, and maps these features to a single score through the pooling process. As mentioned above, it is not easy to design an accurate prediction model because 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, it is essential to predict the appropriate HVS-related prior knowledge in conjunction with the prediction task. In fact, since there are limitations in studying the HVS biologically and psychologically as a closed form of the prior knowledge, researchers lean on reverse engineering. Thus, researchers have rigorously measured the response of human visual perception to the contents available on the individual display and to their device characteristics.

Table 1 tabulates the available content types and recent QoE tasks conducted over the display devices covered in this paper. In the table, the circle mark indicates the available content type for each display, and which QoE task can be applied to. For measurement of experience on 2D display, image quality assessment (IQA) and video quality assessment (VQA) have been actively pursued 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. For example, in the IQA task, the structural similarity (SSIM) [5] is formulated. Associated with the human perception on the spatial information, its value implies a correlation with the phenomenon through the divisive normalization process in the receptive field.

For various QoEs, the image quality evaluation has been diversely evolved depending on the application such as the sharpness of object or contour relative to the background from various visual perspectives. Thus, recently, contrast IQA and sharpness IQA [6,7] 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 concentrated by user has also been widely used as a factor for predicting the target QoE more accurately [8–11]. 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 [12,13]. 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.

Nowadays, the service is gradually evolving into an interactive form through being customized toward satisfying the personal need, and into realizing the stereoscopic effect in order to maximize the presence feeling. In keeping with this trend, S3D display enables to provide a virtual 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 [14]. 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 [15–18]. For this research, modeling has been performed from various angles to analyze stereoscopic recognition of human brain processes [19].

Recently, virtual reality (VR) or augmented reality (AR) using HMD has been actively studied. There are two main types of HMDs. One is a 360-degree VR content type that

Table 1. Comparison of the available content types for each display device with related QoE tasks.

	Available content types					Major QoE tasks					
Display device	Image	Video	S3D image or video	360 Image or video	3D modeled scenario	IQA/ VQA	Contrast IQA	Sharpness IQA	Visual presence	Visual-discomfort assessment	VR sickness assessment
2D display	O	O	–	–	–	O	O	O	O	–	–
S3D display	–	–	O	–	–	O	–	–	O	O	–
HMD	–	–	–	O	O	O	–	–	O	O	O

allows viewing 360-degree panorama image/video w.r.t. the original point. The other is a computer graphic (CG)-based VR content type that experiences a three-dimensional space defined in the 3D modeling platform using Unity or Unreal engine. For 360-degree VR, the tasks of I/VQA are actively performed because they are closely relevant to the topics done for 2D I/VQA in nature. However, in the CG-based 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 [20,21].

Overall, in this paper, we focus on overviewing trends in the QoE from the viewpoint of display technology, and on discussing the details of individual tasks. There have been a small number of QoE-related review papers in the past, but there has been a lack of research to various other QoEs in a holistic way according to display type [1,22,23]. The remainder of this paper is organized as follows. Section II introduces an overview of QoE assessment for general QoE tasks. Then, after reviewing the QoE assessment on image/video content including I/VQA and the visual preference assessment in Section III, we describe the QoE assessment on 3D stereoscopic content including visual discomfort and image/video quality in Section IV. Then, Section V discusses the QoE assessment on HMD device dealing with VRSA and 360-degree VR content I/VQA. In addition, the future trends of QoE and the conclusion are presented in Sections VI and VII, respectively.

II. OVERVIEW OF QOE ASSESSMENT

A) Subjective methodology

It is common sense that the most reliable method for measuring QoE is subjective assessment conducted by human subjects. Traditionally, subjective assessment is conducted in several forms such as explorative research, psychophysical scaling, or questionnaire to gather focus group opinions from subjects after showing them test sequences [3]. The methods referred in international recommendation such as DSCQS and SSCQE have been widely employed in the QoE field [3,24,25]. In recent years, more dynamic contents have been produced by adding additional depth domain, i.e., 3D. In order to capture such dynamics of psychophysical features during watching 3D video, it is necessary to develop more interactive subjective method. For this, the multimodal interactive continuous scoring of quality (MICSQ) method was presented to acquire more reliable quality scores by maximizing convenience of the user scoring process in the middle of entertaining contents, which can be used even in a dark room [26]. Therefore, it is important to choose an appropriate method subjective to the purpose of assessment and to adequate constraints of contents. Details of each method are as follows

1) DSCQS

In the DSCQS method, a pair of reference and assessment contents are successively displayed to the subject in random order. At the end of each second viewing sequence, the subject assessment process is carried out in continuous quality scale between bad to excellent followed by calculating the difference of scores being obtained after watching the reference and assessment contents. In many cases, the score gap is larger when there is significant distinction in time or space domain. Finally, DMOS is obtained by averaging differential scores from all subjects. A less value of DMOS indicates that the subjective score of the test content is near to that of the reference content, so the test content has good quality as close as the reference content. The most significant advantage of DSCQS is robustness to context effect, since DSCQS is a reference content-aware method, i.e., viewers always watch a pair of contents. However, this implies that DSCQS cannot be applied to no-reference (NR) I/VQA where reference contents do not exist. Therefore, DSCQS is suitable for full-reference (FR) where both target and reference contents exist together [27].

2) SSCQE

In contrast to DSCQS, SSCQE is devised to conduct the time-efficient and referenceless subjective assessment for general viewing environments. In this protocol, the subject experiences the long sequence at a time, and evaluates it by the continuous quality scale in real-time. The subjective score is recorded by using a device such as slider or fader. Also, the subject can monitor the scored values for more reliable testing. Especially, while most subjective assessment methods provide only one quality rating for a single content, SSCQE can produce a temporal scoring output. However, as the side effect, there are still issues about accuracy. Since the subject experiences only target content, the contextual effect can be reflected to the scoring judgment. Thereby, there is a possibility that the annotated score can be drifted according to the test sequence due to the user dependency. Moreover, this drawback becomes much worse when the target QoE gets a severe impact on the local region over the temporal domain [27].

3) MICSQ

MICSQ is a more user-friendly way compared to SSCQE [26]. The motivation of MICSQ is a separation of viewing experience from assessment by involving hearing and touching in addition to seeing. Toward this, an additional interaction is employed to minimize distraction from visual immersion. In the cases of SSCQE and DSCQS, no suggestions are given on how the assessment interface is presented on the screen. However, MICSQ utilizes a sub-display such as tablet, to separate the assessment view from the content view, which allows viewers to focus on watching contents while continuously carrying out the assessment process. This helps the subject to score more reliably by providing an environment that they can fully concentrate on visual cues. During the subjective testing, a haptic cue (periodic vibration) and an auditory cue (beeping) are utilized together to prevent the subject from losing sight of core values being recorded, and to enhance the credibility of assessment results. Nevertheless, there is still weakness. The existence of delay and reaction speed variance between subjects may degrade accuracy. This method is intrinsically designed to find specific sections where intense drift of quality, discomfort, or presence exists, so it shows strength in measuring the response of human perception in face with irregularity of stimuli triggered by immersive contents.

B) Objective methodology

The objective QoE assessment is broadly divided into three evaluation manners according to the availability of the reference information: FR, reduced-reference (RR) and NR. FR metrics are generally designed to measure the distance from a target image to the reference image such as mean-squared error (MSE) or peak-signal-to-ratio (PSNR). When the pristine unimpaired stimulus is given, the information of the reference is fully utilized, and the prediction performance is generally higher than the others. The RR approach is applicable to scenarios for image/video communication or transmission. This evaluates QoE by relying on incomplete reference information (e.g., visual feature information) for a given target content. The NR assessment remains the only scheme to be used for the general-purpose application when there exists no reference content. Commercially, this case is the most prevailing case. The NR assessment has been designed by formulating a QoE metric, or by developing an evaluation model from data-driven perspective.

In general, to verify the performance of QoE assessment, researchers have followed three standard measures, i.e., Pearson's linear correlation coefficient (PLCC), Spearman's rank order correlation coefficient (SROCC), and Kendall's rank order correlation coefficient (KROCC) by following the recommendation from the video quality experts group [28]. PLCC is obtained by

$$\mathrm{PLCC} = \frac{\sum_i (q_i - \bar{q}) \cdot (o_i - \bar{o})}{\sqrt{\sum_i (q_i - \bar{q})^2 \cdot \sum_i (o_i - \bar{o})^2}}, \tag{1}$$

where o_i and $\bar{o}$ are the ith subjective score and the mean of o_i. q_i and $\bar{q}$ are the ith predicted score and the mean of q_i.

SROCC is a method of measuring the correlation between two variables by the non-parametric method:

$$\mathrm{SROCC} = 1 - \frac{6}{k(k^2 - 1)} \sum_{i=1}^{k} d_i^2, \tag{2}$$

where d_i is the difference between the subjective and predicted scores for the ith image rank, and k is the image index of the testing set. From equation (2), SROCC can be obtained based on the rank of each difference of subjective and predicted scores. Therefore, even if there is less linear-relationship or regularity, it can derive the correlation as long as the tendency to the ranking is clear. However, it may not operate well if there are several outliers in the difference or if its variance is small.

KROCC is similar to SROCC, while the difference is that KROCC is designed to capture the association between two ordinal variables, not the order itself. KROCC quantifies discrepancy between the number of concordant and discordant rank pairs. This means that KROCC gives stronger penalty to non-sequential cases compared to SROCC. KROCC can be obtained as

$$\mathrm{KROCC} = \frac{N_c - N_d}{\frac{1}{2}N(N-1)}, \tag{3}$$

where N is the number of total rank pairs, N_c and N_d are the numbers of concordant and discordant pairs in the dataset, respectively. This method has more robust performance than SROCC in cases when the sample size is small or multiple ties in rank order exist.

III. QOE ON 2D DISPLAY

A) QoE trend on 2D display

With the evolution of 2D display, image/video content has become the most familiar medium for users. In addition, by virtue of recent advances in network transmission and spread of smartphones, the application of 2D image/video content has become an increasingly important medium for acquiring data and for communicating with others. However, due to limitations of access device, storage, and transmission equipment, digital images can be easily degraded during acquisition, compression, and transmission. For this reason, it is particularly relevant to identify and quantify image distortions since the perceptual distortion severely affects the human understanding of 2D content. This trend has led to the emergence of numerous QoE assessments.

More recently, the demand for high-quality image/video has steadily grown. Nowadays, "high-quality" simply goes beyond the quality of information against loss, i.e., artifact, and more implies aesthetic sense. Toward this, in many studies, post-processing and domain-transfer techniques such as image generation and style transfer have been actively presented to satisfy user expectation of high quality from the aesthetic point of view [29]. Accordingly, new assessments such as contrast IQA and sharpness IQA, which quantify visual preferences of enhanced 2D content, appear as recent core topics to ensure the image quality. In addition, with the advent of ultra-high definition displays, it enables to accommodate greater immersion experience over a wider screen [30]. Therefore, in order to evaluate the QoE afforded by high-resolution image/video, major studies have been conducted while covering the variation of visual perception according to viewing geometry (viewing distance, viewing angle, etc.).

Furthermore, there have been QoE-related studies to model the HVS for more precise 2D QoE assessment. These works contribute to clarify various human perceptual factors dealt in the fields of psychology and neuroscience using formula, and enable to quantify the QoE with a broader understanding of the perceptual process. The following items are introduced as major factors for 2D QoE: foveation, viewing geometry, and visual saliency.

- *Foveation*: The distribution of photoreceptors in the human eye is not uniform and decreases away from the center of the fovea [12,13]. This characteristic is defined as foveation and has been employed as a spatial weight of the 2D domain in many existing studies [7,12,13,31–33]. For example, when a viewer gazes at a fixation point, as shown in Fig. 2(a), peripheral regions of the foveal region Ω can be blurred due to non-uniform distribution of photoreceptors.
- *Viewing geometry*: The perceptual resolution shown by the display varies w.r.t. viewing geometry factors (viewing distance, display resolution, display size, and display types: flat or curved). Therefore, many existing QoE studies have applied viewing geometry to design a prediction model that reflects perceptual resolution [7,33–35]. Figure 2(b) geometrically depicts an example of perceived pixel according to display type. When the viewing position is straight in front, the number of pixels is 4 for 1° of the viewing angle, where the perceived pixel length corresponds to the geometric length of each pixel for a given viewing distance. In contrast, when the viewing position moves to the side of the display as shown in right, the number of pixels is seven for 1° of the viewing angle, so that the perceived pixel length is relatively reduced compared to the previous case. This means the QoE is reduced in the ratio of the perceived pixel length. In contrast, for the curved display, the perceived pixel length variation is smaller because each pixel is relatively close to the viewer due to the curved shape, which becomes a good reference how to make a consumer product through quantifying QoE in terms of viewing geometry.
- *Visual saliency*: Visual attention can be characterized by how much user focuses on visual information on the region of display. The strength is termed "visual saliency" which can be used as an important criterion to measure QoE by figuring out the most critical information in 2D content. The research on saliency is determined in two ways of utilizing bottom-up and top-down visual cues. The bottom-up method is triggered by stimuli of low-level features obtained at the resolution of pixel. Thus, saliency is captured as the distinction of image regions or objects by analysis of low-level signals such as intensity, color, gradient, and shape. In general, image processing techniques have been mainly applied to find visual cues. In contrast, the top-down visual attention is inspired by recognition of objects in daily life from the computer vision perspective. Hereby, top-down saliency models utilize prior knowledge, expectations, or rewards as high-level visual cues to identify the target of interest. Overall, visual saliency prediction is to model the fixation selection behavior as well as biological interest mechanism of the HVS. Therefore, there have been studies to predict the visual saliency objectively through the HVS-based content analysis [8,36].

Figure 2(c) shows the heat-map traced by using an eye-tracker and its saliency predicted map. As shown in the figure, users tend to focus on specific local regions more clearly. Therefore, when user undergoes a specific QoE, it can be seen that the QoE is likely to be induced from the concentrated area. For this reason, in many studies, the saliency prediction has been implemented through the saliency weighting on the target QoE [7,33].

B) QoE tasks on 2D display

1) Image/video quality assessment

2D I/VQA databases

As mentioned in Section II, the I/VQA database containing subjective assessment data plays an important role in measuring the performance of the objective assessment. In the meanwhile, a number of public I/VQA databases have been proposed. In this section, we introduce major databases based on numerous existing studies including six IQA databases: LIVE IQA [37], TID2008 [38], CSIQ IQA [39], LIVE-MD [40], TID2013 [41], and LIVE-Challenge [42], and three VQA databases: LIVE VQA [43], CSIQ VQA [44], and IVP VQA [45].

Table 2 tabulates the comparison of major 2D IQA databases. The LIVE IQA database is one of well-utilized IQA databases containing 29 reference images and 799 distorted images with five distortion types: JP2K compression, white noise (WN), Gaussian blur (GB), and Rayleigh fast-fading (FF) channel distortion. Although most databases have focused on specific distortion components such as compression artifacts and transmission errors, the TID2008 database includes various types of distortion. TID2008 consists of 25 reference images and 1700 distorted images with 17 different distortions at four levels of degradation. Moreover, the TID2013 database is expanded to the dataset of having 3000 distorted images with 24 distortion types at

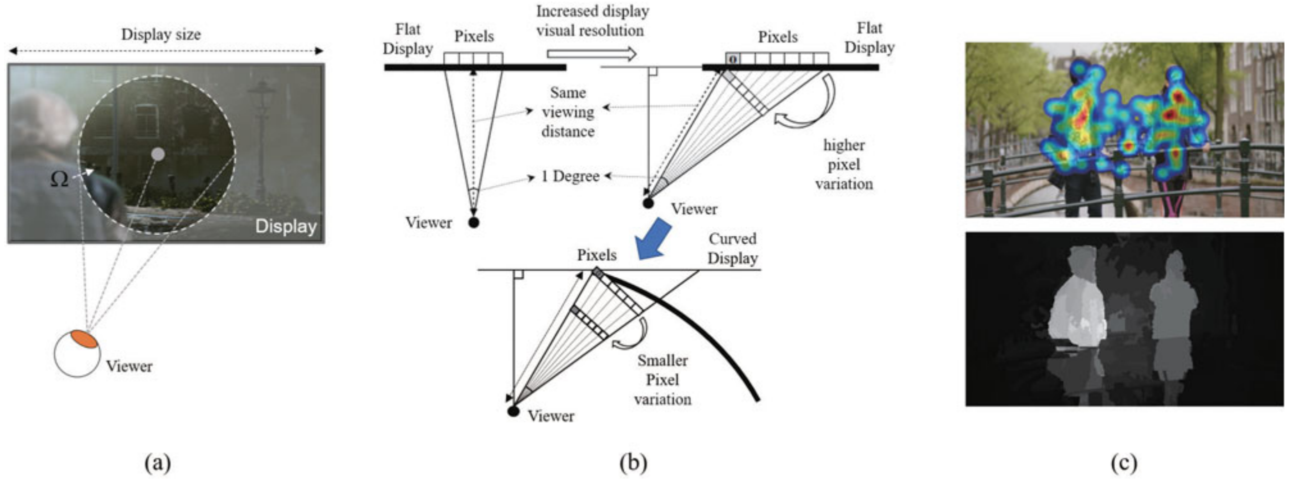

Fig. 2. Representation of related studies for 2D QoE assessment. (a) Foveation, (b) viewing geometry, and (c) visual saliency.

Table 2. Comparison of major 2D image quality assessment databases.

Database	Ref.	Dist.	Dist. types	Res.	Score type	Published year
LIVE IQA [37]	29	799	5	Various	DMOS	2006
TID2008 [38]	25	1,700	17	512 × 384	MOS	2008
CSIQ IQA [39]	30	866	6	512 × 512	DMOS	2009
LIVE-MD [40]	15	405	2	1280 × 720	MOS	2012
TID2013 [41]	25	3,000	24	512 × 384	MOS	2013
LIVE-Challenge [42]	N/A	1,162	Numerous	1280 × 720	MOS	2015

Table 3. Comparison of major 2D video quality assessment databases.

Database	Ref.	Dist.	Dist. types	Res.	Frame rate	Score type	Published year
LIVE VQA [43]	10	150	4	768 × 432	25 or 50	DMOS	2009
IVP VQA [45]	10	128	4	1920 × 1088	25	DMOS	2011
CSIQ VQA [44]	12	216	6	832 × 480	24, 25, 30, 50, 60	DMOS	2014

five levels of degradation. The CSIQ IQA database includes 30 reference images and 866 distorted images with six distortion types: JPEG, JP2K, WN, GB, pink Gaussian noise (PGN), and contrast distortion (CTD). The LIVE MD database includes 15 reference images and 405 distorted images degraded by two multiple types of distortion. One is associated with images corrupted by GB followed by JPEG (GB+JPEG) and the other one is associated with images corrupted by GB followed by WN (GB+WN). Finally, the LIVE challenge database includes almost 1200 unique image contents, obtained by a variety of mobile camera devices under highly diverse conditions. As such, the images were subjected to numerous types of authentic distortions during the capture process such as low-light blur and noise, motion blur, camera shake, overexposure, underexposure, a variety of color errors, compression errors, and many combinations of these and other impairments.

For 2D VQA, three VQA databases are tabulated in Table 3. Unlike the IQA database, the video sequences have high complexity, hence the number of sequences is limited compared to the IQA database. The LIVE VQA database contains 10 references and 150 distorted videos with four distortion types: wireless, IP, H.264, and MPEG-2 compression distortions. The IVP VQA database contains 10 references and 128 distorted sequences with four distortion types: MPEG, Dirac-wavelet, H.264, and packet loss. The CSIQ VQA database includes 12 references and 216 distorted videos with six distortion types: motion JPEG (MJPEG), H.264, HEVC, wavelet compression using the SNOW codec, packet-loss in a simulated wireless network, and additive white Gaussian noise (AWGN).

Major 2D IQA approaches

For the FR-IQA method, most existing works have focused on discovering specific visual characteristics, and on mathematically formulating them from a top-down perspective. Here, we introduce seven well-known FR-IQA approaches. The most intuitive way is calculating error signals between a reference image and its distorted image using the PSNR. Later then, researchers have found that the HVS is a more important factor in perceptual quality. Based on this, Wang *et al.* established a SSIM metric which utilizes divisive normalization and accords with the normalized response of the HVS [5]. Similarly, a variety of studies have been proposed. Lai and Kuo proposed haar wavelet transform-based approach to address HVS-related perceptual distance [46]. The Visual Information Fidelity (VIF) model calculates the information distance between natural scene statistics (NSS)

[47]. FSIM embodies phase coherency in an SSIM-like computation [48]. In addition, a simple and high-efficient model has been proposed using gradient magnitude similarity deviation [49]. With the development of deep-learning technique, Kim. *et al.* proposed a convolutional neural network (CNN)-based FR-IQA model which infers the visual sensitivity map as an intermediate training target of the CNN [50]. Different from existing CNN-based IQA works, this provides visual analysis of the HVS and demonstrates state-of-the-art performance.

For NR-IQA, since not having a reference image, only the statistics of target images is available. There have been several attempts using perceptually relevant low-level feature extraction mechanisms associated with parametric fitting. Here, we introduce six NR-IQA methods. BRISQUE is one of the well-utilized models, which uses NSS features [51]. Ye *et al.* proposed a codebook representation method through learning for NR image assessment (CORNIA) [52]. Zhang *et al.* proposed an effective algorithm called ILNIQE. To characterize structural distortions, they deployed quality-aware gradient statistics feature [53]. Moreover, in [54], dictionary learning was applied to capture useful features from the raw patches. Recently, several deep-learning approaches have been introduced in the literature. BIECON [55] is a novel CNN approach where a new data augmentation method is presented. More recently, they proposed a more reliable model called DIQA by predicting the sensitivity of human perception, and then weighted the sensitivity onto the predicted error map [56].

Major 2D VQA approaches

Similar to 2D FR-IQA, FR-VQA has also focused on the formulation of the distance metric from the HVS point of view. For this reason, some IQA metrics, e.g., PSNR, SSIM, still have been applied to the FR-VQA task accompanied by a temporal pooling method. Recently, Vu *et al.* proposed STMAD which takes into account the visual perception of motion artifacts. STMAD employs a concept of spatio-temporal frame which enables to quantify motion-based distortion [57]. Similarly, they continued their work to ViS3 [58], while separating estimates of perceived degradation into spatial distortion and joint spatio-temporal distortion. More recently, Kim *et al.* proposed a new deep-learning-based FR-VQA approach by learning the spatio-temporal sensitivity map [59]. This work effectively addresses the motion masking effect and provides an attention mechanism-based temporal pooling strategy.

For NR-VQA, V-BLIIND [60] employed a statistical approach from frame difference signals as done in NSS-based approach [51]. In the paper, the authors utilized the frame difference signal which contains temporal variation to extract DCT coefficients, and calculated statistical features. One well-known NR-VQA approach is MOVIE which has been proposed in [61]. By using a spatio-temporal Gabor filter family, they computed quality index as done in SSIM. MOVIE provides significant performance improvement in comparison with existing works. Recently, Li *et al.* proposed a deep-learning-based NR-VQA model called SACONVA [62]. Interestingly, they combined hand-crafted features from 3D shearlet transform, and regressed the CNN output onto the subjective score. Powered by the CNN's strong predictive performance, SACONVA achieved the highest performance in NR-VQA.

2) Visual preference assessment

2D contrast/sharpness IQA database

To compare the performance of the contrast IQA, the CID2013 and CCID2014 databases were presented. The CID2013 consists of 400 contrast distorted images with six contrast changed distortion types. The CCID2014 database is a large-scale contrast-changed database including 655 images with eight contrast-changed distortion types [6].

The camera-shaken image (CSI) database was opened in public for 2D sharpness IQA [63]. The database contains camera-shaken images with resolutions of 1024 × 768 and 1092 × 728. This database is classified into two categories. Category I has 11 natural images and 99 blurred images by linear camera shake. Category II consists of 25 blurred images impaired by complex camera movement. The camera aperture varies in the range from f /2.2 to f /32, and the exposure time range is from *1/40* to *1* second, where f means the focal length of a lens.

Major sharpness IQA approaches

Existing studies for measuring the sharpness of an image can be categorized into two methods. The first is measuring the spread of edges within an image. The edge width is calculated by fitting the distribution of edge region in the blurred image to the Gaussian function [64,65] or by measuring the distance between the start and endpoints of edges [66–68]. Another method is the spectral-based method, which assumes that edges and textured regions comprise high-frequency energy. This method measures the sharpness of images by analyzing the statistical characteristics of coefficients obtained by Fourier transform or discrete cosine transform on the image [69]. However, existing approaches lack consideration on the perceived resolution change according to viewing geometry. To overcome this, Kim *et al.* [33] proposed a sharpness assessment metric that takes into account various factors that affect the perceived resolution.

Major contrast IQA approaches

Recently, Wang *et al.* [70] estimated the perceptual distortion for each component by decomposing the image patches into mean intensity, signal strength, and signal structure. In Gu *et al.* [6], they proposed the RR contrast IQA technique based on phase congruency and information statistics from image histogram. However, since most contrast enhancement techniques do not have reference images, NR methods have been more actively studied in contrast IQA research. Feng *et al.* [71] proposed a blind quality assessment method based on NSS in terms of mean, standard deviation, skewness, and kurtosis. Chen *et al.* [72] extracted feature vectors from feature descriptors and color motions, and then used a regression algorithm to measure the final quality score. The

Table 4. SROCC and PLCC comparison on the five 2D IQA databases. *Italics* indicate the deep-learning-based methods.

		LIVE IQA		CSIQ IQA		TID2013		LIVE-MD		Live challenge	
Type	Method	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
FR	PSNR	0.876	0.872	0.806	0.800	0.636	0.706	0.666	0.704	–	–
	SSIM [5]	0.948	0.945	0.876	0.861	0.637	0.691	0.745	0.767	–	–
	VIF [47]	0.963	0.960	0.920	0.928	0.677	0.772	0.765	0.826	–	–
	FSIMc [48]	0.962	0.962	0.932	0.920	0.851	0.877	0.863	0.818	–	–
	GMSD [49]	0.960	0.960	0.957	0.954	0.804	0.859	0.867	0.890	–	–
	DeepQA [50]	0.981	0.982	0.961	0.956	0.939	0.947	0.937	0.940	–	–
NR	BRISQUE [51]	0.939	0.942	0.775	0.817	0.572	0.651	0.897	0.921	0.607	0.645
	CORNIA [52]	0.942	0.943	0.714	0.781	0.549	0.613	0.900	0.915	0.618	0.662
	ILNIQE [53]	0.902	0.908	0.821	0.865	0.521	0.648	0.902	0.914	0.594	0.589
	HOSA [54]	0.948	0.949	0.781	0.841	0.688	0.764	0.902	0.926	0.659	0.678
	BIECON [55]	0.958	0.962	0.825	0.838	0.721	0.765	0.912	0.928	0.595	0.613
	DIQA [56]	0.975	0.977	0.884	0.915	0.837	0.861	0.939	0.942	0.703	0.704

authors of [73] devised a machine-learning-based algorithm that extracts feature vectors by calculating contrast, sharpness, brightness, colorfulness, and naturalness of images.

3) Benchmarking on 2D QoE tasks

The performance of each 2D I/VQA task was benchmarked by means of PLCC and SROCC. The correlation coefficients were calculated by using the objective/subjective scores. In the case of support vector regressor (SVR) or neural networks (NNs)-based model, the predicted scores have closely fitted to the subjective scores, so no other post-fitting is needed to evaluate the performance. In contrast, for metric-based model, a metric index was developed, but in different scale from the subjective score. Thus, it is necessary to use a logistic function to fit the objective scores to MOS (or DMOS) in order to account for quality rating confining at the extremes of the test range and to prevent the overfitting problem. For this scale conversion, four or five parametric logistic functions have been broadly utilized to fit objective prediction score to subjective quality score as shown in [47,48]. In this benchmark, all the experimental settings followed their origin literature.

Major I/VQA benchmarking

Firstly, to compare the performance of existing 2D IQA methods, five IQA databases were used: LIVE IQA, CSIQ IQA, TID2013, LIVE-MD, and LIVE challenge. Table 4 reports the SROCC and PLCC of the compared FR/NR-IQA algorithms with the five different databases. Here, 12 existing methods are compared by means of six FR-IQA metrics: PSNR, SSIM [5], VIF [47], FSIMc [48], GMSD [49], DeepQA [50], and of six NR-IQA methods: BRISQUE [51], CORNIA [52], ILNIQE [53], HOSA [54], BIECON [55], and DIQA [56]. As listed in Table 4, the best performance is done by using the deep-learning approach. For both FR/NR IQA benchmarking, DeepQA and DIQA show the best performance for the overall databases. Generally, FR-IQA methods show higher performance than NR-IQA methods since reference images can be utilized as additional information for FR-IQA. For the TID2013 database, conventional metrics such as SSIM and VIF do not perform well since TID2013 has more widespread types of distortion than other databases. For the LIVE challenge, because the database was designed for the NR-IQA, the FR is not evaluated. As it can be seen, the overall performance is lower than those using the other databases due to broader types of distortion added to the reference images. Nevertheless, it is noted that DIQA still yields higher performance on the LIVE challenge database.

Secondly, we compared 2D VQA methods using three VQA databases: LIVE VQA, CSIQ VQA, and IVP VQA. Table 5 tabulates the SROCC and PLCC of the compared FR/NR-VQA methods. We benchmarked eight existing methods including five FR-VQA approaches: PSNR, SSIM [5], STMAD [57], ViS3 [58], DeepVQA [59], and three NR-VQA methods: V-BLINDS [60], MOVIE [61], and SACONVA [62].

In our experiment, as shown in Table 5, the highest SROCC and PLCC of overall distortion types are achieved by DeepVQA which takes full advantage of deep-learning and reference information in all the databases. Also, SACONVA achieves competitive performance even though it is an NR-based model. Since the metrics of PSNR and SSIM were designed for FR-IQA, it shows lower performance over all the databases. However, the other recent FR-VQA algorithms show improved performances. Interestingly, V-BLINDS shows higher performance than non-deep-learning-based FR-VQA works even it is an NR-VQA approach. Overall, it can be concluded that the deep-learning-based model can demonstrate powerful performance in correlation with the subjective scores.

Major visual preference assessment benchmarking

To verify the performance of the sharpness IQA methods, we used the CSI database [63]. For the contrast IQA, the CID2013 and CCID2014 databases were used. For image sharpness IQA, the following methods were benchmarked: Marziliano *et al.* [66], Narvekar *et al.* [67], Ferzli *et al.* [68], Caviedes *et al.* [69], and Oh *et al.* [32]. Also, we compared the several contrast IQA methods: FSIM [48], PCQI [70], RIQMC [6], FANG [71], and BIQME [73].

Table 6 shows the results for the sharpness IQA methods, where the performance of Oh *et al.* [32] is superior to the

Table 5. SROCC and PLCC comparison on the three 2D VQA databases. *Italics* indicate the deep-learning-based methods.

		LIVE VQA		CSIQ VQA		IVP VQA	
Type	Method	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
FR	PSNR	0.750	0.696	0.714	0.704	0.805	0.789
	SSIM [5]	0.788	0.721	0.763	0.762	0.661	0.646
	STMAD [57]	0.878	0.830	0.825	0.822	0.881	0.886
	ViS3 [58]	0.825	0.816	0.810	0.803	0.831	0.831
	DeepVQA [59]	0.895	0.915	0.913	0.912	–	–
NR	V-BLINDS [60]	0.843	0.832	0.849	0.859	0.848	0.832
	MOVIE [61]	0.811	0.790	0.789	0.811	–	–
	SACONVA [62]	0.871	0.857	0.867	0.864	0.886	0.870

Table 6. PLCC and SROCC comparison of sharpness IQA on the CSI database.

Method	SROCC	PLCC
Marziliano [66]	0.194	0.261
Narvekar [67]	0.366	0.505
Ferzli [68]	0.508	0.486
Caviedes [69]	0.526	0.216
Oh [32]	0.732	0.773

Table 7. Performance comparison of contrast IQA methods on the two contrast IQA databases.

	CID2013		CCID2014	
Method	SROCC	PLCC	SROCC	PLCC
FSIM [48]	0.849	0.857	0.766	0.820
PCQI [70]	0.926	0.925	0.875	0.889
RIQMC [6]	0.900	0.900	0.847	0.873
FANG [71]	0.801	0.790	0.782	0.789
BIQME [73]	0.902	0.900	0.831	0.859

other methods. The scene classification method by object and camera movement plays an important role in improving the performance of image sharpness IQA [32].

Table 7 shows the results for the contrast IQA methods. The contrast quality dedicated models [6,70,73] are superior to the conventional IQA metric FSIM [48]. Mostly, PCQI and RIQMC achieved higher performance than the other methods. Interestingly, BIQME which is an NR-IQA model shows competitive performance, and even higher than RIQMC in the CID2013 database.

IV. QOE ON STEREOSCOPIC 3D DISPLAY

A) QoE trend on S3D display

Different from 2D display, S3D display enables to accommodate another dimension of QoE by providing enhanced sense of reality through depth provision to viewers. However, as the side effect, entertaining S3D contents make viewers perceive the optical illusion effect induced from disparity of left and right images. This effect causes abnormal interaction of the oculomotor and crystalline lens control system, which results in feeling of discomfort to viewers [74–76]. Therefore, VDP needed to assess the level of fatigue has emerged as an important topic in addition to the quality assessment as done for 2D contents. As a result, several researches have been conducted to predict quality and visual discomfort more objectively for S3D contents.

Besides, there are several studies which investigated HVS-related factors observed when watching S3D contents. Since QoE can be measured after displaying S3D content over the display rather than by analyzing just content itself as done for 2D I/VQA, the performance of the QoE prediction model can be dramatically improved by considering these QoE-related factors. One of them is understanding of the depth perception-related mechanism including the process of receiving visual information through the HVS. The following items introduce major QoE-related studies: accommodation–vergence mismatch and binocular rivalry and suppression.

- *Accommodation–vergence mismatch*: This mismatch occurs due to the side effect of optical illusion. When watching S3D contents, the accommodative stimulus is fixed on the stereoscopic screen while the vergence stimulus fluctuates according to disparity over the screen as shown in Fig. 3(a). These artificial decoupling phenomenon has been commonly known as an important factor of visual discomfort [77,78].
- *Binocular rivalry and suppression*: The binocular rivalry occurs when the left and right images are mismatched, e.g., when one of the images is severely distorted, the stereopsis by both eyes is imperfectly established. In this case, the image information recognized at the same retinal location is different from our expectation used to be in daily life. The failure of binocular matching induces the binocular rivalry, which occurs in various forms, such as a sense of failed fusion or bi-state alternation between the eyes [79] as shown in Fig. 3(b). The binocular suppression is a special case of the binocular rivalry, in which rivalrous fluctuations do not occur between two images. In other words, only one image is perceived in the brain while viewing mismatched stereo stimuli, as shown in Fig. 3(b) [80].

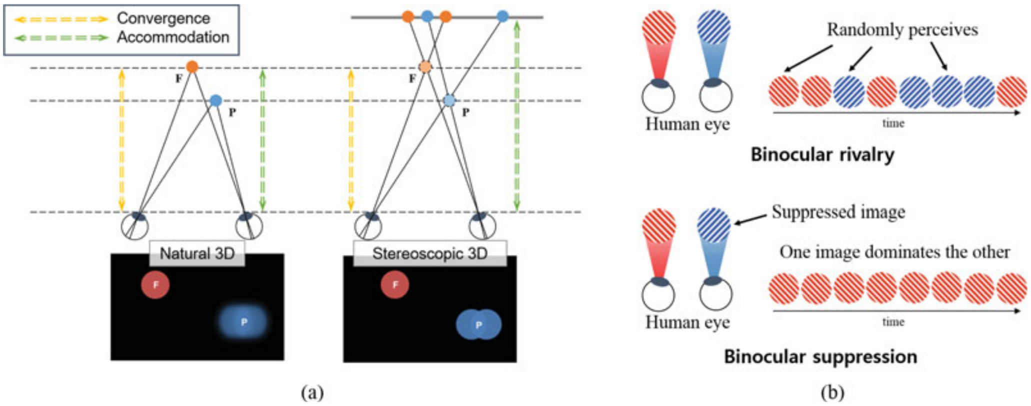

Fig. 3. Representation of related studies for S3D QoE assessment. (a) Accommodation–vergence mismatch, and (b) binocular rivalry and suppression.

B) QoE tasks on S3D display

1) Visual discomfort assessment

Visual discomfort databases

The databases for VDP consist of the left and right images and the corresponding subjective visual discomfort scores. The IEEE-SA database [81,82] and IVY LAB database [83] are the representative public databases for the performance comparison of VDP models. The IEEE-SA database [81,82] consists of 800 S3D image pairs with a resolution of 1920 × 1080. The images of the IEEE-SA database are classified into eight categories (e.g., indoor/outdoor, no salient/salient and large/small) according to the distribution of depth. The IVY LAB database [83] consists of 120 S3D image pairs with a resolution of 1920 × 1080, and includes indoor and outdoor scenes which contain various objects (e.g., humans, trees, buildings, etc.).

Major S3D VDP approaches

In previous studies, various visual factors have been found including crosstalk, keystone effects, window violations, and optical distortion, which mainly cause visual discomfort when viewing S3D contents [84–86]. However, neuronal and oculomotor conflicts arising from the accommodation–vergence mismatch are known to be the most important cue of visual discomfort [17,19,75].

Early VDP models focused on extracting statistical features related to distribution of disparity, such as mean and variance [87–90]. More recently, an advanced VDP model based on human visual perception has been proposed. Jung *et al.* [83] developed a saliency-based VDP model based on saliency-weighted disparity and disparity gradient features. Park *et al.* [18] developed a VDP model in terms of retinal resolving power and optics.

In addition to spatial information, S3D videos additionally contain temporal information over the frame sequence. Therefore, it is necessary to take into account various temporal characteristics when designing the VDP model for S3D video. The authors of [91] devised a metric for predicting discomfort from motion magnitude in the salient region under the assumption that the motion information is an important factor for visual discomfort. Lambooij *et al.* [25] classified the motions in S3D video into static/planar/in-depth motions according to depth variation and motion size. They extracted a feature vector to predict visual discomfort through the regression process.

In recent years, some researchers attempted to apply deep learning to the VDP task. However, deep-learning approaches have shown a limited performance improvement due to the insufficient databases for visual discomfort. To resolve this problem, Oh *et al.* [92] used S3D images in patch unit, and employed the existing VDP model [19] to derive the proxy ground-truth score for each patch. The authors of [20] showed the state-of-the-art performance by proposing a binocular fusion network mimicking binocular fusion model of human.

2) S3D image quality assessment

S3D I/VQA databases

The databases of the S3D I/VQA include both of left and right images and their corresponding subjective quality scores. There have been four official S3D IQA databases: LIVE S3D IQA (Phase I) [93], LIVE S3D IQA (Phase II) [94], IVC S3D IQA (Phase I) [95], and IVC S3D IQA (Phase II) [95], and three S3D VQA databases: EPFL S3D VQA [96], IRCCYN S3D VQA [97], and the QI S3D VQA [98].

Table 8 tabulates the comparison of major S3D IQA databases. The LIVE S3D IQA database (Phase I) includes 20 reference image pairs and 365 distorted image pairs including five distortion types: JPEG, JPEG2000 (JP2K), additive white Gaussian noise (AWGN), Rayleigh fast-fading channel distortion (FF), and Gaussian blur (BLUR). The LIVE S3D IQA database (Phase II) [94] contains eight reference image pairs and 360 distorted image pairs with five corresponding distortion types: JPEG, JP2K, AWGN, BLUR, and FF. The IVC S3D IQA database (Phase I) [95] has six reference image pairs and 72 distorted image pairs including three distortion types: AWGN, BLUR, and JPEG. The IVC S3D IQA database (Phase II) [95] consists of 10 reference image pairs and 120 distorted image pairs with three distortion types: AWGN, BLUR, and JPEG.

For the S3D VQA databases, Table 9 tabulates the comparison of major S3D VQA databases. The EPFL S3D VQA database [96] consists of six reference pairs and 24 distorted

Table 8. Comparison of the stereoscopic S3D IQA databases.

Database	Ref.	Dist.	Dist. types	Res.	Score type
LIVE IQA (phase I) [93]	20	385	5	640 × 360	DMOS
LIVE IQA (phase II) [94]	8	368	5	640 × 360	DMOS
IVC IQA (phase I) [95]	6	78	3	1390 × 1080	MOS
IVC IQA (phase II) [95]	10	130	3	1920 × 1080	MOS

Table 9. Comparison of the stereoscopic S3D VQA databases.

Database	Ref.	Dist.	Dist. types	Res.	Frame rate	Score type
EPFL VQA [96]	6	30	3	1920 × 1080	25	MOS
IRCCYN VQA [97]	10	20	5	1920 × 1080	25	MOS
QI VQA [98]	9	459	2	1680 × 1050	25	MOS

video pairs including three distortion types: geometrical alignment, temporal alignment, and color adjustment. The IRCCYN S3D VQA database [97] has 10 reference video pairs with 10 distorted video pairs including five distortion types: H.264, JP2K, down-sampling, sharpening, and down-sampling and sharpening. The QI S3D VQA database [98] includes nine reference pairs and 450 distorted pairs containing two distortion types: H.264 and Gaussian blur.

Major S3D IQA methods

Early studies for FR S3D IQA stemmed from approaches used for 2D IQA. In this way, a 2D FR-IQA metric was applied to the left and right images, and additional metrics were used to derive the final predictive score [99–101]. More recently, various methods have been introduced to reflect the binocular vision into the IQA model. Chen *et al.* [94] proposed a cyclopean model based on the binocular rivalry theory when the human eyes recognize stereoscopic images. In other words, the authors used a linear additive model to calculate the cyclopean images and applied a 2D IQA method for them to predict the quality scores. Lin and Wu [102] proposed a FR-S3D IQA model which includes the neural processing occurring at the visual cortex. Also, the authors of [103] proposed a local quality pooling method that calculates the quality score by dividing the distorted image pair into binocular fusion, rivalry, and suppression regions.

Some researchers have studied S3D IQA metrics without the use of reference image pairs. Sazzad *et al.* [104] proposed a distortion-specific approach which is applicable only to JPEG distortion. After extracting the edge and relative depth information, they modeled the NR S3D IQA metric using a logistic regression function. Chen *et al.* [105] extracted normalized texture information, disparity, and uncertainty maps under the assumption that natural S3D images show statistical regularity, and fitted them to a generalized Gaussian and log-normal distribution. The shape parameters were utilized as features to predict the MOS score through support vector regression.

Along with the development of deep learning, CNN-based methods have been studied in the S3D IQA domain. Zhang *et al.* [106] proposed a CNN model to predict the quality score by using left, right, and difference images as an input to the deep-learning model. Each input (left, right, and difference images) passed through different convolution layers whose weights were shared, and the resulting structural features were concatenated to predict the quality score through the MLP layer. Oh *et al.* [14] presented a two-stage deep-learning framework to solve the database shortage problem when applying deep learning to IQA. Also, Ding *et al.* [107] proposed a CNN network that imitates the process of depth recognition in HVS.

Major S3D VQA methods

In previous S3D VQA researches, disparity and spatio-temporal information have been importantly considered. Han *et al.* [108] extracted spatial-temporal structural information and derived the SSIM-like similarity between adjacent frames. The authors of [97] applied sensitivity and luminance masking to generate videos perceived by the HVS, and calculated the MSE between the reference and distorted videos. Recently, the authors of [109] proposed a video quality evaluation method by extracting spatial information from color and depth maps. More recently, in [110], they proposed a depth perception-based VQA metric. To model a metric closer to the HVS, the authors of [111] proposed a just-noticeable-difference model that can be applied to stereoscopic content, and predicted video quality by using a saliency map as its weighting function. For the learning-based approach, in [112], the authors introduced the features of auto-regressive prediction-based disparity measurement. Jiang *et al.* [113] performed tensor decomposition on S3D video to extract motion features representing time-varying information. In recent years, there have been attempts to use CNN for S3D VQA. Yang *et al.* [114] proposed a 3D CNN framework based on local and global spatiotemporal information.

3) Benchmarking on 3D QoE tasks

For 3D QoE, the benchmark is performed by evaluating the performance using the correlation indices of PLCC and SROCC.

Major S3D VDP benchmarking

To compare the performance, the following S3D VDP models are included:Yano *et al.* [87], Nojiri *et al.* [88], Choi

Table 10. Performance comparison of VDP models on the two S3D VDP databases. *Italics* indicate the deep-learning-based methods.

	IEEE-SA		IVY LAB	
Method	SROCC	PLCC	SROCC	PLCC
Yano [87]	0.403	0.336	0.411	0.346
Nojiri [88]	0.694	0.606	0.703	0.613
Choi [89]	0.672	0.587	0.682	0.598
Kim [90]	0.704	0.617	0.711	0.625
Park [18]	0.852	0.779	0.862	0.781
Oh [92]	0.885	0.816	0.862	0.781
Kim [20]	0.904	0.843	–	–

Table 11. Performance comparison for S3D IQA models on the two S3D IQA databases. *Italics* indicate the deep-learning-based methods.

		LIVE (phase I)		LIVE (phase II)	
Type	Method	SROCC	PLCC	SROCC	PLCC
FR	Chen [94]	0.928	0.883	0.836	0.823
	You [100]	0.805	0.803	0.719	0.731
	Benoit [99]	0.858	0.863	0.770	0.769
	Lin [102]	0.856	0.784	0.638	0.642
NR	Sazzad [104]	0.618	0.624	0.648	0.669
	Chen [105]	0.890	0.881	0.864	0.854
	Zhang [106]	0.943	0.947	0.708	0.763
	Oh [14]	0.935	0.943	0.871	0.863
	Ding [107]	0.942	0.940	0.924	0.930

Table 12. Performance comparison for S3D VQA models on the two S3D VQA databases. *Italics* indicate the deep-learning-based methods.

	IRCCYN		QI	
Method	SROCC	PLCC	SROCC	PLCC
Feng [111]	0.650	0.623	0.842	0.838
PHVS-3D [97]	0.548	0.515	0.708	0.817
SFD [98]	0.597	0.590	0.648	0.663
3D-STS [108]	0.642	0.612	0.648	0.663
MNSVQM [113]	0.855	0.839	0.882	0.857
BSVQE [112]	0.924	0.909	0.939	0.939
Yang [114]	0.948	0.923	0.950	0.942

et al. [89], Kim *et al.* [90], Park *et al.* [18], Oh *et al.* [92], and Kim *et al.* [20]. The IEEE-SA database [81,82] and IVY LAB database [83] were used for training and testing. As shown in Table 10, the methods using HVS-related features [18,90] perform better than those using only statistical features from disparity maps [87–89]. It means that the analysis of 3D factors such as accommodation–vergence mismatch and binocular rivalry is important to predict the visual discomfort. In particular, the deep-learning approaches [20,92] show superior performance compared to the other methods [18,87–90]. This means there is a limitation to cover various factors optimally by existing hand-crafted methods. It also shows that the data augmentation process through the patch-based approach has successfully solved the database shortage problem as experienced in the 2D-IQA approaches.

Major S3D I/VQA benchmarking

To compare the performance of existing S3D IQA models, we used 10 S3D IQA methods: Chen [94], You [100], Benoit [99], Lin [102], Sazzad [104], Chen [105], Zhang [106], Oh [14], Ding [107]. For the training and testing processes, the LIVE S3D IQA database (phase I and II) was used. As shown in Table 11, the FR methods show higher performance than the NR methods since reference images can be utilized as additional information for the FR-S3D IQA. Also, the deep-learning approaches [106,107] outperform the other works [94,105]. This means that the deep-learning models have successfully learned various visual perception characteristics compared to conventional hand-crafted methods. Oh *et al.* [14] proposed a concept of pseudo ground-truth for each S3D patch to overcome overfitting problem. Also, Ding et al [107] considered various HVS factors such as saliency and multiscale disparity map, which produced better performance than previous works.

Secondly, we compared the performance of existing S3D VQA models: Feng [111], PHVS-3D [97], SFD [98], 3D-STS [108], MNSVQM [113], BSVQE [112], and Yang [114]. The IRCCYN 3D video quality database [97] and the Qi stereoscopic video quality database [98] were used for performance comparison. As can be seen in Table 12, the performance of S3D VQA is better when they are trained on the QI database. Because the QI database has fewer distortion types and more training videos, the VQA models can easily learn the distortion features of videos. Also, the machine-learning-based methods [112–114] show higher performance than the metric-based methods [97,98,108,111] which rely on only mathematical equations. This means that the metric-based methods have a limitation to reflect the characteristics of distorted S3D videos. In particular, the deep-learning approach shows higher performance than the other methods while learning the spatial and temporal characteristics of videos more successfully [114].

V. QOE ON HMD DEVICE

A) QoE trend on HMD device

360-degree VR content

360-degree VR content is a new type of visual information that brings users totally immersive experience. Different from usual 2D content displayed on a normal plane, 360-degree content surrounds user spherically so that the content is in his view at any head pose. Equipped with an HMD device, user can enjoy the content using head motion similar to what we do in real life, which provides the immersive and interactive experience. This is a new type of experience so that several new human factors should be identified to assess the quality of the immersive experience. There are many factors that determine the immersive experience of 360-degree content such as resolution, saliency, bitrate, and visual quality. Unfortunately, compared to the 2D counterpart, there is no widely-accepted I/VQA workflow for 360-degree contents. Simply applying 2D IQA metrics to the assessment is also troublesome as these metrics do not consider the spherical nature of 360-degree contents.

In [12,115], Lee et al. first introduced a metric of assessing the image over curvilinear coordinates from Cartesian coordinates after mapping it in accordance with the foveation. Stemming from this concept, researches have been conducted to map the content from Spherical coordinates to Cartesian coordinates, and then applied conventional metrics to assess the QoE. However, this warping may cause some projection errors into the scene, and results in much redundancy, which leads to a decrease in accuracy of the IQA. Thus, reliable I/VQA pipelines have been studied for the continuous integration of the 360-degree contents into our life.

Computer graphic-based VR content

Unlike the 360-degree VR content, the breakthrough of CG-based VR content has led the user to interact in the virtual space. Furthermore, development tools such as Unity and Unreal have facilitated the diversified virtual experiences through a huge amount of CG contents. Accordingly, the user's demand to access sufficient QoE has accelerated the production of CG content in a more realistic way. However, in the CG-based VR contents, heterogeneous visual stimuli through the HMD strongly induces physiological side effects called VR sickness (or cybersickness). With this unexpected symptoms, it has been critical to predict the VR sickness level in order to guarantee users' viewing safety and abundant QoE. In literature, many VR sickness-related human factors have been found such as motion-to-photon latency, flicker, and visual movement pattern. Nevertheless, there is still no definitive conclusion to predict VR sickness due to its complex perceptual mechanism.

To overcome this problem, a few VRSA approaches have focused on analyzing changes in the physiological condition during VR experience, and on analyzing meaningful feature information through statistical analysis. Electrogastrogram (ECG), eye blink, heart period, electroencephalogram (EEG), and galvanic skin response (GSR) are the main measurement tools for observing physiological changes.

B) QoE tasks on HMD device

1) 360-degree I/VQA

360-degree I/VQA databases

There are several notable databases in 360-degree VR content I/VQA literature. One of them for 360-degree VR content IQA is CVIQD2018 [116]. The database consists of 16 pristine images and 544 distorted derivations compressed by JPEG, H.264, and H.265. The resolution of these images is 4096 × 2048. The OIQA [117] also contains 16 pristine images and 336 distorted images which are perturbed by JPEG and JPEG2000 compressions, Gaussian blur, and white Gaussian noise. The resolutions vary from 11332 × 5666 to 13320 × 6660.

For the VQA task, VQA-ODV [118] is the largest one set up to date. The set has 60 pristine sequences and another 600 videos distorted by H.265 compression noise. The resolution ranges from 3840 × 1920 to 7680 × 3840. The second-largest dataset is much smaller [119]. It has only 16 original sequences and 400 distorted videos derived from them. However, there are many more distortion types including VP9, H.264, H.265, Gaussian blur, and box-blur. The resolution is fixed to 4096 × 2048.

Major 360-degree I/VQA approaches

After 360-degree contents are converted to 2D contents by means of projection, and then the 2D QA metrics can be applied as done in some early works [120]. However, this raises a serious problem due to the warping in projection. Depending on the projection, some regions in the spherical 360-degree contents are redundantly projected onto the 2D plane, which significantly affects the QA scores. Also, the QA metrics are not invariant to projection and sampling, which does not agree with our HVS.

In the literature, there exist three main approaches to the I/VQA problem; the sampling, attention, and learning-based methods.

Sampling-based I/VQA methods: To deal with the redundancy and over-sampling problem, in [121], the authors introduced the concept for 360-degree spherical PSNR (s-PSNR). s-PSNR measures the PSNR between points in the spherical domain instead of the usual planar image domain. Points on the sphere were obtained by either nearest neighbor selection or interpolation from the planar image. PSNR can also be calculated in a weighting manner. The weighted to spherically uniform PSNR (WS-PSNR) [122] simply calculates the PSNR between two images but the pixels are weighted differently depending on the sampling area on the corresponding spherical regions. In [123], the authors converted the images into the Crasters Parabolic Projection format, which is similar to the unit sphere, and then applied PSNR, which is called CPP-PSNR, to the transformed images. CPP-PSNR can be applicable for images with different resolutions and projection types. Indeed, this type of sampling-based I/VQA had been researched earlier [12,115], where foveal-PSNR has been presented by reflecting the foveation over the Cartesian coordinates.

Saliency-based I/VQA methods: Many studies resort to saliency for the weighting scheme in different ways. In [124], the authors proposed a saliency detection model and used it to weight the PSNR score. In [125], the authors utilized random forest to predict the 360 attention, and then masked the non-content region. Differently, Yu *et al.* [126] calculated PSNR in the attentive viewports only while Ozcinar *et al.* [127] used ground-truth saliency maps directly. The F-PSNR can be directly applied to this saliency-based scheme as long as the saliency region is identified [12,115]. Thus, it can be stated that both sampling and saliency-based methods stem from the idea published in [12,115].

Learning-based I/VQA methods: Given the popularity of deep learning, recent works also manage to benefit from the phenomenal performance of CNN. Kim *et al.* [128] proposed to extract patch-based positional and visual features from a 360 image using CNN, and then regressed the features onto the ground truth MOS. Similarly, Lim *et al.* [129] proposed latent spatial and positional features, and used adversarial learning to predict the quality score. In [118], the authors predicted head movement and eye movement, and incorporated this information into the I/VQA model of a deep CNN. Li *et al.* [130] proposed the viewport CNN (V-CNN) and a two-stage training scheme to extract proposal viewports, and rated the quality of them. The final score was obtained by integrating the scores of the viewports.

A quantitative benchmark of some spotlight methods is shown in Table 13. The first three rows are bottom-up methods derived from PSNR while the others are all deep-learning-based models. s-PSNR is the best among these metrics. DeepQA [50] is a state-of-the-art deep-learning-based method designed for the 2D IQA task. Interestingly, it is easy to notice that it shows lower performance than those in general IQA task. This emphasizes the difference in characteristics of 2D and 360 stimuli. V-CNN [130] achieves the current state-of-the-art correlation with human opinions. However, compared to the 2D IQA score, the performance is still much behind, which suggests there is still much room for improvement in future work.

Table 13. Performance comparison of major 360-degree VQA models on the VQA-ODV database [118]. Results are reproduced from [130].

Method	PLCC	SROCC
s-PSNR [121]	0.693	0.698
WS-PSNR [122]	0.672	0.684
CPP-PSNR [123]	0.681	0.690
DeepQA [50]	0.694	0.730
Li *et al.* [118]	0.782	0.795
V-CNN [130]	0.874	0.896

2) VR sickness assessment

VR sickness databases

According to our knowledge, there is no public VR sickness database for CG-based VR contents. However, several self-produced databases have been introduced in the literature. Padmanban *et al.* produced a dataset of stereoscopic CG videos and their corresponding subjective sickness scores to quantify their nauseogenicity [131]. The dataset consists of 19 stereoscopic videos extracted with a 60-second clip with the corresponding Kennedy SSQ and the MSSQ-short. Kim *et al.* constructed the VR sickness database and conducted a subjective sickness scoring using the Unity engine [21]. The database includes two reference scenes which are then extended by adjusting parameters: object movements, camera movements, and content components. In addition, it serves the user's head-movement information from Gyroscope sensors in the HMD.

The largest dataset for CG-based VR sickness is perhaps the dataset of [132]. Kim *et al.* constructed a total of 52 scenes by composing various contents in different scenarios, and conducted a subjective test on 200 non-expert users, which is named the ETRI-VR dataset. Furthermore, it contains three types of physiological signals including EEG, ECG, and GSR. For the collection of the dataset, they used the HTC-VIVE, which provides a resolution of 1080 × 1022 per eye, with a refresh rate of 90 Hz and a nominal 110-degree vertical field of view.

Major VRSA approaches

Feature-based VRSA: Recently, it has been found that the VR sickness can be triggered by content factors such as fidelity, global (local) motion, and depth change. Based on the fact that motion is a dominant component that induces sickness, Padmanaban *et al.* developed the first machine-learning-based VRSA model [131]. Here, optical flow is utilized to extract motion representation, and the features are regressed onto the subjective sickness score using the SVR.
Learning-based VRSA: Different from the feature engineering perspective, CNN-based deep-learning methods have been proposed. Kim *et al.* designed a sickness prediction model that utilizes an unsupervised learning manner [20]. To learn the exceptional motion in the content, they defined unexpected motion as the difference between the image input and reconstructed output from the autoencoder. Furthermore, Kim *et al.* proposed the model that reflects individual differences for VRSA by complementary learning of the visual and physiological features [132]. They devised a novel deep learning framework to identify the human cognitive feature space by analyzing brain activity, and then expressed the visual and cognitive features simultaneously in the intermediate state.

Table 14. Performance comparison of major VRSA models on the ETRI-VR database.

Method	PLCC	SROCC
Padmanban [131]	0.532	0.577
Kim [21]	0.621	0.581
Kim [132]	0.773	0.780

In this review, we report benchmarking of major VRSA models on the ETRI-VR dataset. Table 14 shows the performance comparison of existing methods. The first two rows are top-down approaches that compute optical-flow features while the others are CNN-based models. Among the benchmarks, the deep-learning-based model shows the best prediction performance. Moreover, it can be seen that Kim [132] shows higher performance than Kim [20]. Therefore, it is more effective to take into account both of visual and cognitive space in predicting VR sickness compared to the simple motion estimation-based method.

VI. FUTURE TRENDS ON QOE

A) QoE on future displays

On the basis of the recent trend, display technology has been growing for user to entertain high-quality contents. Advances in these technologies continue to raise user satisfaction with larger screens and even sophisticated user interaction. From this ongoing technological development, it can be easily inferred that the resolution of 2D display is getting higher, the S3D display combines a sense of depth on 2D space, and the HMD device is expanded for more interactive experience. As if the sprout comes out of the ground, this has been continuing to evolve toward areas where user is deeply immersive to media just like reality rather than experience. Therefore, it is expected that the future displays, such as AR, 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 with 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.

B) Deep-learning approaches

As aforementioned, the deep CNN has emerged as a core technology while breaking most performance records in the area of QoE via intensive training in accordance with the dataset. Accordingly, there have been many attempts on the deep-learning technique to find out new factors without using prior information. For instance, in recent I/VQA works, the human visual sensitivity has been successfully investigated from the output of hidden nodes attained from the deep-learning mechanism, which enables to provide deep insight on how human perception is responsive to an input image [50,56,59,133].

How can this technology be more generally applied to future QoE metrics? Various deep-learning-based QoE applications could fall in the area of image/video, S3D, and VR contents. 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. For example, in VRSA, the distribution of the spatial texture has a great effect on the motion perception of the HVS. Thereby, it is expected that the motion component of an image and the weighting process of the human visual mechanism extracted from the deep model can be effectively applied to calculate the visually perceived QoE.

VII. CONCLUSION

In this paper, we have examined the QoE assessment and classification of existing displays (ie, 2D display, S3D display, and HMD device) from a comprehensive viewpoint. In addition, QoE assessment approaches that have been utilized in each display was introduced and benchmarked according to QoE types and applications. Based on this, it can be concluded that the objective QoE assessment has been played an influential role in the human satisfaction of the display.

In the future, we expect that a lot of QoE work will be actively accomplished in accordance with display type (i.e., AR, holographic display, and light-field display). Notably, content quality involved in human interaction will emerge as a new paradigm of QoE challenging issues. Toward this, the current valuable insights of image-processing techniques based on data-driven approach will play an important roll in the future.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Brunnström K. *et al.*: QualiNet white paper on definitions of quality of experience (2013).

[2] Laghari K.U.R.; Connelly K.: Toward total quality of experience: a QoE model in a communication ecosystem. *IEEE Commun. Mag.*, **50** (4) (2012), 58–65.

[3] Union I.T.: Recommendation ITU-R BT. 500-13: methodology for the subjective assessment of the quality of television pictures (2012)

[4] Subjective Video Quality Assessment Methods for Multimedia Applications. Recommendation ITU-T P.910 (Sep. 1999).

[5] Wang Z. *et al.*: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, **13** (4) (2004), 600–612.

[6] Gu K. *et al.*: The analysis of image contrast: from quality assessment to automatic enhancement. *IEEE Trans. Cybernet.*, **46** (1) (2015), 284–297.

[7] Kim H. *et al.*: Visual preference assessment on ultra-high-definition images. *IEEE Trans. Broadcast.*, **62** (4) (2016), 757–769.

[8] Kim H. *et al.*: Saliency prediction on stereoscopic videos. *IEEE Trans. Image Process.*, **23** (4) (2014), 1476–1490.

[9] Kim H.; Lee S.: Transition of visual attention assessment in stereoscopic images with evaluation of subjective visual quality and discomfort. *IEEE Trans. Multimedia*, **17** (12) (2015), 2198–2209.

[10] Ahn S. *et al.*: Visual attention analysis on stereoscopic images for subjective discomfort evaluation, in *2016 IEEE Int. Conf. on Multimedia and Expo*, IEEE, 2016.

[11] Nguyen A.-D. *et al.*: Deep visual saliency on stereoscopic images. *IEEE Trans. Image Process.*, **28** (4) (2018), 1939–1953.

[12] Lee S. *et al.*: Foveated video compression with optimal rate control. *IEEE Trans. Image Process.*, **10** (7) (2001), 977–992.

[13] Lee S. *et al.*: Foveated video quality assessment. *IEEE Trans. Multimedia*, **4** (1) (2002), 129–132.

[14] Oh H. *et al.*: Blind deep S3D image quality evaluation via local to global feature aggregation. *IEEE Trans. Image Process.*, **26** (10) (2017), 4923–4936.

[15] Oh H. *et al.*: Stereoscopic 3D visual discomfort prediction: a dynamic accommodation and vergence interaction model. *IEEE Trans. Image Process.*, **25** (2) (2015), 615–629.

[16] Oh H. *et al.*: Enhancement of visual comfort and sense of presence on stereoscopic 3d images. *IEEE Trans. Image Process.*, **26** (8) (2017), 3789–3801.

[17] Kim T. *et al.*: Transfer function model of physiological mechanisms underlying temporal visual discomfort experienced when viewing stereoscopic 3D images. *IEEE Trans. Image Process.*, **24** (11) (2015), 4335–4347.

[18] Park J. *et al.*: 3D visual discomfort prediction: vergence, foveation, and the physiological optics of accommodation. *IEEE J. Sel. Topics Signal Process.*, **8** (3) (2014), 415–427.

[19] Park J. *et al.*: 3D visual discomfort predictor: analysis of disparity and neural activity statistics. *IEEE Trans. Image Process.*, **24** (3) (2014), 1101–1114.

[20] Kim H.G. *et al.*: Binocular fusion net: deep learning visual comfort assessment for stereoscopic 3D. *IEEE Trans. Circuits Syst. Video Technol.*, **29** (4) (2018), 956–967.

[21] Kim J. *et al.*: Virtual reality sickness predictor: analysis of visual-vestibular conflict and VR contents, in *2018 Tenth Int. Conf.on Quality of Multimedia Experience (QoMEX). IEEE*, 2018.

[22] Liu T.-J. *et al.*: Visual quality assessment: recent developments, coding applications and future trends, in *APSIPA Transactions on Signal and Information Processing*, 2013.

[23] Chikkerur S. *et al.*: Objective video quality assessment methods: a classification, review, and performance comparison. *IEEE Trans. Broadcast.*, **57** (2) (2011), 165–182.

[24] IEEE P3333.1. (2012). Standard for the quality assessment of three dimensional displays, 3D contents and 3D devices based on human factors. IEEE Standards Association.

[25] Lambooij M. *et al.*: Visual discomfort of 3D TV: assessment methods and modeling. *Displays*, **32** (4) (2011), 209–218.

[26] Kim T. *et al.*: Multimodal interactive continuous scoring of subjective 3D video quality of experience. *IEEE Trans. Multimedia*, **16** (2) (2013), 387–402.

[27] Park J. *et al.*: Video quality pooling adaptive to perceptual distortion severity. *IEEE Trans. Image Process.*, **22** (2) (2013), 610–620.

[28] VQEG: final report from the video quality experts group on the validation of objective models of video quality assessment, Phase II, in *VQEG*, Boulder, CO, USA, Rep, 2003.

[29] Nguyen A.-D. *et al.*: A simple way of multimodal and arbitrary style transfer, in *IEEE Int. Confe. on Acoustics, Speech and Signal Processing*. IEEE, 2019.

[30] Sugawara M. *et al.*: Research on human factors in ultrahigh-definition television (UHDTV) to determine its specifications. *SMPTE Motion Imag. J.*, **117** (3) (2008), 23–29.

[31] Kim W. *et al.*: No-reference perceptual sharpness assessment for ultra-high-definition images, in *IEEE Inte. Conf. on Image Processing*. IEEE, 2016.

[32] Oh T. *et al.*: No-reference sharpness assessment of camera-shaken images by analysis of spectral structure. *IEEE Trans. Image Process.*, **23** (12) (2014), 5428–5439.

[33] Kim H. *et al.*: Blind sharpness prediction for ultrahigh-definition video based on human visual resolution. *IEEE Trans. Circuits Syst. Video Technol.*, **27** (5) (2016), 951–964.

[34] Oh H.; Lee S.: Visual presence: Viewing geometry visual information of UHD S3D entertainment. *IEEE Trans. Image Process.*, **25** (7) (2016), 3358–3371.

[35] Kim J. *et al.*: Video sharpness prediction based on motion blur analysis, in *IEEE Int. Confe. on Multimedia and Expo*. IEEE, 2015.

[36] Hu S. *et al.*: Objective video quality assessment based on perceptually weighted mean squared error. *IEEE Trans. Circuits Syst. Video Technol.*, **27** (9) (2016), 1844–1855.

[37] Sheikh H. *et al.*: A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.*, **15** (11) (2006), 3440–3451.

[38] Ponomarenko N. *et al.*: TID2008-a database for evaluation of full-reference visual quality assessment metrics. *Adv. Modern Radioelectronics*, **10** (4) (2009), 30–45.

[39] Larson E.C.; Chandler D.M.: Most apparent distortion: full-reference image quality assessment and the role of strategy. *J. Electron. Imaging*, **19** (1) (2010), 19–21.

[40] Jayaraman D. *et al.*: Objective quality assessment of multiply distorted images, in *Proc. Asilomar Conf. Signals, Systems, and Computers*, 2012, 1693–1697

[41] Ponomarenko N. *et al.*: Image database TID2013: peculiarities, results and perspectives. *Signal Process. Image Commun.*, **30**, (2015), 57–77.

[42] Ghadiyaram D.; Bovik A.C.: Massive online crowdsourced study of subjective and objective picture quality. *IEEE Trans. Image Process.*, **25** (1) (2016), 372–387.

[43] Seshadrinathan K. *et al.*: Study of subjective and objective quality assessment of video. *IEEE Trans. Image Process.*, **19** (6) (2010), 1427–1441.

[44] Laboratory of Computational Perception & Image Quality, Oklahoma State University, CSIQ Video Database. [Online]. Available: http://vision.okstate.edu/?loc=stmad

[45] Zhang F.; Li S.; Ma L.; Wong Y.C.; Ngan K.N. IVP Video Quality Database, 2011, [online] Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective/index.html.

[46] Lai Y.-K.; Jay Kuo C-C.: A Haar wavelet approach to compressed image quality measurement. *J. Vis. Commun. Image R.*, **11** (1) (2000), 17–40.

[47] Sheikh H.R.; Bovik A.C.: A visual information fidelity approach to video quality assessment, in *The First Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Vol. 7. SN, 2005.

[48] Zhang L. *et al.*: FSIM: a feature similarity index for image quality assessment. *IEEE Trans. Image Process.*, **20** (8) (2011), 2378–2386.

[49] Xue W. *et al.*: Gradient magnitude similarity deviation: a highly efficient perceptual image quality index. *IEEE Trans. Image Process.*, **23** (2) (2013), 684–695.

[50] Kim J.; Lee S.: Deep learning of human visual sensitivity in image quality assessment framework, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017

[51] Mittal A. *et al.*: Blind /referenceless image spatial quality evaluator, in *2011 Conf. record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*. IEEE, 2011.

[52] Ye P. *et al.*: Unsupervised feature learning framework for no-reference image quality assessment, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*. IEEE, 2012

[53] Zhang L. *et al.*: A feature-enriched completely blind image quality evaluator. *IEEE Trans. Image Process.*, **24** (8) (2015), 2579–2591.

[54] Xu J.; Ye P.; Li Q.; Du H.; Liu Y.; Doermann D.: Blind image quality assessment based on high order statistics aggregation. *IEEE Trans. Image Process.*, **25** (9) (2016), 4444–4457.

[55] Kim J.; Lee S.: Fully deep blind image quality predictor. *IEEE J. Sel. Top. Signal Process.*, **11** (1) (2016), 206–220.

[56] Kim J. *et al.*: Deep CNN-based blind image quality predictor. *IEEE Trans. Neural Netw. Learn. Syst.*, **30** (1) (2018), 11–24.

[57] Vu P.V. *et al.*: A spatiotemporal most-apparent-distortion model for video quality assessment, in *IEEE Int. Confe. on Image Processing*. IEEE, 2011.

[58] Vu P.V. *et al.*: ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electron. Imaging*, **23** (1) (2014), 013016.

[59] Kim W. *et al.*: Deep video quality assessor: from spatio-temporal visual sensitivity to a convolutional neural aggregation network, in *Proc. Eur. Conf. Comput. Vis.(ECCV)*. 2018.

[60] Saad M. *et al.*: Blind Prediction of Natural Video Quality and H. 264 Applications (2013).

[61] Seshadrinathan *et al.*: Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Trans. Image Process.*, **19** (2) (2009), 335–350.

[62] Li Y. *et al.*: No-reference video quality assessment with 3D shearlet transform and convolutional neural networks. *IEEE Trans. Circuits Syst. Video Technol.*, **26** (6) (2015), 1044–1057.

[63] Köhler R. *et al.*: Recording and playback of camera shake: benchmarking blind deconvolution with a real-world database, in *Proc. Eur. Conf. Comput. Vis.(ECCV)*, 2012.

[64] Elder J.H.; Zucker S.W.: Local scale control for edge detection and blur estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **20** (7) (1998), 699–716.

[65] Hu H.; De Haan G. Low cost robust blur estimator, in *IEEE Int. Conf. on Image Processing*. IEEE, 2006.

[66] Marziliano P. *et al.*: A no-reference perceptual blur metric, in *IEEE Int. Conf. on Image Processing*. IEEE, 2002.

[67] Narvekar N.D.; Karam L.J.: A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). *IEEE Trans. Image Process.*, **20** (9) (2011), 2678–2683.

[68] Ferzli R.; Karam L.J.: A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB). *IEEE Trans. Image Process.*, **18** (4) (2009), 717–728.

[69] Caviedes J.; Oberti F.: A new sharpness metric based on local kurtosis, edge and energy information. *Signal Process. Image Commun.*, **19** (2) (2004), 147–161.

[70] Wang S. *et al.*: A patch-structure representation method for quality assessment of contrast changed images. *IEEE Signal Process. Lett.*, **22** (12) (2015), 2387–2390.

[71] Fang Y. *et al.*: No-reference quality assessment of contrast-distorted images based on natural scene statistics. *IEEE Signal Process. Lett.*, **22** (7) (2014), 838–842.

[72] Chen Z. *et al.*: Quality assessment for comparing image enhancement algorithms, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2014.

[73] Gu K. *et al.*: Learning a no-reference quality assessment model of enhanced images with big data. *IEEE Trans. Neural Netw. Learn. Syst.*, **29** (4) (2017), 1301–1313.

[74] Emoto *et al.*: Repeated vergence adaptation causes the decline of visual functions in watching stereoscopic television. *J. Display Technol.*, **1** (2) (2005), 328.

[75] Hoffman D.M. *et al.*: Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *J. Vis.*, **8** (3) (2008), 33–33.

[76] Lee K. *et al.*: 3D visual activity assessment based on natural scene statistics. *IEEE Trans. Image Process.*, **23** (1) (2013), 450–465.

[77] Wann J.P. *et al.*: Natural problems for stereoscopic depth perception in virtual environments. *Vision research*, **35** (19) (1995), 2731–2736.

[78] Okada Y. *et al.*: Target spatial frequency determines the response to conflicting defocus-and convergence-driven accommodative stimuli. *Vision Res.*, **46** (4) (2006), 475–484.

[79] Levelt W.J.M. On binocular rivalry. Diss. Van Gorcum Assen, 1965.

[80] Blake R. *et al.*: What is suppressed during binocular rivalry? *Perception*, **9** (2) (1980), 223–231.

[81] IEEE-SA Stereo Image Database 2012 [online]. Available: http://grouper.ieee.org/groups/3dhf/

[82] IEEE P3333.1.1. Standard for the quality of experience (QoE) and visual comfort assessments of the three dimensional contents based on psychophysical studies. 2015.

[83] Jung Y.J. *et al.*: Predicting visual discomfort of stereoscopic images using human attention model. *IEEE Trans. Circuits Syst. Video Technol.*, **23** (12) (2013), 2077–2082.

[84] Sohn H. *et al.*: Predicting visual discomfort using object size and disparity information in stereoscopic images. *IEEE Trans. Broadcast.*, **59** (1) (2013), 28–37.

[85] Bando T. *et al.*: Visual fatigue caused by stereoscopic images and the search for the requirement to prevent them: a review. *Displays*, **33** (2) (2012), 76–83.

[86] Kim J. *et al.*: Quality assessment of perceptual crosstalk on two-view auto-stereoscopic displays. *IEEE Trans. Image Process.*, **26** (10) (2017), 4885–4899.

[87] Yano S. *et al.*: A study of visual fatigue and visual comfort for 3D HDTV/HDTV images. *Displays*, **23** (4) (2002), 191–201.

[88] Nojiri Y. *et al.*: Measurement of parallax distribution and its application to the analysis of visual comfort for stereoscopic HDTV, in *Stereoscopic Displays and Virtual Reality Systems X*, vol. 5006, *Int. Society for Optics and Photonics*, 2003.

[89] Choi J. *et al.*: Visual fatigue modeling and analysis for stereoscopic video. *Opt. Eng.*, **51** (1) (2012), 017206.

[90] Kim D.; Sohn K.: Visual fatigue prediction for stereoscopic image. *IEEE Trans. Circuits Syst. Video Technol.*, **21** (2) (2011), 231–236.

[91] Jung Y.J. *et al.*: Visual comfort assessment metric based on salient object motion information in stereoscopic video. *J. Electron. Imag.*, **21** (1) (2012), 011008.

[92] Oh H. *et al.*: Deep visual discomfort predictor for stereoscopic 3D images. *IEEE Trans. Image Process.*, **27** (11) (2018), 5420–5432.

[93] Moorthy A.K. *et al.*: Subjective evaluation of stereoscopic image quality. *Signal Process. Image Commun.*, **28** (8) (2013), 870–883.

[94] Chen M.-J. *et al.*: Full-reference quality assessment of stereopairs accounting for rivalry. *Signal Process. Image Commun.*, **28** (9) (2013), 1143–1155.

[95] Wang J. *et al.*: Quality prediction of asymmetrically distorted stereoscopic 3D images. *IEEE Trans. Image Process.*, **24** (11) (2015), 3400–3414.

[96] Joveluro P. *et al.*: Perceptual video quality metric for 3d video quality assessment. in *2010 3DTV-Conf.: The True Vision-Capture, Transmission and Display of 3D Video*. IEEE, 2010.

[97] Jin L. *et al.*: 3D-DCT based perceptual quality assessment of stereo video. in *IEEE Int. Conf. on Image Processing*. IEEE, 2011.

[98] Lu F. *et al.*: Quality assessment of 3D asymmetric view coding using spatial frequency dominance model, in *2009 3DTV Conf.: The True Vision-Capture, Transmission and Display of 3D Video*. IEEE, 2009.

[99] Benoit A. *et al.*: Quality assessment of stereoscopic images. *EURASIP J. Image Video Process.*, **2008** (1) (2009), 659024.

[100] You J. *et al.*: Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis, in *Proc. Int. Workshop Video Process. Quality Metrics Consum. Electron*, Vol. 9. 2010.

[101] Yang J. *et al.*: Objective quality assessment method of stereo images, in *3DTV Conf.: The True Vision-Capture, Transmission and Display of 3D Video*. IEEE, 2009.

[102] Lin Y.-H.; Wu J.-L.: Quality assessment of stereoscopic 3D image compression by binocular integration behaviors. *IEEE Trans. Image Process.*, **23** (4) (2014), 1527–1542.

[103] Lee K.; Lee S.: 3D perception based quality pooling: stereopsis, binocular rivalry, and binocular suppression. *IEEE J. Sel. Topics Signal Process.*, **9** (3) (2015), 533–545.

[104] Sazzad Z.M. *et al.*: Objective no-reference stereoscopic image quality prediction based on 2D image features and relative disparity. *Adv. Multimedia*, **2012**, (2012), 8.

[105] Chen M.-J. *et al.*: No-reference quality assessment of natural stereopairs. *IEEE Trans. Image Process.*, **22** (9) (2013), 3379–3391.

[106] Zhang W. *et al.*: Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network. *Pattern Recognit.*, **59**, (2016), 176–187.

[107] Ding Y. *et al.*: No-reference stereoscopic image quality assessment using convolutional neural network for adaptive feature extraction. *IEEE Access*, **6**, (2018), 37595–37603.

[108] Han J. *et al.*: Stereoscopic video quality assessment model based on spatial-temporal structural information, in *2012 Visual Communications and Image Processing*. IEEE, 2012.

[109] Malekmohamadi H. *et al.*: A new reduced reference metric for color plus depth 3D video. *J. Vis. Commun. Image Representation*, **25** (3) (2014), 534–541.

[110] Zhu H. *et al.*: A stereo video quality assessment method for compression distortion, in *2015 Int. Conf. on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2015.

[111] Qi F. *et al.*: Stereoscopic video quality assessment based on visual attention and just-noticeable difference models. *Signal Image Video Process.*, **10** (4) (2016), 737–744.

[112] Chen Z.; Zhou W.; Li W.: Blind stereoscopic video quality assessment: from depth perception to overall experience. *IEEE Trans. Image Process.*, **27** (2) (2017), 721–734.

[113] Jiang G. *et al.*: No reference stereo video quality assessment based on motion feature in tensor decomposition domain. *J. Vis. Commun. Image Representation*, **50**, (2018), 247–262.

[114] Yang J. *et al.*: Stereoscopic video quality assessment based on 3D convolutional neural networks. *Neurocomputing*, **309**, (2018), 83–93.

[115] Lee S. *et al.*: Foveated image/video quality assessment in curvilinear coordinates, in *Int'l. Workshop on Very Low Bitrate Video Coding*, Urbana, IL, USA, October 1998, 189–192.

[116] Sun W. *et al.*: A large-scale compressed 360-degree spherical image database: from subjective quality evaluation to objective model comparison, in *2018 IEEE 20th Int. Workshop on Multimedia Signal Processing*, 2018, 1–6.

[117] Duan H. *et al.*: Perceptual quality assessment of omnidirectional images, in *IEEE Int. Symp.on Circuits and System*, 2018, 1–5.

[118] Li C. *et al.*: Bridge the gap between vqa and human behavior on omnidirectional video: a large-scale dataset and a deep learning model. arXiv preprint arXiv:1807.10990 (2018).

[119] Zhang B. *et al.*: Subjective and objective quality assessment of panoramic videos in virtual reality environments, in *IEEE Int. Conf. on Multimedia and Expo Workshops*, 2017.

[120] Qian F. *et al.*: Optimizing 360 video delivery over cellular networks, in *Proc. the 5th Workshop on All Things Cellular: Operations, Applications and Challenges*, 2016, 1–6.

[121] Yu M.; Lakshman H.; Girod B.: A Framework to Evaluate Omnidirectional Video Coding Schemes. *IEEE Int. Symp. on Mixed and Augmented Reality*, 2015, 31–36.

[122] Sun Y.; Lu A.; Yu L.: AHG8: WS-PSNR for 360 video objective quality evaluation. document JVET-D0040 (2016).

[123] Zakharchenko V. *et al.*: Quality metric for spherical panoramic video, in *Proc. SPIE 9970, Optics and Photonics for Information Processing X*, 2016

[124] Luz G. *et al.*: Saliency-driven omnidirectional imaging adaptive coding: modeling and assessment, in *IEEE 19th Int. Workshop on Multimedia Signal Processing*, 2017, 1–6.

[125] Xu M. *et al.*: Assessing visual quality of omnidirectional videos. *IEEE Trans. Circuits Syst. Video Technol.* (2018), 1–1.

[126] Yu M. *et al.*: A framework to evaluate omnidirectional video coding schemes, in *2015 IEEE Int. Symp. on Mixed and Augmented Reality*, 2015, 31–36.

[127] Ozcinar C. *et al.*: Visual attention-aware omnidirectional video streaming using optimal tiles for virtual reality. *IEEE J. Emerging Sel. Topics Circuits Syst.*, **9** (1) (2019), 217–230.

[128] Kim H.G. *et al.*: Deep virtual reality image quality assessment with human perception guider for omnidirectional image. *IEEE Trans. Circuits Syst. Video Technol.*, (2019).

[129] Lim H.-T. *et al.*: VR IQA NET: deep virtual reality image quality assessment using adversarial learning, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2018, 6737–6741

[130] Li C. *et al.*: Viewport Proposal CNN for 360deg Video Quality Assessment, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, 2019, 10177–10186.

[131] Padmanaban N. *et al.*: Towards a machine-learning approach for sickness prediction in 360 stereoscopic videos. *IEEE Trans. Vis. Comput. Graph.*, **24** (4) (2018), 1594–1603.

[132] Kim J. *et al.*: A deep cybersickness predictor based on brain signal analysis for virtual reality contents, in *Proc. of the IEEE Int. Conf. on Computer Vision*, 2019.

[133] Kim J. *et al.*: Deep blind image quality assessment by learning sensitivity map, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. IEEE, 2018.

Woojae Kim received the B.S. degree in electronic engineering from Soongsil University, Korea, in 2015. He is currently pursuing the M.S. and Ph.D. degrees with the Multidimensional Insight Laboratory, Yonsei University. He was a Research Assistant under the guidance of Prof. Weisi Lin with the Laboratory for School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore in 2018. His research interests include image and video processing based on the human visual system, computer vision, and deep-learning.

Sewoong Ahn received the B.S. degree in fusion electronic engineering from Hanyang University, Korea, in 2015. He is currently pursuing the M.S. and Ph.D. degrees with the Multidimensional Insight Laboratory, Yonsei University. His research interests include 2D/3D image and video processing based on human visual system, 3D virtual reality, and deep-learning.

Anh-Duc Nguyen received the B.S degree in automatic control from the Hanoi University of Science and Technology, Vietnam in 2015.He is currently pursuing the M.S. and Ph.D. degrees with the Multidimensional Insight Laboratory, Yonsei University. His research interests are computer vision, image/video analysis, and machine learning.

Jinwoo Kim received the B.S. degree in electrical and electronic from Hongik University, South Korea, in 2016. He is currently pursuing the M.S. and Ph.D. degree with the Multidimensional Insight Laboratory, Yonsei University. His research interests are in the area of quality assessment, computer vision, and machine learning.

Jaekyung Kim received the B.S. degree in electrical and electronic engineering from Yonsei University, Korea, in 2017. He is currently pursuing the M.S. and Ph.D. degrees with the Multidimensional Insight Laboratory, Yonsei University. His research interests include perceptual image/video/VR/AR processing, computational photography, and deep-learning.

Heeseok Oh received the B.S., M.S., and Ph.D. degrees in electrical and electronics engineering from Yonsei University, Korea, in 2010, 2012, and 2017, respectively. He is currently with ETRI, Korea. His current research interests include 2D/3D image and video processing based on human visual system, machine learning, and computational vision.

Sanghoon Lee received the B.S. degree from Yonsei University, Korea, in 1989, the M.S. degree from the KAIST, Korea, in 1991, and the Ph.D. degree from The University of Texas at Austin, Austin, TX, USA, in 2000, all in E.E. From 1991 to 1996, he was with Korea Telecom, Korea. From 1999 to 2002, he was with Lucent Technologies, NJ, USA. In 2003, he joined the Department of Electrical and Electronics Engineering, Yonsei University, as a Faculty Member, where he is currently a Full Professor. His current research interests include image/video quality assessment, computer vision, graphics, and multimedia communications. He has been currently serving as the Chair of the APSIPA IVM Technical Committee since 2018 and also the Chair of the IEEE P3333.1 Quality Assessment Working Group since 2011.