Review of subjective quality assessment methodologies and standards for compressed images evaluation

Michela Testolina and Touradj Ebrahimi

Multimedia Signal Processing Group (MMSPG)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
Email: firstname.lastname@epfl.ch

ABSTRACT

Acquiring and exchanging a large number of pictures has become nowadays a common practice. Therefore, new image compression solutions to optimize the storage resources are in constant demand. In this context, it is essential to have a solid methodology to evaluate the performance of compression techniques. Such performance is usually measured through objective quality metrics, which are fast and inexpensive but not always reliable. However, performance is best assessed through subjective image quality assessment experiments, which are expensive and time-consuming, but reliable as based on the subjective opinion of a large number of subjects. These experiments are usually conducted in a controlled laboratory environment, with high-quality monitors and controlled lighting conditions. Recently, encouraged by the COVID-19 pandemic, crowdsourcing-based subjective image quality experiments are gaining popularity, and have demonstrated to be a faster and cheaper alternative to traditional approaches in subjective quality assessment. In this paper, different methodologies for subjective image quality assessment experiments are examined, including a review of the released standards as well as a list of publicly available tools. Moreover, the analysis is extended to novel plenoptic imaging techniques, i.e. point clouds and light fields, and to visually lossless quality assessment approaches. Authors hope that this work will help researchers interested in conducting subjective experiments for assessing the quality of compressed media, to better select the appropriate methodology for their use cases.

Keywords: Perceptual quality assessment, subjective quality, image compression, crowdsourcing, point clouds, light fields, visually lossless compression

1. INTRODUCTION

On a daily basis, billions of pictures are acquired by digital devices and generally compressed, e.g. with JPEG, before storage or delivery. Image compression plays, in fact, a fundamental role, making possible efficient storage and delivery of large amount of rich media such as images and video. While ensuring a lower storage space , a lossy image compression introduces distortions that might become visible to the human eye, and in a variable amount depending on the content, the degree of compression and the context such as the environment and the type of display. Each compression method produces different artifacts in images, including blocking, blurring or ringing artifacts, color shift, and others. In order to standardize new and efficient lossy compression, it is fundamental to conduct perceptual quality assessment experiments to assess the severeness of the introduced visual artifacts. The visual quality can be assessed objectively through a number of objective quality metrics, or subjectively by collecting the individual opinion on the presented compressed content from a large number of people. While objective quality assessment is fast and inexpensive, subjective quality assessment is slow and costly, but it has the advantage of being more reliable because it is based on the opinion of human observers. In order to standardize the subjective quality assessment experiments, some recommendations have been proposed by both ITU-T and ITU-R. Most such methods are focused on controlled laboratory environment and for traditional image modalities. In recent years, motivated by the confinements caused by the COVID-19 pandemic, many subjective quality experiments have been conducted using an uncontrolled crowdsourcing approach, in which subjects are hired remotely, conducting the subjective experiment on their own environments. While this type of approach is not as popular as the controlled environment, some best-practices documents have been presented.

In recent years, and with the increasing number of high quality capture and display devices accessible to general public, the interest to visually lossless image compression has grown. In this context, the Joint Photographic Experts Group (JPEG) released standardized methodologies for assessing the quality of visually lossless approaches in AIC Part-2. Also, novel imaging techniques, such as point clouds and light fields, have gained in popularity, but there is still a lack of standardized methodology for subjective quality assessment of such new modalities.

The aim of this paper is to review a wide range of different subjective quality assessment methodologies and standards, focusing on image compression applications. Other similar works have been proposed previously, e.g. Lee et al.,[1] which compared 3 different subjective quality assessment methodologies in a controlled laboratory environment. Another similar work has been proposed by Pinson et al.,[2] which compared six different subjective video quality assessment methodologies. However, a review of different subjective quality techniques, both in controlled and based on crowdsourcing uncontrolled environments, as well as approaches for novel imaging techniques, has not yet been proposed.

The paper is organized as follows. In Section 2 different subjective quality assessment methodologies for the controlled environment are proposed. Section 3 introduces the problem of crowdsourcing quality assessment and the proposed recommendations. Section 4 reviews the standardized methodologies for visually lossless image quality assessment. Finally, Section 5 summarizes the proposed subjective quality assessment methodologies for new imaging modalities, i.e. point clouds and light fields. Conclusions are drown in Section 6.

2. CONTROLLED ENVIRONMENT SUBJECTIVE QUALITY ASSESSMENT

The most popular approach to assess the quality of compressed images is the controlled environment subjective quality assessment, where the experiment is conducted in a test laboratory with controlled lighting conditions. It is in fact critical to create ideal conditions in order to avoid noise and bias that can cause fluctuations in the subjective scores. Different standards have been proposed to recreate such a testing environment for subjectively assessing the visual quality of images. As an example, the International Telecommunication Union (ITU) presented multiple public recommendations documents, namely Recommendation ITU-R P.910,[3] Recommendation ITU-R P.913[4] and Recommendation ITU-R BT.500-14,[5] which provide different methodologies for visual assessment of image quality. These documents describe the ideal testing conditions, i.e., low room illumination with peak luminance $70 - 250cd/m^2$, monitor contrast ratio less or equal to 0.02 and a ratio of about 0.15 between the luminance of background behind the picture in the monitor and the peak luminance of the picture.

A crucial element for a reproducible subjective quality assessment is in the selection of the monitor, and therefore the choice of the visualization device is essential in order to obtain reliable results. BT.500-14[5] provides a detailed description of the main characteristics that a monitor suitable for a subjective quality evaluation experiment should have. Figure 1 shows an example of a test room with controlled environment, isolated from natural light, and with an adjustable artificial lighting system. The monitor is selected depending on the specific type of subjective quality experiment.

Another key factor is the choice and the number of subjects participating in the experiment. In particular, the subjective quality experiment might be conducted by *experts*, i.e. people who are already familiar with the type of image artifacts assessed in the experiment, or *non-experts/ naïve* subjects, i.e. people that have no previous experience in the type of artifacts or in image compression in general. In any case, ITU-R recommends choosing a number of viewers equal or greater than 15. It is also important that all the subjects have a normal or corrected visual acuity. Prior to the session, all the subjects should be introduced to the experiment and familiarized with the objective of the experiment, grading scale and timing. The test sessions should not be longer than half an hour, in order not to fatigue the subjects, who might produce unreliable results afterward. Moreover, the images should be presented to the subjects in random order.

A crucial element in all experiments involving human subjects is the data protection and respect of privacy. Although, the exact details often depend on the legislation and best practices of the country or organization where the experiments are carried out, anonymization of subjects and conditional access to the collected data, including various restrictions in the type of personal data that can be collected, are common elements among many and should be strictly adhered to when running experiments.

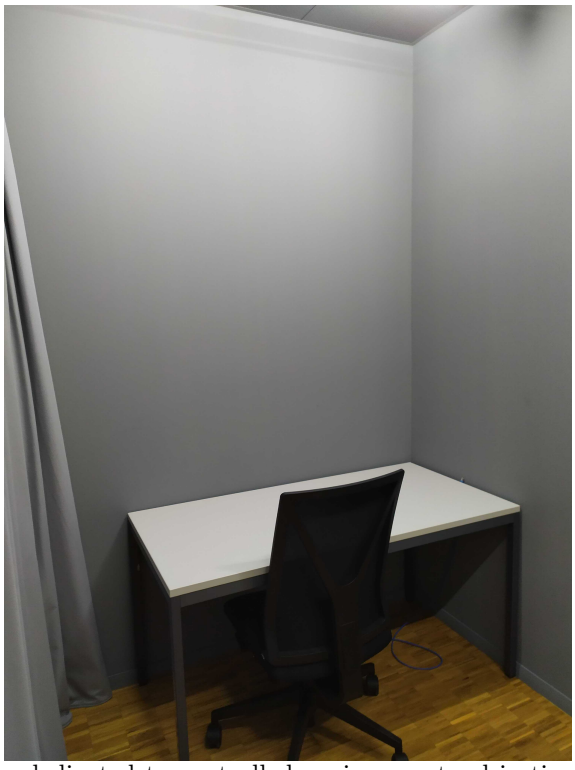

Figure 1: Example of test room dedicated to controlled environment subjective quality assessment evaluation. The room does not have any natural light and the artificial lighting conditions are adjustable. The monitor can be selected depending on the experiment type.

ITU-R reports different experimental methodologies, that can be separated into two main macro-categories: single stimulus (SS) and double stimulus (DS) methodologies. The main difference is in the number of stimuli that are presented to subjects: in particular, in the SS the subjects rate the quality of a single image, while in the DS methodology the subjects rate the impairment between two images, shown side by side. Each of these two macro-categories then includes their specific experimental methodologies, each with a different grading and scale approach.

A common practice at the beginning of an experiment is to introduce the subjects to the objectives of the experiment, acquaint them with the types of artifacts and familiarize them with the grading scale through a training session. This phase is crucial in order to limit the chances of misunderstanding and to obtain reliable results. It's therefore important to dedicate a suitable number of images to this task and to give clear and easy to carry out instructions to subjects. The next subsections present in more details the different subjective quality assessment methodologies.

2.1 Single stimulus methodologies

The single stimulus (SS) methods consist in presenting to the participants in the experiment a sequence of images, one at a time, asking them to rate their visual quality. The grading scale varies from experiment to experiment, and a training phase is usually performed at the beginning of the experiment. This subjective methodology is a popular choice due to its simplicity and low number of steps. As an example, the single stimulus methodology has been used in several experiments such as those conducted by Sheikh et al.[6] or by Cheng et al.[7]

Some of the most popular methodologies for SS subjective quality assessment are:

Absolute Category Rating (ACR): is a type of single stimulus subjective quality experiment where the test stimuli are presented one at a time and the subjects are asked to rate the visual quality of the images on a

discrete scale from 1 to 5, namely:

1. Bad
2. Poor
3. Fair
4. Good
5. Excellent

The advantages of such a method are the simplicity in its design and in the computation of the subjective scores, but, it usually requires a long training session in order to acquaint the subjects with the grading scale. Moreover, it has been observed that subjective opinion is occasionally influenced by subjects' opinions on the content of the stimulus. In order to mitigate the influence of the image content on the subjective scores, the ACR-HR methodology is used.

Absolute Category Rating with Hidden Reference (ACR-HR): as introduced above, it is a variation of the ACR where the original image is "hidden" among the distorted stimuli, without informing the subjects of such occurrence. This experimental methodology allows to remove the variance due to the subjects' personal opinion on the content, and allows to compute the differential mean opinion score (DMOS) rather than the mean opinion score (MOS), obtaining a more precise evaluation of the quality of the stimuli. Due to the trade-off between its simplicity and accuracy, this methodology is widely used in the state of the art. As an example, this methodology was utilized for the large-scale and publicly available subjective quality assessment study conducted by the Laboratory for Image and Video Engineering (LIVE)[6] and in Cheng et al.[7] to evaluate the perceptual quality of learning-based image compression methods.

Single Stimulus Continuous Quality Evaluation (SSCQE): is an alternative type of single stimulus subjective quality experiment similar to the ACR, using a continuous evaluation scale rather than the discrete one. As suggested in BT.500-14,[5] an electronic recording handset connected to the computer should be used, but recently slide-bars displayed directly on the screen, similar to that shown in Figure 2, on the left, are more popular. The advantage of the continuous quality scale is its similarity with the continuous grading scale of objective quality metrics, in order to have a more accurate comparison with such type of quality evaluation approach.

2.2 Double stimulus methodologies

The double stimulus (DS) method is a different type of subjective quality experiment, consisting of showing to subjects a couple of stimuli, displayed side-by-side, while they are asked to evaluate the impairment between the two images. Also in this case, the grading scale differs from experiment to experiment and will be introduced in the following paragraphs. In general, the double stimulus methodologies take a longer time when compared to single stimulus, but are generally more accurate on specific types of artifacts, for example, in the case of shifts in the colors of the stimuli. For this reason, the method has been used recently in multiple subjective quality experiments, e.g. in Ascenso et al.[8] and in Testolina et al.[9] to evaluate the quality of learning-based image coding solutions.

Some of the most popular methodologies for DS subjective quality assessment are:

Figure 2: Example of quality rating scale employed in subjective experiments with continuous quality scale, i.e. SSCQE and DSCQS. On the left, an example of continuous quality rating slide-bar for single stimulus experiments, on the right an example for double stimulus experiments. The continuous quality rating slide-bar includes a bar (in red) that can be moved by the subject with the mouse in order to match the desired quality level.

Double Stimulus Impairment Scale (DSIS): also known as Degradation Category Rating (DCR), the DSIS consists of showing to subjects two stimuli side-by-side, asking to rate the amount of impairment of one against the other taken as reference, according to the following discrete grading scale:

1. Very annoying
2. Annoying
3. Slightly annoying
4. Perceptible but not annoying
5. Imperceptible

The main disadvantage of this methodology is that for the same amount of time, the DSIS produces fewer ratings compared to the ACR as the subjects are requested to observe two stimuli rather than one, and therefore it is considered a slower option. However, it has the advantage that the scores are in general not influenced by the subjects' opinion of the content, as the valuation regards the impairment between images rather than the general quality. Moreover the DSIS facilitates, compared to the ACR , the detection of color impairment between two images. The reference stimulus is always presented in the same position, known to the grading subject.

The DSIS is widely used in the field of subjective quality assessment of compressed images, for example in Ascenso et al.,[8] where the authors used this methodology to compare different learning-based compression methods.

Double-Stimulus Continuous Quality-Scale (DSCQS): is a subjective quality assessment methodology similar to the DSIS, in which subjects are asked to rate the overall quality of both presented stimuli using a continuous quality rating scale, as shown in Figure 2, on the right. In this methodology the reference stimulus is displayed in a random position, unknown to the subject. As the subjects are asked to rate the quality of two stimuli at every step, this method is the slowest among those that have been presented above. Such a method is particularly useful for evaluating learning-based compression methods, as such methods might

include some image processing operations that produce images, at the highest bitrates, with higher quality than the original. An example, the DSCQS method was utilized for the subjective quality assessment experiment conducted to evaluate the submission to the JPEG AI Call for Evidence, co-organized in conjunction with the IEEE MMSP'2020 Challenge on Learning-Based Image Coding, and presented in the work proposed by Testolina et al.[9]

Double Stimulus Comparison Scale (DSCS): in the DSCS subjective quality experiment, also known as pair comparison (PC), the subjects are asked to evaluate at each step the visual quality of the first stimulus based on the second taken as reference. The grading scale is discrete, and the grades are as follow:

-3. Much worse

-2. Worse

-1. Slightly worse

0. The same

1. Slightly better

2. Better

3. Much better

In this case, subjects rate all reference and test stimuli of the same content between themselves, in a randomized order. This is therefore the experiment with the largest number of tasks, and therefore also the longest. While this method is the most accurate in evaluating performance of different compression methodologies in terms of quality, it has the disadvantage that the bitrates of the compared stimuli should be as close as possible, in order to guarantee a fair comparison. Other variants of scaling in this method could include three (Better, The same, Worse) or even only two (Better, Worse). Last but not least, partial comparisons among stimuli have been proposed in order to reduce the duration of the tests using this methodology.

A general summary of the methods introduced in this section is available in Table 1. In particular, an analysis of the advantages and disadvantages of each method is presented.

3. CROWDSOURCING-BASED SUBJECTIVE QUALITY ASSESSMENT

Besides subjective quality assessment methods in controlled environments, subjective image quality assessment can be conducted based on crowdsourcing. In this methodology, the subjects are hired remotely and are able to conduct the experiment directly in their environment. The experiment is therefore performed in an uncontrolled environment, but it is more likely to be comparable to real viewing conditions of digital media in a more realistic set up.

Such an approach has been adopted since the first decade of the year 2000, e.g. in the experiment presented in 2009 by Chen et al.[10] However in recent years, due to the worldwide confinements caused by the COVID-19 pandemic, this approach has shown increasing interest and popularity. Moreover, the European Network on Quality of Experience in Multimedia Systems and Services (Qualinet) [*], has worked towards the definition of a number of guidelines for crowdsourcing subjective image quality assessment. Specifically, in 2014 the Qualinet Task Force on Crowdsourcing produced a whitepaper on the best practices and recommendations for Crowdsourced QoE.[11] More recently, the ITU-T SG12 is working toward a set of recommendations to carry subjective quality assessment based on crowdsourcing.

In the following, the main recommendations proposed in the paper are summarized:

[*] http://www.qualinet.eu

Table 1: Summary of the different methods for subjective quality assessment, and their advantages and disadvantages.

Method	Type	Scale Type	Advantages	Disadvantages
ACR	SS	Discrete	Fast and simple	Scores influenced by the subjects' opinion on the content
ACR-HR	SS	Discrete	Allows to remove the variance due to the subjects' personal opinion on the content	Requires a long training procedure to acquaint the subjects with the artifacts
SSCQE	SS	Continuous	Comparable to the continuous grading scale of objective quality metric	Requires a long training procedure to acquaint the subjects with the artifacts
DSIS	DS	Discrete	Not influenced by the subjects' opinion on the content, reliable in evaluating color impairment	Slower than ACR
DSCQS	DS	Continuous	Both the original and impaired stimuli are graded	Slower than DSIS
DSCS	DS	Discrete	Compares all the different stimuli among themselves	Biggest number of comparisons, bitrate matching is critical

- Utilize a **user-friendly software**, easy-to-use and without requiring admin installations. In this way, more subjects will be able to participate in the experiment. A popular solution to the problem, is the usage of web-based applications, where the subjects only need to connect to a web server without installing any software on their machine.
- A crowdsourcing experiment is more diverse in terms of spoken languages and cultural background for its nature, and it is therefore essential to use **simple and direct questions** in order to minimize the chances of misunderstandings by the subjects attending the experiment.
- It is important to perform a subjective experiment of the **proper duration**, to avoid fatigue in the subjects. For this reason, it's suggested to have even shorter sessions than those in the controlled environment. In addition, the reward should be proportional to the experiment time, to encourage more participants to take part in the experiment. Moreover, participation of subjects with direct relationship with the organizer of the experiment should be avoided.
- As it is not possible to directly get immediate feedback from the subject, it is important to provide a suitable and exhaustive **training sessions** in order to avoid poor quality of the subjective scores due to misunderstandings. It is also useful to address well known issues that have been experienced in previous experiments in controlled or uncontrolled environments.
- It is advisable to collect **feedback** from the subjects, in order to improve the experiment or to correct common issues identified among participants.
- Collecting **event logging**, and therefore information of what occurs during the experiment, is critical in order to evaluate the quality of the submitted subjective scores. For example, the clicking behavior, the window resizing operations, the page reloading, the switching of the tab are all important factors in order to understand the behavior of the subject during the experiment.

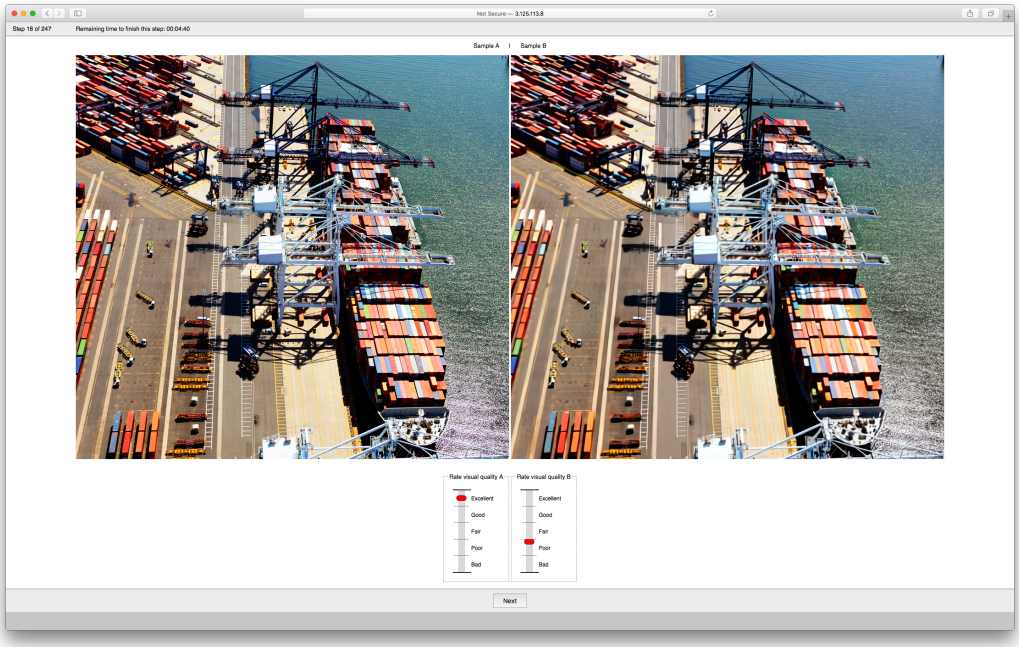

Figure 3: Example of interface of a crowdsourcing-based DSCQS experiment.

- Including **honeypot questions** is also advisable in order to evaluate the amount of attention that the subjects are paying to the experiment. As an example, the semantic content of the previous image or simple general questions (e.g. "Five plus 2 = ?") can be asked during the experiment. In this way, it is possible to detect the subjects who have not focused adequately on the experiment, or that are answering randomly. The reliability questions can be asked during or after the experiment.

Most of the crowdsourcing approaches are web-based frameworks that can be run thought a widely used browser, and that therefore don't require the installation of any software. Egger et al.[12] presented different crowdsourcing frameworks for subjective quality assessment, summarizing the advantages and disadvantages of each. Among the most popular one we can find Euphoria,[10] Crowd MOS,[13] Quality Crowd 2,[14] WESP,[15] BeaqleJS,[16] in-momento[17] and Crowdee.[18]

Multiple subjective quality assessment experiments based on crowdsourcing have been conducted and reported over the years. Among the most recent experiment, Testolina et al.[9] presented the results of the crowdsourcing-based subjective experiment conducted to assess the performance of the submissions to the JPEG AI Call for Evidence, co-organized with the IEEE MMSP'2020 Challenge on Learning-Based Image Coding. The experiment was conducted on Amazon Mechanical Turk by means of the Quality Crowd 2 framework,[14] using the DSCQS methodology. Figure 3 shows the interface that was displayed by the subjects during the experiments.

In Table 2, the main characteristics of the controlled environment and crowdsourcing subjective quality assessment are summarized.

4. SUBJECTIVE QUALITY ASSESSMENT FOR VISUALLY LOSSLESS QUALITY

The previously mentioned subjective quality assessment methodologies and standards are designed for web-quality applications, i.e. applications with a limited or variable bitrate requirement, and therefore images that present visual artifacts that are usually easily perceivable by the human eye. In recent years, the number of applications that target high, or perceptually lossless, visual qualities is increasing.[19] In the context of storage applications, in fact, limited memory consumption is no longer the main requirement, thanks to cheap and large portable storage devices or to cloud storage services. Therefore users demand compression methods that maximize the visual quality of the images, rather than minimizing their bitrate consumption. In this context,

Table 2: Summary of the main characteristics of the controlled environment subjective quality assessment method vs the crowdsourcing subjective quality assessment method

Controlled environment	**Crowdsourcing**
Expensive and time consuming	Fast and cheaper
It is possible to test one variable at the time	It is conducted in an uncontrolled but realistic environment
Low diversity in the participants	High diversity in the participants
Relevant for professionals who work in a controlled environment	Relevant for media broadcasting applications

the standard double stimulus experiments are not accurate enough, as it is difficult to assess some types of subtle artifacts with such methodologies. As an example, it is usually difficult to detect slight shifts in the colors when images are presented alone or even side by side.

To address this issue, the Joint Photographic Experts Group (JPEG) committee has released the standard ISO/IEC 29170-2 (AIC Part-2),[20,21] which includes standardized methodologies for subjective visual quality assessment of lossless or nearly lossless visual qualities. In particular, two different methodologies have been proposed, relatively in the Annex A and B of the standard, notably:

- Methodology A: two distorted test images, along with the original image, are presented to the subjects, which are asked to choose the closest test image to the original one. The subjects have 4 seconds to select their preference.
- Methodology B, or "Flicker" Test: one distorted test image along with the original image are presented to subjects, side-by-side. The test image is interleaved at a certain frequency with the original; therefore, in the case in which the test image presents some degradation perceivable in the quality, the test image will appear "flickering". The position of the original and "flickering" stimuli is random and unknown to the evaluating subjects. The subjects are asked to choose which one of the two stimuli presents the flickering. If the image presents visible distortions, the flickering will be visible by the human eye and the subjects are able to detect the flickering image correctly. If the distortions in the test images are not perceivable by the human eye, the subjects are not able to correctly detect the "flickering" image and will answer randomly.

As the visually lossless approaches to compression has gained popularity only recently, these methodologies are not as popular as those using controlled environment discussed earlier, but have been occasionally used in a few studies. For example, Willeme et al.[22] adopted such methodologies for visual quality assessment of the JPEG XS.

5. PLENOPTIC IMAGING QUALITY ASSESSMENT

In recent years new emerging imaging technologies, like point cloud and light field imaging, have gained popularity and increasing interest. These types of imaging systems generally require a greater amount of storage space when compared to conventional images, and therefore image compression is essential to make these approaches available to a larger public. However, only a few compression methods have been proposed, due to their recent interest from the community. In this context, it is essential to have a reliable and repeatable subjective quality assessment methodology, in order to accurately assess the performance of each proposed method. Furthermore, no standardized methodology for subjective quality assessment of such subjective methods has been proposed yet. Instead, multiple subjective quality assessment experiments exploring different methodologies have been reported in the state of the art for both. In the next subsections, the most relevant subjective methodologies for point clouds and light fields are presented and analyzed.

5.1 Point cloud quality assessment

A point cloud is a cluster of points in the space, each of them defined by their coordinates X, Y and Z on a certain coordinate system, and optionally a color component R, G and B, generally used for the rendering of 3D models in virtual environments. This technology gained its popularity only in recent years, and therefore the research in subjective quality assessment methodologies is not as mature as for conventional images. In particular, currently no standardized methodology has been proposed. Instead, multiple subjective methodologies have been attempted, taking inspiration from subjective quality assessment standards for conventional images.

A simple and popular strategy for point clouds quality assessment consists of converting them into a video sequence, which captures all the possible viewpoints, as if it was recorded from a virtual camera rotating around the 3D objects.[23–25] In this way, the videos can be displayed on a regular 2D monitor. As the visualized stimuli are simple videos, any of the single stimulus methodologies in Section 2.1 or double stimulus in Section 2.2 can be used, and the ITU-R recommendations for controlled environment subjective quality assessment can be followed. Additionally, due to its simplicity, this approach is also suitable for crowdsourcing-based subjective point clouds quality experiments.

Recently more immersive and interactive methodologies for point cloud quality assessment have been proposed. As an example, Mekuria et al.[26] proposed to assess the quality of point clouds in a 3D tele-immersive system where the subjects were represented by their own 3D avatar and were able to move across a virtual room and interact with other 3D avatars. The experiment was conducted on a standard 2D display, and the subjects were able to explore the scene with a standard mouse. Another interactive approach was proposed by Alexiou et al.,[27] who proposed to leave the subjects free to interact with the point clouds by moving the mouse and with no time limit. The experiment was conducted using the double stimulus methodology, and therefore both the point clouds were rotated simultaneously. Successively, the same authors also proposed a novel augmented reality (AR) experiment methodology where the subjects evaluated the point clouds through a head mounted device[28] and on a 3D display.[29]

Table 3 summarizes the main features of the point cloud subjective quality experiments cited in this paper.

Table 3: Summary of the strategies adopted in subjective point clouds quality evaluation.

Experiment	Methodology	Display Type	Interaction
Javaheri et al.[23]	DS	2D Display	Passive
Zerman et al.[24]	DS	2D Display	Passive
Su et al.[25]	DS	2D Display	Passive
Mekuria et al.[26]	SS	2D Display	Active
Alexiou et al.[27]	DS	2D Display	Active
Alexiou et al.[28]	DS	Head Mounted Device	Active
Alexiou et al.[29]	DS	3D Display	Active

5.2 Light filed quality assessment

Another new emerging imaging technology, light field, aims at collecting the light rays coming from multiple spatial directions. Light fields are typically collected with multi-camera arrays or with plenoptic cameras, e.g., Lytro and Raytrix, incorporating an array of micro-lenses. As a result, they collect multiple densely sampled views, or images that slightly differentiate between each other, with the disadvantage of requiring a huge amount of storage space. For this reason, light field compression plays a fundamental role, in order to make this new imagining type more easily accessible. To evaluate the performance of compression methods, it is important to design a proper subjective quality assessment methodology, taking into consideration the increasing complexity of the problem. In fact, the diversity of the acquisition techniques, rendering methods and distortion types make light field subjective quality evaluation a complex task. Currently, no standardized methodology has been proposed.

Different strategies can be adopted for a light field subjective quality experiment:

- Methodology: as for images, the experiment can be conducted using SS or DS methodologies, and in particular all the mythologies in Section 2 can be used. As an example, the MPI-LFA dataset[30] was conducted using the ACR methodology followed by a novel double stimulus methodology, the VALID subjective light field dataset[31] was collected using the DSIS methodology, the LF dataset[32] using the DSCQS methodology and the SAMRT dataset[33] using the DSCQS methodology.

- Display type: light fields should preferably be displayed on Light Field displays, but such technology is still under development and fairly expensive. Therefore, a popular choice is to use the standard 2D monitor or a stereo display instead, using one of the visualization strategies presented below.

- Light field visualization strategy: depending on the display type in which the light field is displayed, different visualization techniques can be adopted. As it was introduced above, a popular choice for the visualization device is the 2D monitors or the stereo displays. When visualizing the point clouds on such devices, different visualization strategies can be applied, and in particular among the most popular strategies we find:

  - *Pseudo-Video (PV)*, in which the different views are combined in a video. As an example, the LF dataset[32] utilized this visualization strategy;
  - *Interactive Visualization*, in which the observers are free to interact and navigate in the scene through devices like a simple mouse, as in VALID dataset,[31] or a head tracking device, as in the MPI-LFA dataset;[30]
  - *Refocused-pseudo-video (RV)* in which a series of images obtained by refocusing the light field image at different depth planes are combined in a video;
  - *All-in-Focused Image* in which one single image, with all the depth planes in focus, is shown to the observers using a classic subjective image quality assessment experiment. This visualization strategy was adopted in the SAMRT dataset;[33]
  - *Refocused Images* where some of the images with a single depth plane on focus are selected and used in a classic subjective image quality assessment experiment, dividing them into multiple experiments if necessary.

Table 4 summarize some subjective quality assessment experiments presented in the state of the art and their characteristics.

Table 4: Summary of the strategies adopted in some subjective light field quality evaluation experiments.

Experiment	**Methodology**	**Display Type**	**LF Visualization**
MPI-LFA[30]	ACR and a novel DS methodology	3D stereo	Interactive visualization (webcam)
VALID dataset[31]	DSIS	2D	Interactive visualization (mouse)
LF dataset[32]	DSCQS	2D	Pseudo video
SMART[33]	DSCS	2D	All-in-focused image

Moreover, the impact of the different light field subjective quality assessment strategies have been assessed in multiple works, e.g. in Paudyal et al.,[34] in Battisti et al.[35] and in Ribeiro et al.,[36] in which a variance of the subjective scores when applying the different methodologies was observed. Paudyal et al.[34] also assessed that the most reliable and consistent method is the *Pseudo-Video (PV)* visualization strategy.

6. CONCLUSIONS

In this paper a number of subjective quality assessment methodologies and standards, specific to the problem of image compression, have been reviewed. The survey includes the most popular methodologies for controlled

environment subjective quality assessment as well as the recommendations for crowdsourcing quality assessment and the novel methodologies for visually lossless qualities. Finally, the latest attempts of subjective quality assessment for the novel point cloud and light field imaging technologies have been reviewed. This work could guide the authors interested in conducting a subjective quality experiment to assess the quality of compressed media towards the selection of the proper experimental methodology for their specific problem.

ACKNOWLEDGMENTS

The authors would like to acknowledge support from the Swiss National Scientific Research project entitled "Advanced Visual Representation and Coding in Augmented and Virtual Reality" under grant number 200021_178854.

REFERENCES

[1] Lee, C., Choi, H., Lee, E., Lee, S., and Choe, J., "Comparison of various subjective video quality assessment methods," in [*Image Quality and System Performance III*], **6059**, 605906, International Society for Optics and Photonics (2006).

[2] Pinson, M. H. and Wolf, S., "Comparing subjective video quality testing methodologies," in [*Visual Communications and Image Processing 2003*], **5150**, 573–582, International Society for Optics and Photonics (2003).

[3] Recommendation ITU-T P.910, "Subjective video quality assessment methods for multimedia applications," *International Telecommunication Union* (2008).

[4] Recommendation ITU-T P.913, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," *International Telecommunication Union* (2016).

[5] Recommendation ITU-T BT.500-14, "Methodologies for the subjective assessment of the quality of television images," *International Telecommunication Union* (2019).

[6] Sheikh, H. R., Sabir, M. F., and Bovik, A. C., "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on image processing* **15**(11), 3440–3451 (2006).

[7] Cheng, Z., Akyazi, P., Sun, H., Katto, J., and Ebrahimi, T., "Perceptual quality study on deep learning based image compression," in [*2019 IEEE International Conference on Image Processing (ICIP)*], 719–723, IEEE (2019).

[8] Ascenso, J., Akyazi, P., Pereira, F., and Ebrahimi, T., "Learning-based image coding: early solutions reviewing and subjective quality evaluation," in [*Optics, Photonics and Digital Technologies for Imaging Applications VI*], **11353**, 113530S, International Society for Optics and Photonics (2020).

[9] Testolina, M., Upenik, E., Ascenso, J., Pereira, F., and Ebrahimi, T., "Performance evaluation of objective image quality metrics on conventional and learning-based compression artifacts," *13th International Conference on Quality of Multimedia Experience* (2021).

[10] Chen, K.-T., Wu, C.-C., Chang, Y.-C., and Lei, C.-L., "A crowdsourceable qoe evaluation framework for multimedia content," in [*Proceedings of the 17th ACM international conference on Multimedia*], 491–500 (2009).

[11] Hoßfeld, T., Hirth, M., Redi, J., Mazza, F., Korshunov, P., Naderi, B., Seufert, M., Gardlo, B., Egger, S., and Keimel, C., "Best practices and recommendations for crowdsourced qoe-lessons learned from the qualinet task force" crowdsourcing"," (2014).

[12] Egger-Lampl, S., Redi, J., Hoßfeld, T., Hirth, M., Möller, S., Naderi, B., Keimel, C., and Saupe, D., "Crowdsourcing quality of experience experiments," in [*Evaluation in the crowd. Crowdsourcing and human-centered experiments*], 154–190, Springer (2017).

[13] Ribeiro, F., Florêncio, D., Zhang, C., and Seltzer, M., "Crowdmos: An approach for crowdsourcing mean opinion score studies," in [*2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*], 2416–2419, IEEE (2011).

[14] Keimel, C., Habigt, J., Horch, C., and Diepold, K., "Qualitycrowd—a framework for crowd-based quality evaluation," in [*2012 Picture coding symposium*], 245–248, IEEE (2012).

[15] Rainer, B., Waltl, M., and Timmerer, C., "A web based subjective evaluation platform," in [*2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*], 24–25, IEEE (2013).

[16] Kraft, S. and Zölzer, U., "Beaqlejs: Html5 and javascript based framework for the subjective evaluation of audio quality," in [*Linux Audio Conference, Karlsruhe, DE*], (2014).

[17] Gardlo, B., Egger, S., Seufert, M., and Schatz, R., "Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based qoe testing," in [*2014 IEEE International Conference on Communications (ICC)*], 1070–1075, IEEE (2014).

[18] Naderi, B., Polzehl, T., Beyer, A., Pilz, T., and Möller, S., "Crowdee: mobile crowdsourcing micro-task platform for celebrating the diversity of languages," in [*Fifteenth Annual Conference of the International Speech Communication Association*], (2014).

[19] Testolina, M., Upenik, E., and Ebrahimi, T., "Comprehensive assessment of image compression algorithms," in [*Applications of Digital Image Processing XLIII*], **11510**, 1151020, International Society for Optics and Photonics (2020).

[20] ISO/IEC 29170-2:2015 Information technology — Advanced image coding and evaluation — Part 2: Evaluation procedure for nearly lossless coding.

[21] Stolitzka, D. F., Schelkens, P., and Bruylants, T., "New procedures to evaluate visually lossless compression for display systems," in [*Applications of Digital Image Processing XL*], **10396**, 103960O, International Society for Optics and Photonics (2017).

[22] Willème, A., Mahmoudpour, S., Viola, I., Fliegel, K., Pospíšil, J., Ebrahimi, T., Schelkens, P., Descampe, A., and Macq, B., "Overview of the jpeg xs core coding system subjective evaluations," in [*Applications of Digital Image Processing XLI*], **10752**, 107521M, International Society for Optics and Photonics (2018).

[23] Javaheri, A., Brites, C., Pereira, F., and Ascenso, J., "Subjective and objective quality evaluation of 3d point cloud denoising algorithms," in [*2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*], 1–6, IEEE (2017).

[24] Zerman, E., Gao, P., Ozcinar, C., and Smolic, A., "Subjective and objective quality assessment for volumetric video compression," *Electronic Imaging* **2019**(10), 323–1 (2019).

[25] Su, H., Duanmu, Z., Liu, W., Liu, Q., and Wang, Z., "Perceptual quality assessment of 3d point clouds," in [*2019 IEEE International Conference on Image Processing (ICIP)*], 3182–3186, IEEE (2019).

[26] Mekuria, R., Blom, K., and Cesar, P., "Design, implementation, and evaluation of a point cloud codec for tele-immersive video," *IEEE Transactions on Circuits and Systems for Video Technology* **27**(4), 828–842 (2016).

[27] Alexiou, E. and Ebrahimi, T., "On subjective and objective quality evaluation of point cloud geometry," in [*2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*], 1–3, IEEE (2017).

[28] Alexiou, E., Upenik, E., and Ebrahimi, T., "Towards subjective quality assessment of point cloud imaging in augmented reality," in [*2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*], 1–6, IEEE (2017).

[29] Alexiou, E., Pinheiro, A. M., Duarte, C., Matković, D., Dumić, E., da Silva Cruz, L. A., Dmitrović, L. G., Bernardo, M. V., Pereira, M., and Ebrahimi, T., "Point cloud subjective evaluation methodology based on reconstructed surfaces," in [*Applications of Digital Image Processing XLI*], **10752**, 107520H, International Society for Optics and Photonics (2018).

[30] Kiran Adhikarla, V., Vinkler, M., Sumin, D., Mantiuk, R. K., Myszkowski, K., Seidel, H.-P., and Didyk, P., "Towards a quality metric for dense light fields," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 58–67 (2017).

[31] Viola, I. and Ebrahimi, T., "Valid: Visual quality assessment for light field images dataset," in [*2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*], 1–3, IEEE (2018).

[32] Shan, L., An, P., Meng, C., Huang, X., Yang, C., and Shen, L., "A no-reference image quality assessment metric by multiple characteristics of light field images," *IEEE Access* **7**, 127217–127229 (2019).

[33] Paudyal, P., Battisti, F., Sjöström, M., Olsson, R., and Carli, M., "Towards the perceptual quality evaluation of compressed light field images," *IEEE Transactions on Broadcasting* **63**(3), 507–522 (2017).

[34] Paudyal, P., Battisti, F., Le Callet, P., Gutiérrez, J., and Carli, M., "Perceptual quality of light field images and impact of visualization techniques," *IEEE Transactions on Broadcasting* (2020).

[35] Battisti, F., Carli, M., and Le Callet, P., "A study on the impact of visualization techniques on light field perception," in [*2018 26th European Signal Processing Conference (EUSIPCO)*], 2155–2159, IEEE (2018).

[36] Ribeiro, F. M., de Oliveira, J. F., Ciancio, A. G., da Silva, E. A., Estrada, C. R., Tavares, L. G., Gois, J. N., Said, A., and Martelotte, M. C., "Quality of experience in a stereoscopic multiview environment," *IEEE Transactions on Multimedia* **20**(1), 1–14 (2017).