Perceptual quality evaluation of immersive multimedia content : HDR, Light Field and Volumetric Video

Ali Ak

DOCTORAT BRETAGNE LOIRE / MATHSTIC

Nantes Université

Thèse de doctorat de

NANTES UNIVERSITÉ

École Doctorale Nº 601
Mathématiques et Sciences et Technologies
de l'Information et de la Communication
Spécialité : *Traitement des images et du signal*

Par
Ali AK

Perceptual quality evaluation of immersive multimedia content : HDR, Light Field and Volumetric Video

Thèse présentée et soutenue à Nantes, le 24 Janvier 2022
Unité de recherche : Laboratoire des Sciences du Numérique de Nantes (LS2N)

Rapporteurs avant soutenance :

Maria MARTINI	Professeure, Kingston University, Angleterre
Aladine CHETOUANI	Maître de conférence, HDR, Université dÓrleans, France

Composition du Jury :

Président :	Frédéric DUFAUX	Directeur de Recherche CNRS, Paris Saclay, France
Examinateurs :	Søren FORCHAMMER	Professeur, Technical University of Denmark, Danemark
	Frédéric DUFAUX	Directeur de Recherche CNRS, Paris Saclay, France
	Federica BATTISTI	Ass. Professeure, University of Padova, Italie
Dir. de thèse :	Patrick LE CALLET	Professeur, Université de Nantes, France

ACKNOWLEDGEMENT

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 765911 (RealVision)

TABLE OF CONTENTS

INTRODUCTION

Context

Human Visual System (HVS) process a wider range of color and light than traditional standard dynamic range images. Various cues are used to understand the geometry and depth of the scene. These processes take place at any given moment in our daily lives. Thanks to the advancements in imaging and display technologies in the last decades, we are now able to provide a more realistic viewing experience to the users. However, this is still far from the hyper-realistic experience we experience on a daily basis. To this end, various immersive multimedia formats emerged—each bringing us a few steps closer to hyper-realism.

High Dynamic Range (HDR) imaging provides a wider color gamut and intensity compared to traditional media. Light field represents a scene with a multi-dimensional parallax providing geometric information. The volumetric video relies on 3D meshes or point clouds to represent an object with geometry and color information across time. One thing in common among all representations is the additional dimensionality of the represented data. While providing a more realistic experience to the users, additional dimensionality often results in extra processing steps in the imaging pipeline. Image quality may be altered in each processing step. Moreover, the additional dimensionality of these multimedia formats results in specific impairments on the content. Consequently, quality assessment of immersive multimedia content faces unique challenges based on the use cases and content types.

To this end, this thesis focuses on some of the unique challenges in quality evaluation of immersive multimedia content, including tone mapped HDR images, light fields, Volumetric Videos (VV). More specifically, we explored the following points contributing to subjective and objective quality assessment of immersive multimedia content:

1. **Investigating challenges and motivation behind transferring laboratory experiments to crowdsourcing. Proposing a set of tools for designing crowdsourced experiments and post-processing subjective annotations.**
2. **Exploring just noticeable differences and their relation to image quality.**

Proposing a quality metric that utilizes just noticeable distortion to predict image quality on a continuous scale.

3. **Studying the influencing factors and impairments for the quality assessment of light field content and proposing a light field image quality metric.**

4. **Investigating the impact of temporal sampling on the objective quality assessment of volumetric video content.**

Organization of the Thesis

Part I of the thesis is dedicated to subjective quality assessment in crowdsourcing. More specifically, we investigate the advantages, disadvantages of crowdsourcing platforms for subjective quality evaluation of tone mapped image quality assessment.

In **Chapter 1**, we provide an introductory picture of subjective image quality assessment by dividing it into a number of stages. Furthermore, we introduce the high dynamic range imaging and tone mapping operators and explore the existing work on subjective quality evaluation of tone mapped images.

Chapter 2 discusses the effect of content selection on subjective quality assessment, which has a higher impact on crowdsourcing experiments due to the large amount of data used for evaluation. To this end, we propose a modular content selection strategy designed for pairwise comparison of tone mapped images.

Chapter 3 investigates the possibility of using crowdsourcing platforms for subjective quality evaluation of tone mapped images. Three subjective experiments were conducted with controlled differences in laboratory conditions and crowdsourcing platforms to analyze the effect of uncontrolled experiment conditions and participant pools.

Based on the findings of the preceding chapters, in **Chapter 4**, we design and collect the largest publicly available tone mapped image quality evaluation dataset (RV-TMO) in the literature. Furthermore, we provide a benchmark of existing tone mapped image quality metrics on the collected subjective preferences.

Chapter 5 investigates the observer screening methodologies for crowdsourced pairwise comparison experiments. Commonly used behavioral tools were used and analyzed on the RV-TMO dataset introduced in Chapter 4. Moreover, we propose a novel outlier detection methodology for pairwise comparison experiments.

Part II focuses on objective quality evaluation of a number of multimedia types. Based on the focus of the corresponding chapter, we provide an introduction to domain specific problems and present our contributions regarding the subject.

Chapter 6 introduces the necessary concepts for understanding the proceeding chapters. We introduce state of the art objective quality metrics regarding traditional image quality assessment, light field quality assessment, and point cloud quality assessment. Furthermore, we provide an introduction to objective quality performance measures.

Chapter 7 investigates the just noticeable differences and their relation to image quality on a continuous scale. To this end, we also propose a learning-based image quality metric that utilizes the first just noticeable difference step information.

Chapter 8 focuses on objective quality assessment of light field content. We first investigate the epipolar plane images and visibility of light field specific distortions on epipolar plane images. Based on the findings, we propose a no-reference objective quality metric for light field content.

Chapter 9 analyzes the effect of temporal sampling on the objective quality metrics. The impact of the sub-sampling rate and several pooling methods were investigated.

The thesis ends with a summary of the experimental efforts, our findings, and contributions.

Part I

Subjective Quality Assessment in Crowdsourcing

Chapter 1

Subjective Image Quality Assessment

Subjective assessment is the most reliable way to evaluate image quality. In subjective experiments, a set of observers is asked to evaluate the presented stimuli according to the designed procedure. Desired procedure and the set of observers may vary according to the task, but often we want a composition of observers with balanced demographics and a simple and reproducible experiment design.

To increase the reliability of the experiment, standards and recommendations are constantly being updated, such as ITU-R recommendations [48] and ITU technical report [50] on crowdsourcing experiments.

Conducting a subjective experiment often contains three stages: designing, data collection, processing. Many design choices need to be made before conducting the experiment. One of the most important aspects is the experiment procedure, broadly categorized as rating and ranking tasks. We introduce and compare these procedures in Section 1.1. Experimental conditions also play a crucial role in subjective experiments. These are often regulated by the ITU recommendations [48] in the literature. However, recent advancements in crowdsourced studies bring numerous challenges in controlling such environmental factors. Although there are a several reports and recommendations [50, 31] on conducting crowdsourced studies, they are not complete, and they don't guarantee a smooth transition from laboratory experiments to crowdsourcing.

With the increasing popularity of learning-based approaches in objective quality metrics, the need for large-scale image quality datasets also increased. Since conducting a subjective experiment in a laboratory is often costly and time-consuming, researchers showed great interest in crowdsourced subjective experiments. Although crowdsourced studies allow researchers to reach a wide range of audiences at a lower cost, they also bring challenges. We further discuss the details regarding crowdsourced subjective quality experiments in Section 1.5.

Subjective experiments conducted in crowdsourcing platforms often utilize many stimuli to take full advantage of the modality. Choosing appropriate content for such experiments was not a challenge for laboratory experiments. Often content selection was used to ensure a representative set of images was used in the experiment. In order to provide a challenging benchmark and a desirable baseline for developing learning-based objective quality metrics properties of a good dataset evolved. We further discuss the content selection in Section 1.2.

Processing collected subjective annotations is necessary to interpret the outcome of the experiment. Depending on the experiment procedure, processing may vary. Subjective annotations acquired through rating tasks are often interpreted as MOS/DMOS along with the confidence intervals. Subjective annotations acquired with ranking tasks can be transformed into a quality scale, or direct interpretation can be made through PCM. We explain some of the popular methodologies in Section 1.3.

Prior to processing collected subjective annotations, outliers need to be removed from the experiment. Standards provide recommendations and tools for observer screening in rating experiments. However, there is no well-established observer screening methodology for ranking experiments other than a few attempts. We introduce the existing methodologies in Section 1.4.

The design of the experiment, processing tools, and observer screening may vary significantly between QoE tasks. Therefore, decisions should be made considering the aim of the subjective experiment. As it is the main QoE scenario in the first part of the thesis, we discuss the quality assessment of tone mapped images in Section 1.6.

1.1 Rating & Ranking Methodologies

For the mainly used methodologies in subjective image quality assessment, ITU standards [48] provide a comprehensive summary. This section will categorize these methodologies as rating (direct) and ranking (indirect) tasks and introduce some popular alternatives from each category.

There are specific differences between rating and ranking methodologies. The main difference, in rating methodologies, observers are asked to rate the quality of stimuli (in comparison to a pristine image, or not) directly on a predefined scale. In ranking methodologies, observers are presented with at least two stimuli and asked to rank them in terms of quality. As comically depicted in Figure 1.1, asking participants to rate the

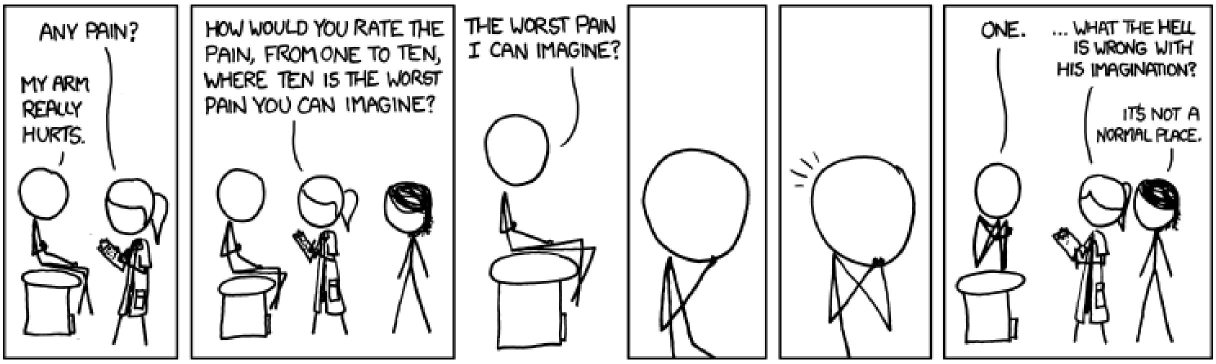

ANY PAIN?
MY ARM REALLY HURTS.
HOW WOULD YOU RATE THE PAIN, FROM ONE TO TEN, WHERE TEN IS THE WORST PAIN YOU CAN IMAGINE?
THE WORST PAIN I CAN IMAGINE?
ONE.
... WHAT THE HELL IS WRONG WITH HIS IMAGINATION?
IT'S NOT A NORMAL PLACE.

Figure 1.1 – Pain rating comic from xkcd by Randall Munroe (comic id:883).

image quality may not mean the same thing from one to another. Therefore, ranking methodologies often make the process simpler to interpret for participants (*e.g.,* which memory from these two is more painful?).

1.1.1 Rating methodologies

As mentioned earlier, rating methodologies collect subjective annotations from observers on a predefined scale. Scale values may be presented to observers as numerical (*e.g.,* 1,2,3,4,5) or categorical (*e.g.,* bad, poor, fair, good, excellent). Collected annotations later represented as mean opinion scores (MOS) or differential mean opinion scores (DMOS). MOS values indicate the perceptual quality of the corresponding stimulus. For DMOS, MOS of the reference stimulus is considered the highest possible quality, and the difference between the distorted and pristine stimuli indicates the perceptual quality of the corresponding distorted stimulus. Rating methodologies often require a training session where the range of quality differences is shown to the observer prior to the experiment, and this helps observers understand the expected quality range during the test.

Absolute Category Rating (ACR): methodology presents the stimuli to observers one by one in random order. Often a natural gray color is displayed between each stimulus. Pristine images can be displayed to observers depending on the evaluated QoE task. Generally, a five grade scale is utilized, i.e. 1 (bad), 2 (poor), 3 (fair), 4 (good), 5 (excellent). MOS scores range between 1 and 5 with this scale.

This method is the simplest to collect subjective annotations since it requires the least amount of time per stimuli. However, the accuracy of the collected scores may vary depending on the evaluated QoE task[73].

Double Stimulus Impairment Scale (DSIS): is another popular alternative rating methodology. Observers are shown with pristine and distorted images, and the effect of the degradation on the perceived quality is evaluated. The observers evaluate the degradation on a predefined scale, such as 1 (very annoying), 2 (annoying), 3 (slightly annoying), 4 (perceptible but not annoying), 5 (imperceptible). Similarly, MOS scores collected with this scale can range between 1 and 5. Due to displaying pristine stimulus with each distorted stimuli increases the time spent per stimulus.

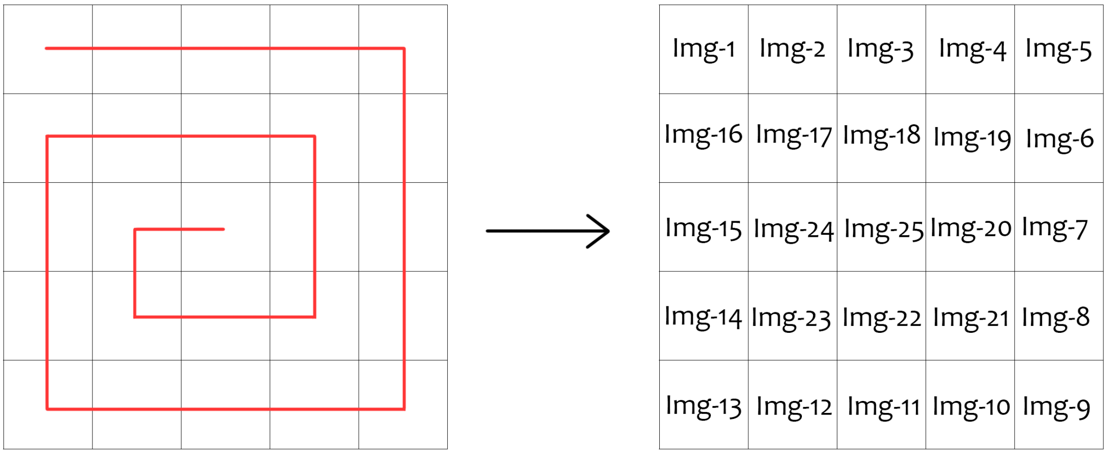
1.1.2 Ranking methodologies

As introduced earlier, ranking (indirect scaling) methodologies present multiple stimuli and ask for the ordering of the presented stimuli in terms of quality. Especially when the number of stimuli presented is low, HVS is quite efficient at ordinal tasks, increasing the reliability of the collected annotations. The main drawback of the ranking tasks is mapping the collected rankings into a quality scale. There are a number of particular methodologies to compensate for this drawback, which will be briefly introduced below.

Pairwise Comparison (PC): is the most popular indirect scaling methodology utilized in subjective image quality assessment. It breaks down the ranking task into small chunks. Observers are presented with a pair of stimuli from the same source image(SRC) and asked to choose which image in the pair has a higher quality. Although the observers might be given the option to state the images in the pair have the same quality in some instances, this is not allowed for two-alternative forced-choice (2AFC). The assumption with 2AFC is that the distribution of preferences will be close to $50\% - 50\%$ for pairs with the same quality.

Subjective annotations acquired with 2AFC experiments are represented in matrices called Paired Comparison Matrix (PCM). PCM is a square matrix with size $k \times k$ where k is the number of images generated from a given SRC. We can display each pair in a cell in PCM_{S_m} for a given SRC S_m. For a given pair P_{ij}, each cell (i, j) contains the number of observers who prefers image i (IMG_i) over image j (IMG_j). Sum of cell (i, j) and cell (j, i) is equal to the number of observers which evaluated the pair P_{ij}. In order to evaluate all pairs within an SRC, $(k/2)(k-1)$ comparisons are required. Note that the order of the images in a pair is redundant. The number of required comparisons increases exponentially with k (*i.e.,* number of HRCs). Comparing all possible pairs is known as Full PC design.

Adaptive Square Design (ASD): is one of the alternative designs developed for PC experiments to reduce the number of required comparisons while preserving the reliability



Img-1	Img-2	Img-3	Img-4	Img-5
Img-16	Img-17	Img-18	Img-19	Img-6
Img-15	Img-24	Img-25	Img-20	Img-7
Img-14	Img-23	Img-22	Img-21	Img-8
Img-13	Img-12	Img-11	Img-10	Img-9

Figure 1.2 – Placement of stimuli on a spiral in square matrix

of the collected subjective annotations[65]. We can illustrate ASD with an example. For a given SRC with 25 HRCs, we can position each HRC in a 5 × 5 square matrix as shown in Figure 1.2. After placing the stimuli onto a square matrix as shown, we compare each neighboring stimuli on each row and column of the square matrix. This procedure significantly reduces the number of required comparisons.

The term adaptive in ASD comes from the fact that the placement of stimuli on the spiral is updated (*adapted*). Ideally, after collecting subjective preferences from each observer, the placement of the stimuli on the spiral is updated. According to the current and previous observers ' preferences, stimuli are rearranged from the highest quality to the lowest. Since prior to the experiment, quality scores of the stimuli were not known, initial placement of the stimuli on the square matrix is either done via previously known information (estimated quality, pre-test, etc.) or just randomly.

Such adaptive design provides a set of pair comparisons with the potential to provide the highest information. Most of the time, comparing the lowest quality image to the highest quality does not provide any information. By omitting these comparisons from the experiment, ASD provides an efficient pair comparison design while keeping the reliability and discriminative power of pairwise comparisons[65].

1.2 Content selection

Selecting source content for the subjective experiment is an integral part of the experiment preparation, and it is crucial to use content that matters for the QoE scenario of the experiment. In the literature handful of features have been frequently used. A brief list of the commonly used features for content selection of 2D images is given below.

Spatial information (SI): is a feature based on the magnitude of edges in spatial domain[120]. It is often used with Sobel kernels [104] and normalized in spatial resolution. Extraction of edges is done on the luminance channel.

Colorfulness (CF): measures the variety and intensity of color information for a given image content[120]. It is defined in RG (Red-Green) and YB (0.5(Red+Green)-Blue) components of an image and relies on colorfulness estimation proposed by Hasler[43].

Image complexity (IC): is a feature based on the compression complexity[124]. It is calculated as the inverse of the lossless compression ratio of an image. Variations are also proposed in the same study.

Dynamic range (DR): is calculated on the pixel level based on the maximum and minimum luminance values for a given image[47]. Prior to calculation, 1_{st} and 99_{th} percentiles of the luminance values are excluded.

$$DR = \log_{10}(L_{max} - L_{min}) \tag{1.1}$$

Image key (IK): is a measure of the average image brightness where values range between zero and one[47]. It is defined as follows:

$$IK = \frac{\log L_{avg} - \log L_{min}}{\log L_{max} - \log L_{min}} \tag{1.2}$$

In addition to the features described above, there are many others used in a number of datasets. For example, uniform distribution of descriptive categories (indoor, outdoor, night, day, etc.) can be ensured to reduce bias towards certain categories.

Defining desired features for the subjective experiment became even more crucial in the last decade with the increasing popularity of learning-based processing tools. Any imbalance in the dataset can affect the performance of the developed models on the dataset. For example, an objective quality metric developed on a dataset with super-threshold distortions may perform poorly when evaluated on supra-threshold distortions.

1.3 Processing subjective annotations

Interpretation of the subjective experiment requires a processing stage after collecting subjective annotations from the observers. It is assumed that, with enough observers, we can estimate the general population's opinion on a stimulus by using a small subset of observers.

For rating tasks, it is often done by simply taking the mean of the collected observer ratings[48]. MOS and DMOS are also used with confidence intervals (CI) to reflect the quality range of the stimuli. In the literature, generally, 95% CI is utilized. For a given stimulus i, CI_i is defined as follows:

$$CI_i = [MOS_i - \sigma_i, MOS_i + \sigma_i] \tag{1.3}$$

where σ is calculated as follows based on the standard deviation (STD) and the number of observers who rated the stimuli i:

$$\sigma_i = 1.96 \frac{STD_i}{\sqrt{N}} \tag{1.4}$$

For a large enough number of observers, MOS sufficiently reflects the general opinion of the target audience. However, it is often challenging to ensure that the number of observers is large enough. For certain tasks, it is practically impossible to recruit large enough participants for the experiment due to higher costs, experiment conditions, etc. A more sophisticated approach, Estimated Population Mean Opinion Score (EPMOS), proposed in recent years to overcome this assumption[84]. EPMOS shows a better estimate of the target population's average perception of the two video quality subjective experiments.

There are two main methods to interpret the PC results. One can directly use PC results as represented with PCM and estimate the statistical significance of the pairwise preferences with statistical tests such as Barnard's [12] or Fisher's [27] exact test. On another front, quality scores of each stimulus can be estimated by mapping the PCM into a continuous scale with models such as Thurston-Moesteller [78, 107] or Bradley-Terry [14]. A detailed overview regarding the interpretation of PC data is given in [112].

Every pair in a PCM can be represented as a 2 × 2 contingency table as below:

$$\begin{bmatrix} T_{A>B} & T_{B>A} \\ T_{B>A} & T_{A>B} \end{bmatrix}$$

where $T_{A>B}$ is the number of observers who prefers IMG_A over IMG_B and similarly $T_{B>A}$ is the number of who prefers IMG_B over IMG_A.

It has been shown that Barnard's exact test is more powerful than alternative statistical tests on 2×2 contingency tables [75]. Therefore, in the rest of this work, we rely on Barnard's exact test for determining statistical significance among the pairs.

1.4 Screening observers

Detecting and rejecting the outliers from the experiment results improves the reliability of the collected subjective annotations. Observer screening for rating methodologies is well covered in Section A1-2.3 in ITU-R BT500.14[48]. The procedure is explained for various rating methodologies, and it is recommended to apply the outlier rejection procedure only once to the collected results. In other words, consecutively applying the procedure may falsely identify honest observers as outliers. It is also suggested that the procedure be used in experiments with less than 20 naive observers. As outlier detection in rating experiments is out of the scope of this thesis, we recommend interested readers to refer to the original document [48] for details.

On another front, there is no well-established methodology in standardization documents for observer screening in PC experiments. Due to the binary nature of the pairwise preferences, identifying outliers is more challenging. The commonly rated number of stimuli, number of observers, and the task's subjectivity highly affect the discriminability of the outliers. Only a few efforts are proposing an outlier rejection model for IQA with PC to the best of our knowledge. Among other tools, PWCMP [89] Matlab package includes an outlier rejection method. However, the authors recommend the proposed model to support the experimenters and leave the outlier detection decisions to the experimenter herself.

1.4.1 Inter-observer reliability measures

There are several metrics (sometimes called inter-rater reliability) to measure the reliability of an observer based on the experiment procedure. While some methods provide a reliability measurement for the whole experiment, some provide a measurement for each observer, whereas some provide a similarity rating for each observer pair. We briefly introduce some of the related methods for inter-observer reliability.

Rank correlation coefficients: are the most commonly used method to determine observer reliability in rating tasks. Correlation coefficients such as Spearman [105] or Kendall [55] can be used to determine how similar an observer's ratings are to the rest of the observer. Generally, for a given observer, mean correlation with the rest of the observers is used to determine the reliability. It is not possible to utilize these measures in PC experiments (or ranking experiments in general).

Cohen's kappa: is another popular choice for inter-observer reliability [19]. It is designed for categorical rating experiments, and therefore, the usage is limited in image quality assessment. It provides values in the $[-1, 1]$ range where negative values indicate no agreement and higher values indicate a greater agreement among observers. Although it can be used for pairwise comparison experiments, its performance is questionable for test scenarios with high subjectivity[7].

Krippendorff's alpha: is a flexible metric that works with various data types (binary, categorical, ordinal, etc.) without a sample bias[61]. It can also handle missing data. It provides an agreement value for the tested population, and higher values indicate a greater agreement among observers within the population. While it can be used as an indicator of unreliable data collection, it is not a standalone outlier detection tool[7].

1.4.2 Reliability checks in crowdsourcing

Crowdsourcing experiments require additional attention for observer screening due the to absence of moderation and uncontrolled experimental conditions. The anonymity of the observers may lead lack of attention and cheating. Although screening observers by inter-observer reliability measures (see Section 1.4.1) is one way to identify such behaviors, reliability checks can be included in experiment design. *Reliability check* is a broad term that covers many approaches such as consistency checks, content-based attention checks, verification tests, golden units, vote patterns, and vote speed checks.

Verification tests: are basic captcha-like questions. It can include simple math problems as "what is 4 plus 4?" or simply an off-the-shelf captcha implementation can be used. It helps to identify spammers who rely on automated software(bots) to do the task for them.

Content based attention checks: are questions related to the content of the displayed image/video. Questions such as "What color was the building in the last image?" can be used to check how attentive the user is to the experiment. Questions need to be carefully selected and should be relatively simple in order to prevent false identifications.

Consistency checks: aims to determine observers' attention by repeating a few stimuli throughout the experiment. Observers are expected to give the same response to repeated questions at each time. However, familiarization with the stimuli/task may alter observers' opinions towards the end of the experiment.

Vote speed check: can be used to identify spammers who rapidly finish the task. It requires the subjective annotations to be recorded with time stamps. The challenging side of this check is to determine the threshold for a "too fast" annotation. Depending on the task, this threshold may vary significantly.

Vote pattern check: measures how biased towards a particular vote each observer is. In PC experiments, one can check if an observer is repeatedly voting for the same position (*e.g.,* left or right in left/right presentation) despite the stimuli. Similarly, finding a threshold for unreliable behavior depends on the number of stimuli in the experiment. Probabilistic approaches can be used to determine such thresholds[6].

Golden units: relies on stimuli with the answer known prior to the experiments. In a pairwise comparison setup, a heavily distorted stimulus can be shown alongside a pristine image, and incorrect answers may be used to identify spammers[6].

1.5 Crowdsourcing Subjective IQA

Crowdsourcing gained popularity over the last decade to outsource laboratory experiments to a wide range of audiences via the internet. It allows to reach a diverse participant pool and provides fast turnover of large-scale experiments for a reduced cost. However, transferring IQA experiments to crowdsourcing platforms is not a straightforward process[50, 31, 22]

In crowdsourcing, the attention span of the participants is much lower compared to laboratory experiments. Due to uncontrolled environmental conditions and the absence of moderation, participants are more likely to get distracted during the experiment. Therefore, shorter test duration and simplification of the tasks are recommended to increase the reliability.

The motivation of the participants also brings complexity for crowdsourcing experiments[30]. Volunteers provide more reliable answers overall but are less likely to finish the experiment. Participants may be motivated to maximize their profit and consequently minimize their time and effort during the experiment[44].

Technical limitations of crowdsourcing are also an influencing factor for subjective experiments. Most crowdsourcing experiments are conducted on browsers. In fact, relying on third-party applications or asking participants to download additional software is not recommended[45]. One of the major technical limitations, however, is the uncontrolled experimental conditions. For example, conducting contrast threshold experiments showed that measured contrast thresholds were higher in crowdsourcing experiments than the laboratory experiment[98].

Many limitations lead to unreliable data collection in crowdsourcing. Lower reliability can be overcame by proper experiment design and observer screening[44]. Reliability checks can be incorporated during the experiment design. Inter-observer reliability methods can be used to analyze observer behaviors after the experiment. Experiment parameters such as task length and monetary compensation can be optimized[18]. After improving experiment design and incorporating analysis tools into the processing stage, pilot studies shall be conducted to determine the possibility of transferring the experiment from laboratory to crowd[33]. In the end, not all QoE tasks are the same, and each task needs to be addressed individually. Methodologies and tools may increase the reliability for one QoE task whereas harming some others.

1.6 Quality Assessment of Tone Mapped Images

Many aspects regarding experiment design, processing tools, and observer screening methodologies heavily depend on the QoE scenario being tested. Since the first part of this thesis focuses on subjective quality evaluation of tone mapped images as the QoE scenario, we introduce the HDR images and tone-mapping operators to provide the necessary background.

1.6.1 High dynamic range imaging

An image's dynamic range(DR) is measured as the log of the difference between the maximum and minimum lightness values. High Dynamic Range (HDR) images have higher differences between their brightest and the darkest points than traditional 8-bit per channel images. The typical DR of a real-world scene is around 1:10000 (even higher for scenes containing direct light source), whereas most of the imaging systems use a ratio of around 1:100 due to 8-bit limitations.

HDR images have gained popularity over the last decades, thanks to advancements in

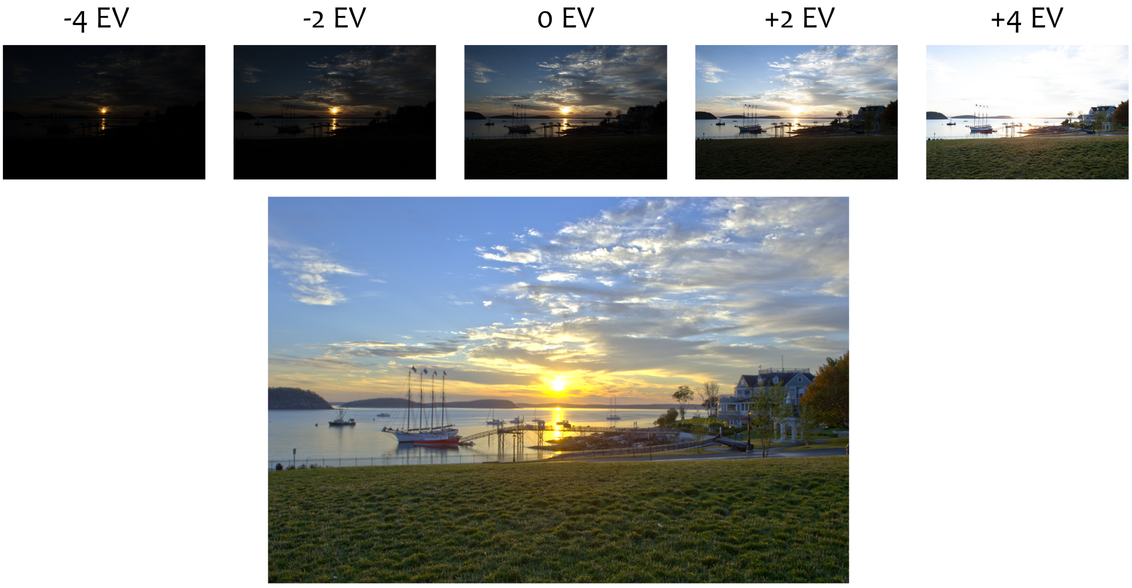


Figure 1.3 – HDR bracketing example from 5 images with varying exposure settings.

image acquisition methods and display devices. Improvement in DR capabilities of recent cameras or exposure bracketing techniques allows us to capture HDR images represented by 16-bit or 32-bit per pixel per channel values. Utilization of a higher bit rate for a given pixel allows representing colors and luminance values in a much more realistic and convincing way.

Acquisition of HDR images can be made in multiple ways. A decade ago, commercially available camera sensors could not capture higher than 8 bits per pixel/channel. Capturing HDR images was mainly done by stacking multiple images with incremental exposure values. In other words, same scene is photographed several times and each capture uses a different exposure setting*, e.g. under-exposed(-1 EV), normal(0 EV), over-exposed(+1 EV).* After capturing individually, we can merge the images into a single HDR image. Under-exposed images can bring in details from bright regions of the scene, while over-exposed images reveal the darker regions' details. This procedure is also called *bracketing* and an example of it is depicted in Figure 1.3. Five images with varying exposure levels are merged into a single HDR image. The acquired HDR image is tone mapped for visualization purposes. Recent advancements in acquisition technologies made it possible to have camera sensors to capture HDR images in a single shot. Some of the new mobile phone cameras have made 10-bit per-pixel per-channel capturing possible with a single shot.

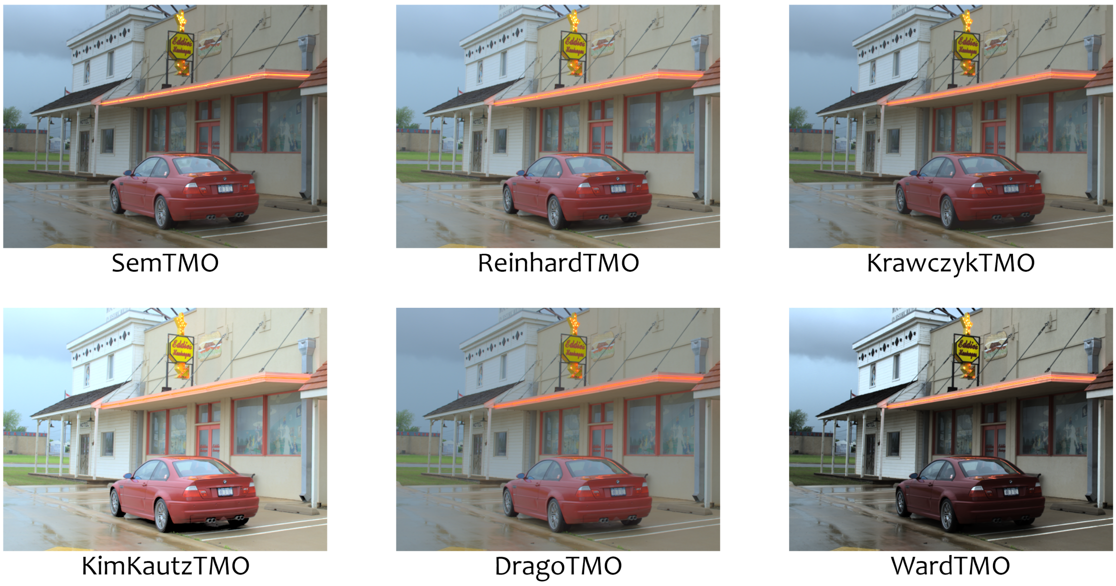
Another critical step in the HDR imaging pipeline is the display. One needs a display device capable of reproducing brighter white and darker blacks to benefit from high dynamic ranges fully. Although HDR displays are commercially available, their adoption is still not widespread, mainly due to the lack of available content. Consequently, HDR images are often converted into 8-bit images prior to display. The functions that allow this conversion is called tone mapping operators, introduced in the next section.

1.6.2 Tone mapping operators

When the DR of the image is higher than the display's DR, mapping the pixel values to the lower DR is necessary. The mappings are done through sophisticated functions called tone mapping operators(TMO). TMOs can be categorized into two main groups as global and local. Global TMOs use the same function to map every pixel in the image into the lower DR. Local TMOs adjust the mapping function based on each pixel and its neighboring pixels. Although the list is not exhaustive, we introduce widely used TMOs in the literature below based on their categories(global/local).

Global TMOs often require less computational power while struggling with challenging scenes due to the loss of details on highlights or shadows. In other words, it maps the HDR image into lower DR by preserving only the detectable contrast levels by the HVS. Drago et al. later proposed a TMO (DragoTMO [21]), which adapts the simplest form of tone mapping (logarithmic tone mapping) by adapting the logarithmic operation based on the pixel luminance. ReinhardTMO [90] mimics the HVS by modeling the photoreceptors with sigmoid different functions. WardTMO[63] is also another TMO that takes benefits from the existing knowledge about HVS. It uses a downsampled version of the image to create a histogram to guide the tone mapping. In another work, KimKautzTMO [56] is proposed based on the log-luminance adaption of the human visual cortex.

To overcome the drawback of global TMOs, which lose details in highlights and shadow areas, local TMOs were introduced. Unlike global TMOs, local TMOs adjust the mapping function based on the pixel statistics of different regions in the image. KrawczykTMO [60] uses a probabilistic model of the lightness perception of the HVS. SemTMO [35] divides the image into a number of semantic regions and adjusts the mapping function based on the statistics of the semantic category.



SemTMO ReinhardTMO KrawczykTMO

KimKautzTMO DragoTMO WardTMO

Figure 1.4 – Output of 6 TMOs for the same scene.

1.6.3 Quality evaluation of tone mapped images

Each TMO provides a relatively different output for the same scene. Some examples are presented in Figure 1.4. Although some are significantly better than others in general, choosing which TMO works better for a particular scene is not straightforward. Choosing the most suitable TMO for a given image requires a thorough evaluation of the image quality to decide. Although there are a number of objective quality metrics to assess tone mapped image quality, their correlations with the subjective opinions are relatively low.

Subjective quality evaluation of tone mapped images may answer different questions depending on the experiment design[58]. The presence of a reference HDR image in the experiment allows measuring the accuracy of the TMO at preserving real-life cues. On the other hand, the absence of reference HDR images provides a purely aesthetic image quality evaluation.

A non-exhaustive list of datasets from the literature for subjective quality assessment of tone mapped images are given in Table 1.1. A more comprehensive review for tone mapped image quality evaluation datasets can be found in [83]. Existing datasets vary significantly in terms of experimental methodologies, and one thing in common is the small number of content used in the evaluation. Due to the expensive and time-consuming nature of laboratory experiments, it is practically challenging to collect a larger dataset

Table 1.1 – Selected tone mapped image quality evaluation datasets from the literature.

Study	**Reference**	**Exp. Procedure**	**Nb of SRC**	**Nb of TMO**
Krasula et al. [58]	HDR & No Ref.	PC	20	5
Ledda et al. [64]	HDR	PC	23	6
Yoshida et al. [123]	Real World	Rating	14	7
Petit et al. [83]	No Ref.	Rating & Ranking	7	4
Cadik et al. [131]	Real World & No ref	Rating & Ranking	3	14

unless the experiment is conducted via crowdsourcing.

To the best of our knowledge, there is only one work on subjective quality evaluation of tone mapped images conducted via crowdsourcing. In their work [62], a subjective experiment was conducted on Amazon Mechanical Turk (AMT [10]) with more than 5000 observers on 605 HDR images with rating methodology. Despite providing the highest number of observations and stimuli in the literature, the dataset is not solely focused on tone mapped image quality evaluation. 4 TMOs, five multi-exposure fusion (MEF) algorithms, and two post-processing effects (Grunge and Surreal) is used to generate the tested stimuli. Each HDR image was only tone mapped with one of the 4 TMOs included in the experiment. Therefore, comparing TMO performances on the same content is not possible from the collected subjective annotations that provide valuable insight for benchmarking existing quality metrics and developing new quality metrics for tone mapped image quality assessment.

Chapter 2

Content Selection Strategy for Subjective Quality Assessment of Tone Mapping Operators with Pairwise Comparison

This chapter proposes a content selection strategy developed for tone mapping image quality evaluation in pairwise comparison experiments. Proposed scheme was briefly introduced in two peer reviewed publications [6, 34] as part of our collaboration with Abhishek Goswami, Wolf Hauser from DxO and Frédéric Dufaux from CentraleSupélec. As part of the same collaboration, a detailed explanation is also given in our recent work which was submitted to IEEE Transaction on Multimedia (currently under review).

Content selection is one of the most crucial steps in developing a desirable dataset for QoE scenarios. Although the methodology may vary according the experiment design and QoE scenario, the ultimate goal is to acquire reliable subjective annotations for a representative set of content. In this chapter, we focus on how to select a representative set of content for tone mapping quality evaluation for pairwise comparison experiments. Challenges and motivation related to content selection are discussed in Section 2.1. Section 2.2 introduces the publicly available HDR image collections. Main contribution of the chapter explained in detail in Section 2.3. Validation of the approach is done in Section 2.4 before concluding the chapter in Section 2.5.

2.1 Challenges & Motivation

Datasets have always played a crucial role in developing image processing tools. In the QoE domain, datasets are often used to develop new objective metrics and benchmark existing approaches for numerous problems. In order to maximize the benefit acquired from subjective experiments, researchers go through a careful experiment design stage. It is crucial to justify how representative the collected data for the QoE scenario being tested. For subjective assessment of image quality, we do not want to evaluate the quality of any image found online. Instead, a desired set of features are defined to select a representative set of images for evaluation.

Furthermore, the last decades brought a surge in the popularity of machine learning models for various computer vision tasks, including objective quality assessment of multimedia content. Training machine learning models often require a large amount of reliable and representative data. For example, a model trained on computer-generated images may perform poorly when used on natural images. This trend towards data-driven models further increased the impact of datasets on the performance of image processing tools.

Like the rest of the subjective experiment design, content selection is also task-dependent. Whereas selecting content for evaluation of compressed algorithms might be done based on uniformly distributed QP levels, it is not straightforward for aesthetic quality evaluation of tone mapped images. Therefore, just like the rest of the experiment design, content selection should be made regarding the targeted use case. Before we discuss the more specific challenges, we will summarize the targeted use case in this work.

This chapter aims to provide a content selection solution for the aesthetic quality evaluation of tone mapped images. Subjective experiment is planned to be conducted on Prolific [88] crowdsourcing platform with pairwise comparison methodology. 4 TMOs (*KimKautzTMO* [56], *KrawczykTMO* [60], *ReinhardTMO* [90] and *SemTMO* [35]) have been selected for generating tone mapped stimuli from 250 source images (SRCs). The required resolution for each tone mapped image is 640 × 480 px to allow a side-by-side presentation on a 1080p display device. Publicly available HDR image collections (introduced in Section 2.2) provide 3840 × 2160 px spatial resolution. Therefore, cropping and/or down-scaling is necessary to acquire the desired resolution. With this subjective experiment, we aim to collect a large number of annotations to provide sufficient data for developing learning-based objective quality metrics for TMO quality evaluation and providing a challenging benchmark for existing quality metrics. More detail regarding the subjective experiment and collected subjective annotations is provided in Chapter 4.

To the best of our knowledge, publicly available HDR images are limited in the literature. We used two datasets containing 229 HDR images to generate desirable content for the subjective experiment. As discussed in the summary of the experiment details above, desired spatial resolution for the stimuli used in the pairwise comparison is 640 × 480 px. Most of the collected HDR images have the spatial resolution of 3840 × 2160 px. The mismatch between the original HDR image resolution and targeted resolution poses both an opportunity and a few challenges. It provides an opportunity to increase the number of SRCs by cropping high-resolution HDR images into target resolution. We can generate thousands of HDR crops from a single HDR image with the given spatial resolutions with a sliding window. The problem arises with the selection of crops from the thousands generated. With small stride values for the sliding window, we generate exponentially more crops; however, we end up with highly similar crops. In some instances, small shifts in composition affect the quality of the tone mapped image considerably. Moreover, due to the significant difference between the original and target image resolution, crops may have a composition containing only a sky patch, grass field, or a building facade. Such unnatural composition neither provides a challenging scene for tone mapping operators nor reflects real-life scenarios. Inevitably, we ask the following question:

"Which HDR crop is better?"

We expect HDR crops to be framed naturally and look like whole images rather than a crop of another image as it helps to create an aesthetic expectation for the observers. HDR crops should also provide challenging scenes for tone mapping. When compared as a pair of tone mapped images, the answer should not be evident for all the pairs. The ambiguity of the tone mapped image pairs should vary. In other words, whereas some pairs are easy to compare (one image in the pair is highly preferable), the dataset should contain also contain difficult pairs (both images in the pair are preferable more or less equally). Moreover, the overall image quality of the tone mapped images acquired from the HDR crops should be above a certain level. Images with heavy distortions are not suitable for aesthetic quality evaluation. Based on these observations, the rest of the chapter aims to formulate our expectations and provide a modular content selection pipeline for the given use case.

2.2 HDR Image Collection

As previously discussed, publicly available HDR image datasets in the literature are limited in number. Fairchild [24] (105) and Artusi [11] (124) datasets have a total of 229 high resolution HDR images. HDR images are generated with seven exposure brackets with 4300 × 2800 px spatial resolution. Each exposure bracket is shot individually with a

Figure 2.1 – Sample HDR images collected for cropping.

DSLR camera. Figure 2.1 presents a sample set of HDR images (tone mapped manually for visualization purposes) from the two datasets. Since the resolution of images required for the subjective experiment is much smaller than the resolution of the original HDR images, multiple crops were generated from the collected 229 HDR images.

2.3 Content Selection Modules

The proposed content selection strategy contains three sequential modules. A general diagram of the proposed scheme is depicted in Figure 2.2. First, collected high-resolution HDR images were cropped with a sliding window to generate candidate crops. Secondly, defined features were extracted from all candidate crops and combined into a cumulative score. Multiple crops with less than %60 overlap were selected from each original HDR based on the cumulative score. Finally, selected HDR crops were clustered in a three-dimensional space where each one of the three axes represents the TMQI score of a tone mapped version. Details regarding each module are given below in their corresponding subsections.

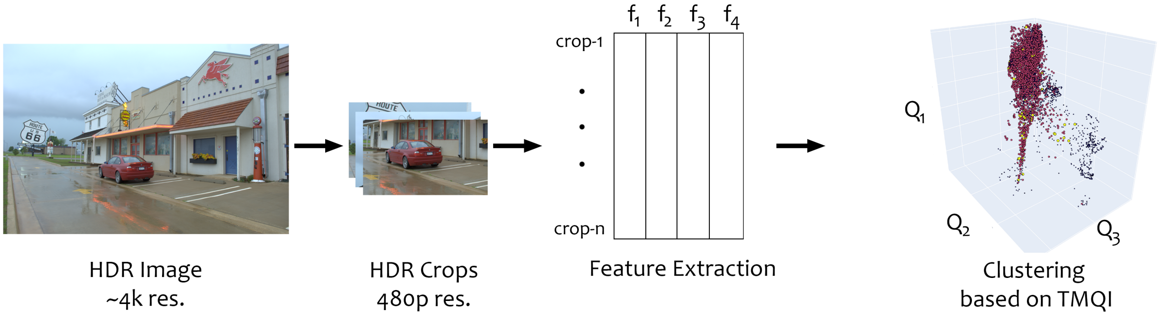


Figure 2.2 – General diagram of the content selection strategy.

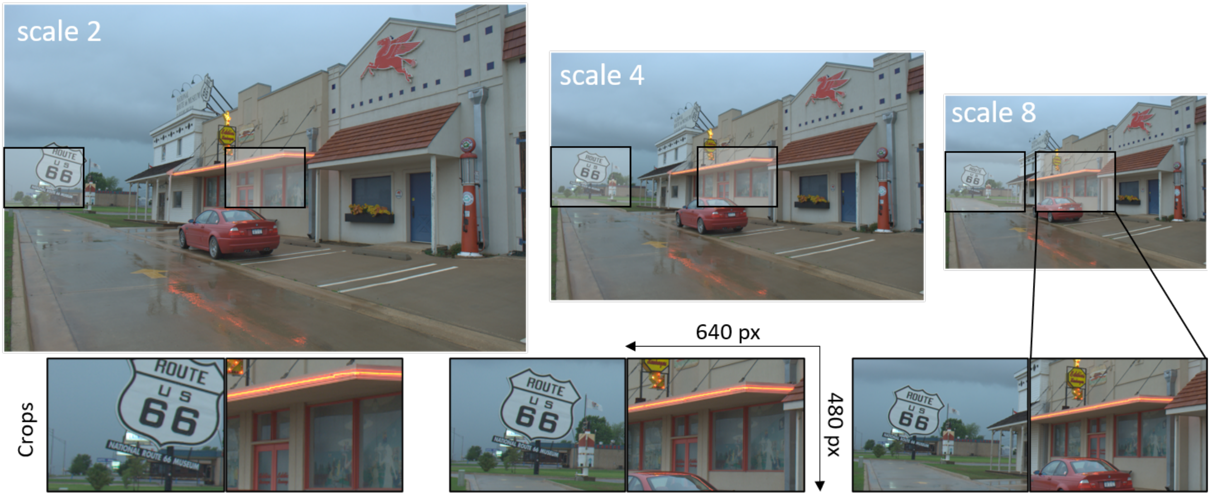


Figure 2.3 – Cropping with a sliding window (100 px stride) on 3 different scales.

2.3.1 Scale down & crop

Original HDR images were scaled down two times, and crops were generated with a sliding window of 100 px stride across all three scales. Scaling down the original HDR images at various rates helps generate diversity on the crops' framing. An example of three scales and sample crops from each scale is depicted in Figure 2.3. Since the target resolution of the crops is 640 × 480 px, over 1000 candidate crops can be extracted from each original HDR image. We end up collecting 167100 candidate crops in total from 229 original HDR images.

2.3.2 Feature extraction

A non-exhaustive list of image features commonly used for content selection in image quality assessment datasets was previously introduced in Section 1.2. A subset of these features was used for the proposed content selection. Below we provide a detailed explanation of each feature and the motivation behind their usage.

Adaptive Dynamic Range(r) and Standard Deviation(d)
Adaptive Dynamic Range (ADR) is defined as the ratio of the brightest point's pixel value to the darkest point's pixel value, and greater difference results in high ADR values. Similarly, the Standard Deviation (SD) of luminance map pixel values quantifies the variation on the crops. 10 and 90 percentile of the pixel values were used for the calculation of both features. Both features help to filter-out crops with dominantly flat textures such as grass, sky, or building facades.

Multi Level Entropy of the Saliency Map(m)
Saliency maps can provide intuition about how interesting and informative an image is. Multi-Level Entropy(MLE) [130] of the saliency maps generated by minimum barrier saliency detection algorithm [128] was used for quantification. Crops with salient regions, *i.e., without a uniformly distributed saliency map* were given a higher preference.

Mean(μ_O) and Variance(σ_O) of Objective Quality Scores
In order to promote crops with higher image quality, three state-of-the-art TMOs were used on candidate crops. Specifically, each HDR crop was tone mapped with *KimKautzTMO* [56], *KrawczykTMO* [60] and *ReinhardTMO* [90]. TMQI [122] was used to calculate the quality score of each tone mapped image version.

Extreme distortions on tone mapped images can be inevitable on certain crops for TMOs without manual tweaking. The arithmetic mean of the TMQI scores was used to promote HDR crops with the higher quality tone mapped images, and this allows to filter-out crops with the highly distorted tone mapped images.

Additionally, the difference between TMQI scores of each tone mapped image is calculated for each HDR crop. This promotes HDR crops, which are difficult for certain TMOs while easy to handle for the rest. This is also a good indication of a challenging to tone map HDR image. Thus, we promote a greater variety in tone mapped image quality while keeping the overall quality of the tone mapped images high.

Calculating Crop Scores
Extracted features do not have the same range. Thus, after extracting features from all 167100 candidate crops, each feature is normalized into the $[0, 1]$ range. Normalized

Algorithm 1: Crop selection based on Q_c and overlap ratio

```
Sort N candidate crops from an HDR image by their crop score $Q_c$;
Start looping over all crops starting from the one with the highest $Q_c$;
Initialize $S$, an empty list of selected crops;
while $i<N$ do
    if Overlap ratio is lower than %60 with other selected crops in $S$ then
        Add candidate crop to selection;
    else
        if Overlapping crops has the same scale then
            Skip the crop;
        else
            Add lower scale crop to the selection and remove higher scale crop
              with lowest score among the overlapping crops in $S$;
        end
    end
end
Result: Return selected crop list
```

features then linearly combined into a *crop score (Q_c)* as follows:

$$Q_c = \hat{r} + \hat{d} - \hat{m} + \hat{\mu_O} + \hat{\sigma_O}, \tag{2.1}$$

Calculated crop scores are stored to be used in the following module.

2.3.3 Selecting crops based on custom score and overlap ratio

Crops obtained from similar locations in an HDR image may have similar crop scores (obtained by equation 2.1). Consequently, selecting crops based solely on the crops score may end up with selected crops with great overlap due to low stride value (100 px). Therefore an empiric threshold of %60 is set to reject any two crops from the same HDR to reduce redundancy. In other words, any two selected crops from the same HDR content are not allowed to have more than %60 overlap. The procedure of eliminating crops based on their Q_c and overlap ratio is summarized in Algorithm 14. This procedure allows us to select candidate crops with minimal redundancy between them. In total, 19540 crops were selected among 167100 candidate crops acquired from 229 HDR images.

2.3.4 Clustering based on objective quality score

Selecting 19540 crops based on the crop score and overlap ratio provides a set of crops with desired quality and no redundancy. 250 SRC images are required for the subjective experiment. Therefore, at this stage, selected crops will be further narrowed down to 250 while promoting pairs with variety in ambiguity. Certain HDR images may be favorable for certain TMOs, and acquired tone mapped image pairs from these SRCs may have low ambiguity (trivial pairwise comparisons). Similarly, certain HDR images may be challenging or easy to tone map. Tone mapped image pairs acquired from these types of SRCs provide high ambiguity (difficult pairwise comparisons). Based on this intuition, we can generate pairs with varying ambiguity if we can classify SRCs into different categories regarding the pair difficulty of the resulting tone mapped images.

To do so, we rely on TMQI [122] scores of tone mapped images. We use the predicted image quality of the tone mapped images to understand HDR difficulty. We tone map each selected HDR crop with three tone mapping operators; ReinhardTMO [90], KimKautzTMO [56], KrawczykTMO [60]. TMQI scores of each tone mapped image were calculated. Then, HDR images can be represented in a 3D space where TMQI scores of the three tone mapped images are each indicated on one of the three axes. It allows us to cluster each SRC into one of the five categories as follows:

- Easy to tone map: TMQI scores of all three tone mapped image is high.
- Difficult to tone map: TMQI scores of all three tone mapped image is low.
- Easy to tone map for ReinhardTMO, difficult for the rest
- Easy to tone map for KimKautzTMO, difficult for the rest
- Easy to tone map for KrawczykTMO, difficult for the rest

After clustering the HDR image crops into described categories, we randomly pick an equal number of crops (50) from each category. This ensures a balanced dataset in terms of TMO performances and a nearly uniform distribution in pairwise comparison difficulty.

2.4 Validation of the proposed strategy

We can validate the initial modules by subjectively inspecting selected crops. Most crops provide a natural framing and a high variety in local brightness within the crop. 8 sample crops among the final selection of 250 is presented in Figure 2.4. In order to evaluate the variety of pair ambiguity, the distribution of pairwise preference' percentages is plotted in Figure 2.5. The vertical axis represents the number of pairs for each bar. The

Figure 2.4 – Sample crops selected with the suggested content selection strategy

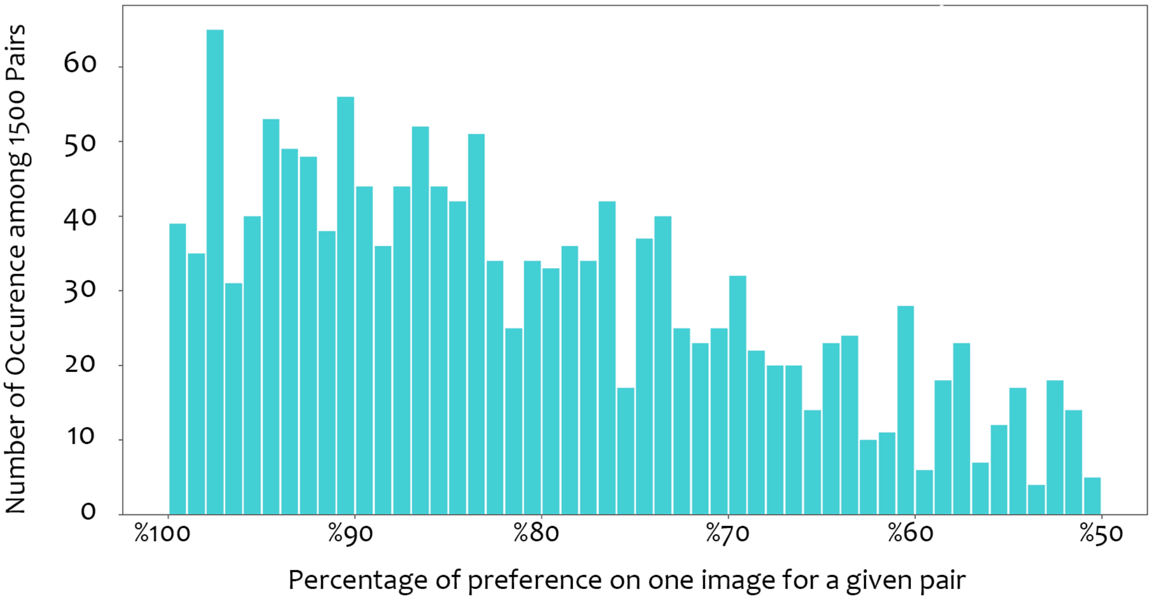


Figure 2.5 – Distribution of pairwise preferences.

horizontal axis represents the pairwise preference percentages. %100 represents the pairs in which every observer preferred the same image while %50 is the pairs where half of the observers prefer one tone mapped image while the other half prefers the opposite. In other words, the ambiguity of the pairs increases from left to right on the horizontal axis. We can observe that a balanced distribution of ambiguity exists in the dataset despite being not perfectly uniform.

2.5 Discussion

The proposed content selection strategy aims to generate HDR crops with 480p resolution from collected HDR images with roughly 4k resolution. Initially, we generated crops with a sliding window on the collected 229 HDR images. This allowed us to generate over 167100 crops. Since no consideration has been made on the quality of the collected crops, most of them were not usable. We implemented two modules that are dedicated to generating natural-looking crops with minimal redundancy in between. A set of features was collected and normalized to generate crop scores. Crop scores were assumed to handle the problem of cropping images with natural framing and content while preserving HDR images, which are challenging for tone mapping. 19540 crops were filtered by maximizing crop scores and minimizing the spatial overlap between crops. Next, we hypothesize that we can select crops with varying difficulty by clustering selected HDR crops in three-dimensional space with TMQI scores laying on the axes. High variance in the ambiguity of pairs is beneficial for developing new objective quality metrics, especially for data-dependent learning-based models. We validated the approach via the pairwise preference distributions.

Chapter 3

Reliability of Crowdsourcing for Subjective Quality Evaluation of Tone Mapped Images

This chapter is dedicated to understanding the challenges involved in using crowdsourcing platforms for the subjective assessment of image quality. Specifically, we will focus on quality evaluation of tone mapped images. We conducted three experiments with varying experimental conditions to understand the impact of the experiment platform and participant recruitment methodologies. The results of this chapter were previously published [34] mainly at IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP 2021, Best Paper Runner-Up) and partially [6] at the Image Quality and System Performance conference in Electronic Imaging Symposium 2020 (EI-IQSP 2020) as part of our collaboration with Abhishek Goswami, Wolf Hauser from DxO and Frédéric Dufaux from CentraleSupélec.

We discuss the challenges of subjective assessment of image quality on crowdsourcing platforms and the motivation behind our work in Section 3.1. Details regarding to subjective experiments are given in Section 3.2. The result of our extensive analysis are shared in Section 3.3 before the concluding the chapter in Section 3.4.

3.1 Challenges & Motivation

As described in Chapter 1, crowdsourcing brings a new set of challenges while providing many advantages. Previous works such as ITU [50] and Qualinet [44] technical reports address many of the challenges and provide recommendations towards transferring QoE experiments conducted in laboratory environments to crowdsourcing.

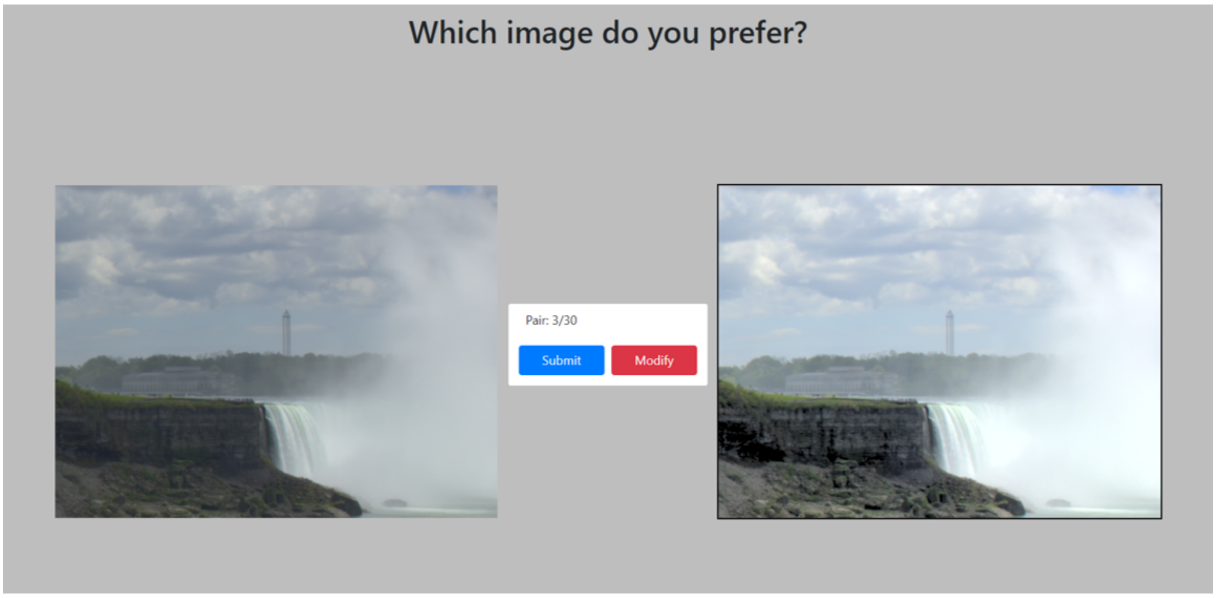
Some of the challenges introduced with crowdsourcing can be dealt with at the experiment design stage. The experiment can be split into shorter chunks (i.e., playlists) to overcome the lack of attention due to the lack of moderation during the experiment. The subjective task can be simplified by using more straightforward methodologies such as pairwise comparison. A certain set of data can be collected during the experiment to identify suspicious behaviors such as observers who are too fast (Section 5.2.1), observers vote by following a specific pattern (Section 5.2.2) or observers with lack of attention (golden units, Section 5.2.3). Although such efforts increase the reliability of the collected subjective annotations, ensuring the experiment will provide a desirable output is necessary.

A straightforward way to determine the significance of the difference between crowdsourced and laboratory experiments is by conducting the same experiment with minimal changes at both platforms. Comparing the subjective annotations collected from two platforms can reveal the effect of crowdsourcing platforms. Therefore, we designed three experiments with minimal differences on subjective quality assessment of tone mapped images. All experiments share the same stimuli, and the only difference between the experiments is the experiment platform and participant pool.

The first experiment(Exp-Lab) was conducted in laboratory conditions at IPI, University of Nantes. Participants were recruited through the IPI mailing list of naive users. The second experiment (Exp-Online) was conducted online via a browser on participants' own devices and desired environments. Participants were recruited through the IPI mailing list. Finally, the last experiment (Exp-Prolific) was conducted on Prolific [88] crowdsourcing platform with platforms' participant pool. By comparing the Exp-Lab with Exp-Online, we can determine the effect of uncontrolled experiment conditions, whereas comparing the Exp-Online vs. Exp-Prolific results can inform us on the reliability of the Prolific participant pool. Moreover, a conclusion can be drawn from the comparison between Exp-Lab and Exp-Prolific regarding the reliability of Prolific crowdsourcing platform in tone mapped image quality assessment scenario.

By comparing the three experiments, we seek an answer to the following questions:

- What are the effects of experimental conditions and participant recruitment methods on subjective preferences?
- Can crowdsourcing platforms be used for TMO evaluation without compromising on the gathered data?
- What is the required number of observers on Prolific to achieve the same level of certainty with laboratory experiment



Which image do you prefer?

Figure 3.1 – Screenshot of the test screen.

3.2 Subjective Experiments' Design

In this section, we will introduce the three experiments designed for comparison of crowdsourcing and in-lab experiments. All three experiments share the same stimuli and subjective task whereas the recruitment and experiment platforms differ between the experiments. QoE task of the experiments are tone mapped IQA. Three experiments are named as *Exp-Lab*, *Exp-Online* and *Exp-Prolific*.

3.2.1 Experiment setup & procedure

For all three experiments, a no-reference pairwise comparison of tone mapped images is used. In other words, two tone mapped images were presented to observers side by side on a single display. Participants were asked to choose the preferred one among the two. An example test screen is presented on Figure 3.1.

Although they simplify the task for observers, pairwise comparison experiments require a higher number of subjective annotations for the same number of content compared to rating tasks. Cross content comparisons and a higher number of testing conditions increase the required number of pairs exponentially. Adaptive designs (see Section 1.1.2 for more details) can be used to reduce the number of required comparisons instead of full PC design. Moreover, cross-content comparisons may not be relevant depending on the QoE

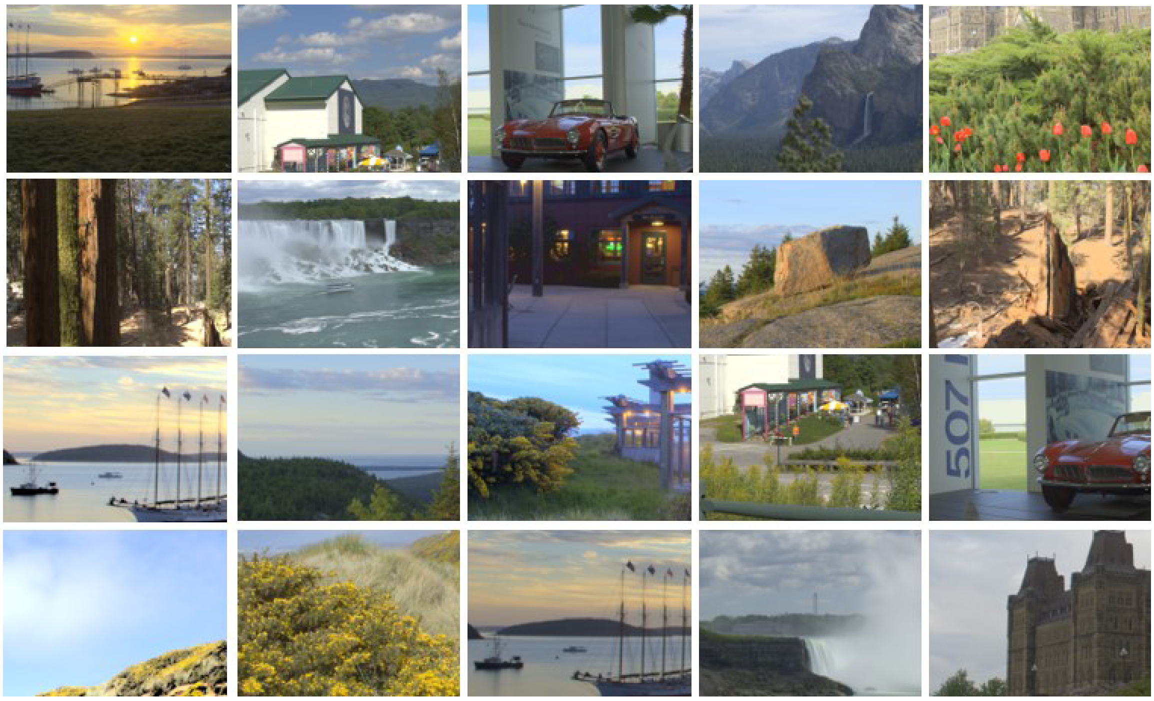


Figure 3.2 – HDR images used for the experiments. Displayed images are tone mapped for visualization purposes.

scenario. Since it is impossible to adopt adaptive designs on the Prolific platform due to lack of API support and cross-content comparisons are irrelevant for tone mapped image quality evaluations, we adopted a full PC design without cross-content comparisons for all experiments. In other words, each possible pair of tone mapped images acquired from the same HDR image are compared in all experiments.

3.2.2 Stimuli & database

20 HDR images were used to generate the tone mapped stimuli. HDR images are generated from Fairchild's HDR dataset [24]. Due to the high spatial resolution (around 4k) of the images in Fairchild's HDR dataset, the cropping strategy described in Chapter 2 is utilized to acquire suitable crops for the experiments. Selected HDR images are shown in Figure 3.2 and they have a spatial resolution of 640 × 480. It allows a side-by-side presentation on 1080p display devices.

4 TMOs were chosen from the literature: *KimKautzTMO* [56], *KrawczykTMO* [60], *ReinhardTMO* [90] and *SemTMO* [35]. 20 HDR images were tone mapped with each

TMOs to create four tone mapped images from each. Tone mapped images generated from an HDR image used to generate pairs: 6 unique pairs per SRC, 120 pairs in total. Cross content comparisons (comparison of two tone mapped images generated from two different HDR images) were not included in the experiment as it does not provide information regarding aesthetic quality evaluation of TMOs.

3.2.3 Experiment Platforms

As previously stated, three subjective experiments differ only in terms of experiment platform and recruitment methodology. *Exp-Lab* was conducted within a controlled laboratory environment at the University of Nantes, and participants were recruited through the local mailing list of naive observers. Experiment conditions were set as recommended by ITU-R BT.500-14 standards [48]. Grundig Fine Arts 55 FLX 9492 SL is used to display the image pairs side-by-side. In total, 40 participants, 22 female and 18 male, who are not experts in the image quality domain were recruited. The average age of the participants was 33.5 years. Each participant was checked for visual acuity with the Monoyer test and color perception with the Ishihara test. Each participant submitted his/her pairwise preferences for all 120 pairs in the dataset with a break after the 60^{th} pair. The average time taken per pair was 7.49 seconds for an observer.

The second experiment, *Exp-Online*, was conducted online with each participants' own display device in their desired viewing environments. The same mailing list used for the *Exp-Lab* experiment is used for recruiting the participants. Display devices allowed in the experiment were limited to devices with 1080p resolution and *Windows* operating system. 50 observers, 28 female and 22 male, were recruited in total. The average age of the participants was 22.6 years. Due to the lower attention span of participants in online experiments [31], we split the initial dataset into four playlists with 30 comparisons in each. Each participant was asked to complete all four playlists at their own pace without limiting the breaks between the playlists. The average time taken for an observer was 4.33 seconds per comparison.

The third experiment, *Exp-Prolific*, was conducted on Prolific [88] crowdsourcing platform. Unlike the first two experiments, observers were recruited through Prolific's participant pool. 400 participants, 116 female and 284 male, from more than 20 countries were recruited with an average age of 28.5 years. Similar to the *Exp-Online* experiment, four playlists of 30 pair comparisons were used for the experiment. 100 unique observers evaluated each playlist. The average time spent per comparison was 3.64 seconds. Table

Table 3.1 – Observer statistics

	Number of Unique Obs.	Mean Age (Years)	Gender Female / Male	Avg. Time Per Comparison (Seconds)
Exp-Lab	40	33.5	22 / 18	7.49
Exp-Online	50	22.6	28 / 22	4.33
Exp-Prolific	100	28.5	116 / 284	3.64

Table 3.2 – Summary of the platforms and recruitment details.

	Platform	Recruitment Pool	Nb. of observers per stimulus
Exp-Lab	Laboratory (Controlled)	IPI mailing list	40
Exp-Online	Online (Uncontrolled)	IPI mailing list	50
Exp-Prolific	Online (Uncontrolled)	Prolific	100

3.1 summarizes the demographics of conducted experiments.

3.3 Comparison of Crowdsourcing vs In-Lab

As introduced earlier, three experiments were conducted with the same design and stimuli. The differences among the three experiments are the platforms used and the recruitment methods. These differences are summarized in Table 3.2. By comparing *Exp-Lab* with *Exp-Online*, we can analyze the effect of experimental conditions (*i.e. controlled vs uncontrolled*) on the collected subjective preferences. Furthermore, by comparing *Exp-Prolific* results with *Exp-Lab* and *Exp-Online* we can understand the effect of both recruitment methods and the experiment platforms. To do so, we first evaluate the similarity between the pairwise preferences among the three experiments. Then, we investigate the inter-observer agreement for each experiment. Finally, we use the permutation test to quantify the effect of the number of observers on the certainty of the collected pairwise preferences. The following subsections go into details of these evaluations and discuss our findings.

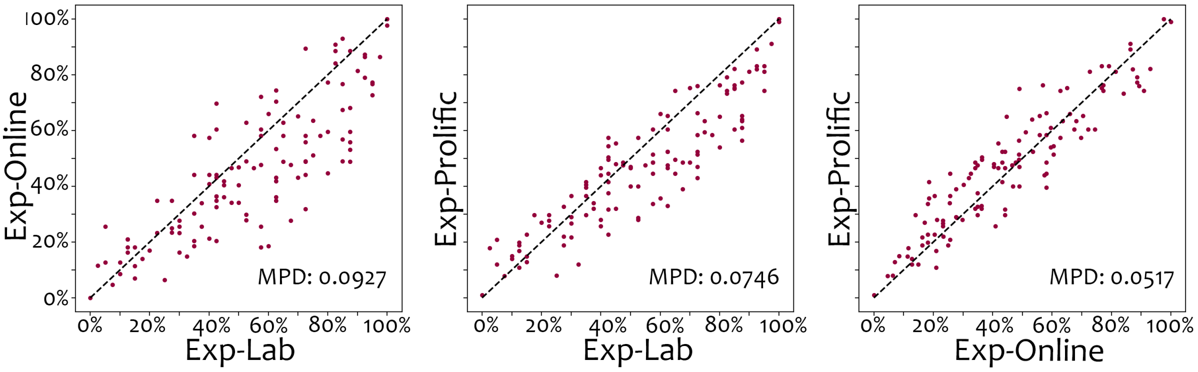


Figure 3.3 – Scatter plot comparisons of pairwise preferences. Each point represents a pair from the dataset. Axis values represent the percentage of votes for the same image in a pair. MPD is the mean of the perpendicular distances of the points from the diagonal.

3.3.1 Pairwise preference similarity between experiments

In this section, the similarity between the experiments is analyzed in terms of collected pairwise preferences. Pairwise preferences can be expressed in terms of percentages. For a given image pair P_{AB}, the percentage of observers who prefers image A (I_A) over image B (I_B) can be used to quantify the quality of I_A in comparison to I_B. Furthermore, Barnard's exact test [12] can be used to determine the statistical significance of the differences between I_A and I_B. For a given pair P_{AB}, number of observers who prefer I_A and I_B, and inversely I_B and I_A is arranged symmetrically on diagonals of a 2×2 matrix as an input to Barnard's exact test. Consequently, we can estimate the statistical significance of the quality difference between I_A and I_B with 95% confidence. This allows us to represent the statistical significance of the difference of a pair as a binary value, *i.e., there is a significant difference or not.* For pairs with a significant difference, we can also identify the better and worse image.

Relative comparison of pairwise preferences:

Initially, we analyze the similarity between the pairwise preferences in terms of percentage of preferences over the same image in a pair, *i.e., the percentage of observers who choose I_A over I_B for a given P_{AB}* . Figure 3.3 presents the result of the analysis. Each plot compares the percentage of observer preferences between two experiments indicated on the axes. Points in each plot correspond to one of the image pairs among the 120 in the dataset. In the case of a perfect agreement between the two experiments' results, each pair should lie on the diagonal. With this intuition, we calculate the Mean Perpendicular

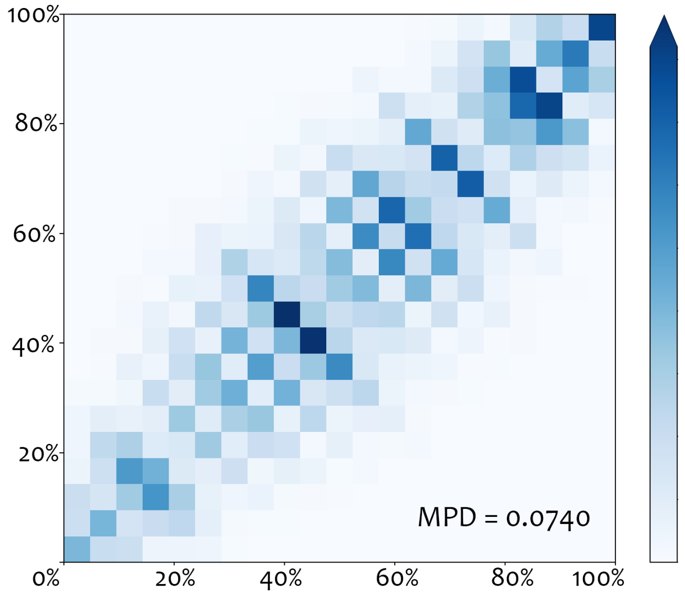


Figure 3.4 – Pairwise preference baseline acquired through 1000 permutations of randomly split halves from the Exp-Lab experiment. MPD value represents the mean perpendicular distance across all permutations.

Distance (MPD) of each point as the mean value of the distances of all points to the diagonal between two experiments. As a result, we quantify the similarity of the collected pairwise preferences between the two experiments. Smaller MPD values indicate a better agreement between the corresponding experiments. MPD values between each experiment are also reported on the plots.

Based on the described criteria for the evaluation, we observe that the distribution between *Exp-Lab* and *Exp-Prolific* experiments are scattered closer to the diagonal in comparison to the *Exp-Lab* and *Exp-Online* plot, which indicates a higher similarity for the former. Similarly, MPD values provide the same conclusion with a lower MPD of 0.0746 between *Exp-Lab* and *Exp-Prolific* compared to MPD of 0.0927 between *Exp-Lab* and *Exp-Online*. Interestingly, we observe even higher similarity between Exp-Online and Exp-Prolific results. This indicates the lesser effect of recruitment methodology on the pairwise preferences in comparison to experiment platforms.

Creating a baseline for pairwise preference comparison:

Although relative comparison of the experiments provides an insight into the effect of experiment platform and recruitment methods, we use a permutation test to determine an expected MPD value. To do so, we split the observers from *Exp-Lab* experiment into two

Table 3.3 – Similarity of the pairwise preferences between experiments with respect to statistically significant differences.

	Agreement	Disagreement	Contradiction
Exp-Lab vs Exp-Online	73	38	9
Exp-Lab vs Exp-Prolific	89	27	4
Exp-Online vs Exp-Prolific	89	31	0

disjoint halves and compare the pairwise preferences between halves. This step is repeated for 1000 iterations, and the average MPD value is calculated as 0.0740. Distribution of the pairwise preferences plotted as a heat map on Figure 3.4. Darker color indicates a higher occurrence. As reported earlier, MPD value between *Exp-Lab* and *Exp-Prolific* is computed as 0.0746, suggesting a desirable similarity of the pairwise preferences between *Exp-Lab* and *Exp-Prolific* when compared to the calculated baseline.

Agreement on the significance of the statistical differences:

For each experiment, the statistical significance of the difference between the two images in each pair is calculated with Barnard's exact test. For each pair, statistical significance results from the three experiments were compared. Table 3.3 presents the findings of this comparison. Each row compares the result of indicated experiments. *Agreement* column represents the number of pairs where both experiments provide the same Barnard's test results, in other words, where both experiment results indicate (or both experiments do not) a statistically significant difference for the pairwise preference of the image pair. Conversely, *disagreement* value represents the number of pairs (P_{AB}) where one experiment finds I_A significantly better over I_B, whereas the other experiment shows the exact opposite, i.e., I_B significantly better over I_A. *Disagreement* value shows the number of pairs where only one of the two experiments indicates a statistically significant difference. Predictably, we observe a similar outcome with the previous analyses. Similarity of *Exp-Lab* and *Exp-Prolific* results is higher than the similarity of *Exp-Lab* and *Exp-Prolific* results.

To sum up, we evaluated the agreement among the experiments in terms of pairwise preference similarities. Relative comparison of pairwise preferences with scatter plots provided insight regarding the effect of the recruitment procedure on the collected data. We also calculate an expected baseline MPD value between disjoint halves of the *Exp-Lab* pairwise preferences across 1000 permutations. This analysis indicates that the similarity between *Exp-Lab* and *Exp-Prolific* results are as high as two different laboratory experiments. Finally, we further analyze the agreement of the statistical significance of the pairs' differences between the experiments. This further confirmed our previous observations. We can conclude that the *Prolific* [88] can be used as a platform for subjective quality evaluation of tone mapped images. Further research might be required to generalize our findings to other QoE tasks.

3.3.2 Inter-observer agreement

Another essential factor in understanding the reliability of the collected data is inter-observer agreement. Due to the higher subjectivity of the aesthetic quality evaluation task, disagreement among observers does not necessarily correlate with these observers' reliability. Nevertheless, comparing the three experiments with the same QoE task and stimuli provides valuable insight regarding the effect of experimental conditions and recruitment procedures. To analyze the inter-observer agreements, we rely on two different measurements: Rogers-Tanimoto dissimilarity [91] and Krippendorff's Alpha [61].

Rogers-Tanimoto Dissimilarity: As previously discussed in the previous section, pairwise preferences are represented in a binary format, *i.e., image A (I_A) is better (1) or worse (0) than image B (I_B).* Traditional correlation analyses fail to capture the agreement among observers in pairwise comparisons. Alternatively, metrics that measure distances between binary vectors can be used [91, 53]. Among the existing binary distances, Rogers-Tanimoto (RT) dissimilarity provides desirable features for observer agreements. RT dissimilarity not only measures the distance between two binary vectors but also allows to weight each to prioritize each observation. It is robust to sample size differences but cannot handle missing entries. RT dissimilarity can be defined as follows:

$$RT_{AB} = \frac{2 \times (v_o)}{v_k + 2 \times (v_o)} \tag{3.1}$$

where v_o is the number of stimuli for which two participants disagree on their pairwise preference, *i.e., one select A over B while other selects B over A.* Conversely, v_k is the number of stimuli where both participants agree on their preference. Additionally, the weight of each pair can be calculated with the following equation to emphasize the effect of pairs with higher agreement on the RT dissimilarity calculation:

$$w_{AB} = \frac{|r_{AB} - r_{BA}|}{N} \tag{3.2}$$

where N is the number of observers ranked the pair P_{AB}. r_{AB} is the number of ob-

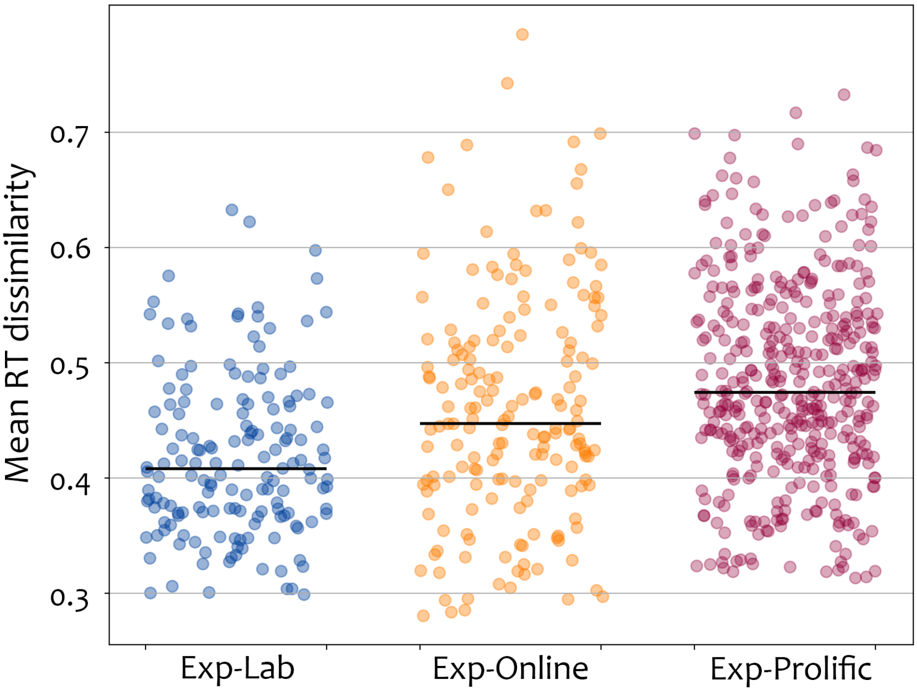


Figure 3.5 – Mean RT dissimilarity distributions of observers for each experiment. Each sample represents the mean RT dissimilarity between an observer and the rest of the observers in corresponding experiment.

servers who prefer image A over image B in pair comparison. Similarly, r_{BA} is the number of observers who prefer image B over image A in pair comparison. This weight calculation allows us to generate weights that are closer to 1 as more observers agree on their preferences for a given pair P_{AB}. Conversely, it generates weights closer to 0 as the ambiguity of the pair increases.

RT dissimilarities are calculated between each observer. In other words, for a given observer, we calculate RT dissimilarity with every other observer who evaluates the same stimuli. RT values range between 0 and 1, and lower values indicate a higher agreement for the corresponding observers.

Figure 3.5 shows the mean RT dissimilarity distribution of observers for each experiment. Each point represents an observer from the corresponding experiment. Note that the number of observers is different in each experiment. Black horizontal lines indicate the median observer dissimilarity for the corresponding experiments. As expected, we observe that the *Exp-Lab* experiment has the highest inter-observer agreement. This can be explained by the controlled experimental conditions and more strict recruitment proce-

Table 3.4 – Inter-observer agreements based on Krippendorff's Alpha coefficient.

	All Pairs				Signf. Diff. Pairs			
	Plist-1	Plist-2	Plist-3	Plist-4	Plist-1	Plist-2	Plist-3	Plist-4
Exp-Lab	0.2244	0.2512	0.2020	0.3229	0.2856	0.3274	0.3214	0.3579
Exp-Online	0.1653	0.2420	0.1571	0.2496	0.2328	0.3157	0.2187	0.4420
Exp-Prolific	0.1576	0.1904	0.1424	0.2224	0.1871	0.2602	0.1958	0.3048

dure in the *Exp-Lab* experiment. On another front, although the distributions are similar, *Exp-Online* has a higher inter-observer agreement than *Exp-Prolific*.

Krippendorff's alpha:

Krippendorff's alpha[61] is a generalized reliability measure that can be used in various scenarios. It works for any number of observers, any scale values (not just pairwise comparisons), and can handle incomplete or missing data. It provides a single reliability measure for a given population over a set of observations. Krippendorff's alpha values range between -1 and 1, where higher values indicate a higher inter-observer agreement.

Since Exp-Online and Exp-Prolific experiments were conducted with four smaller playlists, we divided the Exp-Lab data into the same portions. Although Krippendorff's alpha can work with incomplete data, splitting the dataset into four playlists provides a more fair judgment. Table 3.4 presents the Krippendorf's alpha coefficients of each experiment. The First four columns use all the pairs for calculation, whereas the last four only rely on pairs with a statistically significant difference. For both sets of pairs, we observe a similar outcome to RT dissimilarity. *Exp-Lab* has the highest agreement among observers, *Exp-Online* comes second and *Exp-Prolific* follows with third highest agreement.

In conclusion, both measures indicate that the controlled experimental conditions increase the inter-observer agreement. Furthermore, we observe that the recruitment methodology may impact the inter-observer agreements. Although it provides insight into the effect of isolated factors, inter-observer agreement and observer reliability can not be used interchangeably. Higher variance in observer preferences (the main indication of a low inter-observer agreement) can also occur with reliable observers.

3.3.3 Effect of number of observers

The lower cost of recruitment and wider participant pool makes crowdsourcing platforms attractive for subjective experiments. Although one can recruit infinitely many

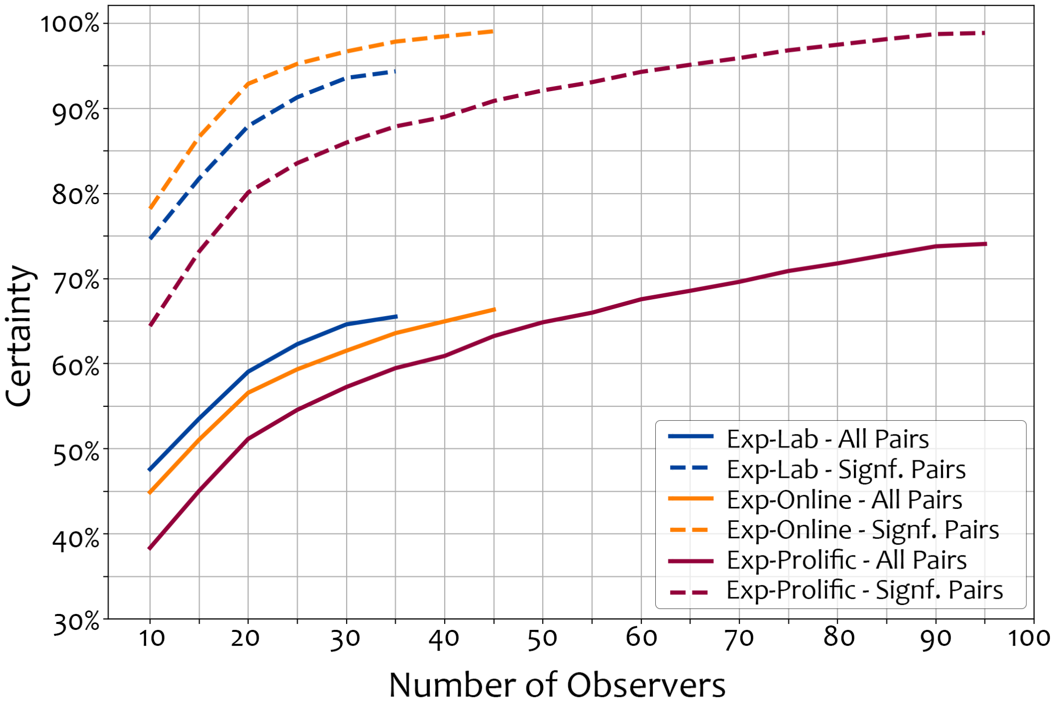


Figure 3.6 – Effect of number of observers on the certainty of the acquired pairwise preferences over 1000 permutations. Horizontal axis is the percentage of pairs which reach to the final conclusion with corresponding number of observers at the vertical axis.

observers with unlimited resources, it is not feasible in the real world. Therefore, it is crucial to determine the required number of observers prior to the subjective experiment. For a given stimulus, there is a threshold that after reaching a certain number of observers, further observations do not affect the evaluation outcome. A common way to estimate such a threshold is by bootstrapping. It allows understanding the effect of the number of observers on the collected subjective preferences.

To do so, we create subsets of observers with incremental sizes from a shuffled list of all observers and evaluate the results of the experiments at each incremental. The evaluation criterion is based on the certainty of the pairwise comparisons. Certainty for a subset of observers was defined as the percentage of iterations which reach the same conclusion with the maximum number of observers about the statistical significance of the difference of image pairs. At each iteration of the bootstrapping, we start by picking five random observers and compare the acquired result with the maximum number of observers. Later, another five random observers are selected and added to the selection until the maximum number of observers is reached. At each incremental, the certainty is calculated. This process is repeated for 1000 iterations. As a result, we acquire 1000 certainty values for every $5n$ observers for a given experiment, where $n = N/5$ and N is the maximum number of observers in the corresponding experiment. This allows us to evaluate the effect of the number of observers robustly on the certainty of the pairwise preferences.

Figure 3.6 illustrates the result of this evaluation. Each line represents the mean certainty value across 1000 permutations. Bootstrapping is done separately with all 120 pairs in the dataset and only with pairs showing statistically significant differences commonly in all three experiments. 100% certainty indicates that, for all the pairs, across all iterations, the number of observers indicated on the horizontal axis is enough to reach to the same conclusion acquired with the maximum number of observers.

When all pairs are considered (solid lines in Figure 3.6), we observe that the *Exp-Lab* requires the least number of observers to reach the same level of certainty ($\sim$ 65%). Considering the higher inter-observer agreement in *Exp-Lab* experiment, this result is not surprising. On another front, when pairs commonly show a statistically significant difference in all three experiments are considered, *Exp-Online* reaches a higher level of certainty with less number of observers than the other two experiments.

To sum up, for all 120 pairs in the datasets, to reach the same level of certainty of the *Exp-Lab* experiment (with 35 observers), *Exp-Online* requires 40 observers, and *Exp-Prolific* requires 50 observers. Similarly, for the statistically significant pairs, in order to reach the same level of certainty of the *Exp-Lab* with 35 observers, *Exp-Online* requires 25, and *Exp-Prolific* requires 60 observers.

3.4 Discussion

In this chapter, we conducted three different experiments with controlled differences to determine the reliability of crowdsourcing platforms for aesthetic quality evaluation of tone mapped images. First, we collected subjective annotations in a controlled laboratory environment. The second experiment was conducted online with the same recruitment channel to isolate the effect of uncontrolled experiment conditions. Finally, we conducted the same experiment on Prolific with the participants pool available on the website to fully investigate the effect of crowdsourcing platforms on the reliability of collected subjective annotations.

Comparing the three experiments revealed that the online experiments have desirable similarity with the laboratory experiment in terms of subjective preferences. Furthermore, effect of Prolific participants pool on the cumulative pairwise preferences is favorable and brings a desirable degree of certainty with enough number of observations per stimuli. We observe a higher variation among observers' subjective preferences in *Exp-Online* and *Prolific*. This is not a surprising outcome considering the uncontrolled environmental conditions of the experiments. Finally, we compared the certainty of the collected subjective preferences with varying number of observers. To reach the desired level of certainty, *Prolific* requires higher number of observers overall when compared to other experiments. Considering the lower cost of recruitment through Prolific and the availability of a wider audience, we find Prolific advantageous in terms of certainty acquired per resource spent.

Hence, through extensive analysis we confirm that Prolific can be safely used to collect subjective preferences on aesthetic evaluation of TMOs. We believe that this conclusion can be generalized to other aesthetic image quality evaluation tasks which do not depend highly on viewing conditions. Finally, we also observe that, depending on the expected certainty compared to the in-lab experiment, the required number of observers to evaluate each pair of stimuli lies between 50 to 60 for a full pair comparison design. We utilize our findings in the large-scale dataset collections which is introduced in Chapter 4.

Chapter 4

Large Scale Aesthetic TMO Quality Evaluation Dataset

Based on the findings of the work presented in the preceding chapters, we collect a large-scale dataset for tone mapped image quality evaluation. To the best of our knowledge, this is the largest publicly available dataset for tone mapped image quality evaluation. An article including the content selection strategy (see Chapter 2), observer screening methodologies (see Chapter 5) with the collected dataset is submitted to IEEE Transaction on Multimedia (TMM) journal and currently under review as part of our collaboration with Abhishek Goswami, Wolf Hauser from DxO and Frédéric Dufaux from CentraleSupélec.

After discussing the challenges of the collecting a large-scale dataset and the motivation behind the work in Section **??**, we present the stimuli collection stage in Section 4.1. Details regarding the subjective experiment design is introduced in Section 4.2. Performance of selected TMOs are analyzed in Section 4.3. Section 4.4 analyzes the performance of tone mapped image quality metrics, and finally, the chapter is concluded with a discussion in Section 4.5.

4.1 Stimuli Generation

Content selection is an essential part of every dataset. The aim of the content selection is to ensure the collection of a representative set of stimuli for the evaluated task, *e.g., tone mapped image quality assessment*. In order to do so, we relied on our previously introduced content selection strategy in Chapter 2.

4.1.1 Source content collection

HDR photography requires time-consuming processing steps and dedicated hardware. It is especially challenging to achieve with moving targets in the frame. On another front, publicly available HDR image datasets can be found in the literature. Therefore, we relied on publicly available HDR images from Fairchild [24] and Artusi [11] HDR collections to generate our stimuli. Both datasets provide HDR images (105 from Fairchild, 124 from Artusi) with nearly 4k resolution by bracketing 5 to 7 exposure levels. However, our subjective experiment design requires a side-by-side display of the image pair on a 1080p display as explained in detail in Section 4.2. Consequently, by following our previously proposed strategy in Section 2.3.1, we reduced the image resolution to the desired size. It crops the original HDR images at various scales to a target resolution of 480p. In the end, we collected 250 HDR images to be used in the final dataset as source images(SRC).

4.1.2 Tone mapping operators

There are many TMOs in literature for HDR images[17] and videos [23]. A brief introduction of various TMOs is also given in Section 1.6.2. Although there is not a single TMO that outperforms the rest in terms of tone mapped image quality for all possible types of content, previous works indicate KimKautzTMO [56] and KrawczykTMO [60] perform slightly better in general[17]. ReinhardTMO [90] is another popular option that is widely adopted in the computer graphics domain and provides tone mapped images with consistent quality over various scenarios. SemTMO [35] is one of the recent TMOs which utilizes semantic information in the scene for tone mapping. Based on the performance and the type of TMOs in the literature we chose 4 of them, namely: SemTMO [35], KimKautzTMO [56], KrawczykTMO [60] and ReinhardTMO[90]. We briefly introduce each TMO below:

SemTMO is a local TMO where segments the scene into various semantic categories and tone maps each segment individually.

KimKautzTMO is a global TMO that follows a gaussian distribution around the average log luminance of the scene for tone mapping.

KrawczykTMO is a local TMO that processes tine image by patches of consistent luminance and calculates the lightness of each patch locally.

ReinhardTMO is a hybrid TMO that provides global and local options for tone mapping. Global scaling of the dynamic range is followed by local operations of dodging and burning.

TMOs with adjustable parameters were optimized based on TMQI [122] scores by a grid search. Each SRC (among 250 SRCs) is tone mapped with the 4 TMOs listed above. We ended up with 1000 tone mapped images.

4.1.3 Final stimuli

The experiment is a pairwise comparison experiment that aims to assess the tone-mapped images' aesthetic quality. Motivation and justification of the design choices are discussed in detail in Section 4.2. Since it is necessary to point out some details to introduce the final stimuli, we will briefly discuss these details below.

The experiment is limited to participants with display devices of 1080p resolution so that the two tone mapped images with 480p resolution can be displayed side by side without down or upscaling. Each of the four tone mapped images for a given SRC is compared to each other this way. We end up with six pair comparisons per SRC. There are no cross-content comparisons, *i.e., no comparison between tone mapped images of different SRCs.*

Conducting subjective experiments on crowdsourcing platforms requires additional care and experimental design choices[31, 44]. One of the most straightforward implications is the lower attention span of the participants in crowdsourced experiments. We can overcome the issue by reducing the length of the experiment sessions. By providing a smaller-sized playlist with approximately 5 minutes length, we can increase the attention paid by participants on each stimulus which provides a more reliable data collection. In our experiment, we divide the collected dataset into 50 smaller playlists with 5 SRCs, 30 pairs to compare in each. Figure 4.1 presents a sample playlist with 5 SRC in each row and four tone mapped versions in each column. 4 tone mapped images at each column are compared to each other in a side by side fashion.

Additionally, to identify participants who pay low attention to the stimuli, we included golden units into each playlist. Golden units are explained in details in Section 5.2.3. Figure 4.2 presents the images which are used as golden units in each playlist. Each column in the figure represents an overexposed image in the first row and a preferable version in the second row. We expected each participant to respond to these pairwise comparisons with the choices in the second row. In the end, with the addition of 3 golden units, we end up with 33 pairs to compare in each playlist.

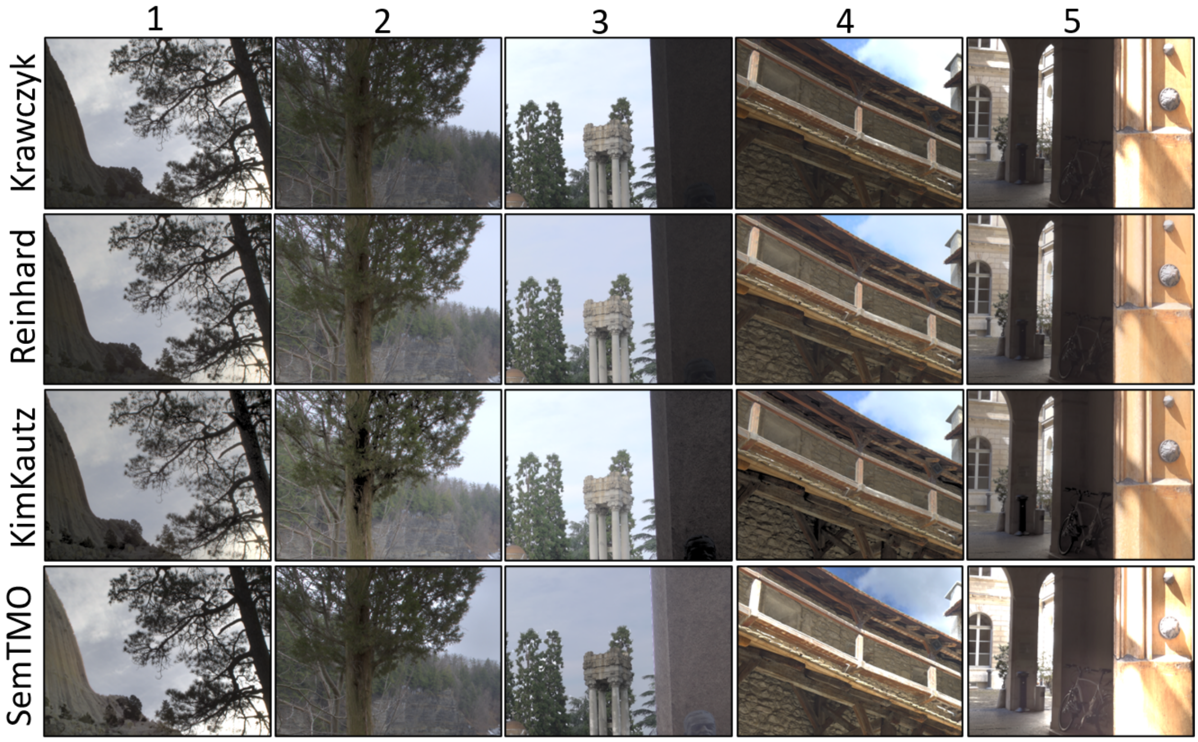


Figure 4.1 – Sample playlist from the experiment.

4.2 Subjective Experiment Design

This section will introduce the experimental design, information regarding participants and crowdsourcing platforms, and strategies adopted to reject unreliable observers.

4.2.1 Experiment setup & Procedure

Subjective quality evaluation of tone mapped images can be categorized as full-reference(FR) and no-reference(NR) based on the presence of HDR reference during the subjective experiment. While FR comparison reveals information regarding the test image's fidelity to the HDR image, NR methodology reveals the overall aesthetic quality by the observer[58]. In this experiment, we focus on aesthetic preferences among the tone mapped images, and consequently, we followed an NR scenario. It is also practically impossible to conduct crowdsourcing experiments for FR quality evaluation of tone mapped images due to the lack of participants with HDR screens available.

Another way to categorize the subjective quality assessment experiments is based on

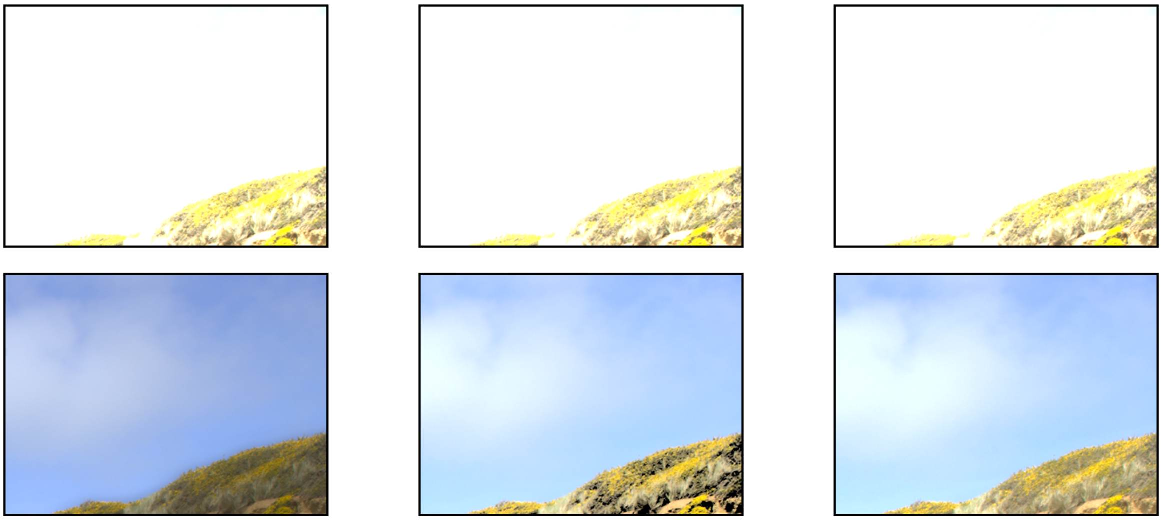
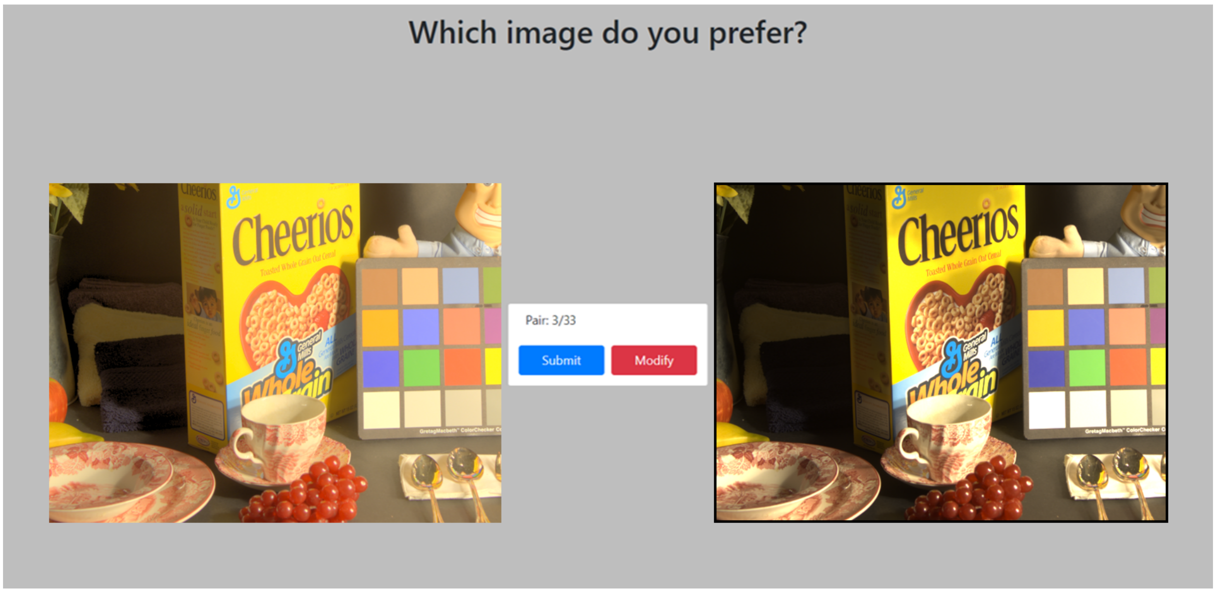


Figure 4.2 – Golden units included in each playlist.

the task presented to participants as rating and ranking tasks. Rating tasks ask the participants to assign quality scores to displayed stimuli based on a predefined scale. This can be achieved by displaying a single stimulus(absolute category rating(ACR)) or two stimuli(double stimulus impairment scale(DSIS)). On the other hand, ranking methods ask the observers to compare two or more stimuli and rank them based on their quality. Pair comparison(PC) is the most commonly used ranking methodology in subjective image quality assessment experiments. Two images are shown to participants each time, and their preference among the two images is requested. PC methodology has the advantage of simplifying the task for participants by eliminating the need to understand the quality scale. Therefore it allows a more reliable quality evaluation compared to rating tasks. On the other hand, the number of pair comparisons required for the experiment increases exponentially with every additional SRC and HRC. The number of required comparisons can be reduced by not comparing all possible pairs. Simple methodologies such as square design(SD) or adaptive square design(ASD) may be used in this regard with minimal loss of accuracy on the collected subjective preferences[66]. Additionally, the number of required comparisons may be reduced if cross-content comparisons are omitted, *i.e., only comparing PVSs from the same SRC.* Although this prevents mapping the ranked stimuli into a global quality scale, it may not be necessary for many tasks such as quality evaluation of tone mapping operators. Novel objective quality metrics proposed in the literature in recent years can utilize pairwise comparison data directly without a need to map onto


Which image do you prefer?

Pair: 3/33

Figure 4.3 – Sample screenshot from the experiment.

a quality scale [86]. Consequently, considering that the aesthetic quality evaluation of the tone mapped images can be significantly affected by the aesthetic quality of the SRCs, we did not include the cross-content comparison. It does not provide a significant benefit for our task. On another front, we followed a full PC design where each possible pair for a given SRC is evaluated. Since there are only 4 HRC in the experiment and the task is highly subjective, the gain from adaptive methodologies such as ASD is lower.

A sample screenshot from the experiment is presented in Figure 4.3. Side by side display of two tone mapped stimuli is shown. After clicking on the preferred image, a confirmation window appears to prevent accidental submissions. Participants are informed about the current pair number and the total number of pairs at each confirmation phase. Additionally, a preferred image is indicated with a black border to decrease any confusion regarding participants' selection. A neutral gray background is used throughout the experiment.

4.2.2 Experiment platform & Participants

We conducted the subjective experiment on Prolific [88] crowdsourcing platform. Participants were recruited from the participant pool provided by Prolific. Prolific provides a reliable participant pool that is governed by ethical concerns. Compared to alternative crowdsourcing platforms, the overall reliability of participants in the Prolific platform is higher[82]. On the contrary, Prolific does not provide an Application Programming Interface(API) which makes certain design implementations,*e.g., SD, ASD, etc.*, practically challenging. Considering the experiment design introduced in Section 4.2.1, this is not a concern for our study.

70 unique observers evaluated each stimulus. The number of observers per stimuli is decided based on the recommendations on the pilot study introduced in Chapter 3. Specifically, in Section 3.3.3, we observe that the certainty achieved with 35 observers in a controlled laboratory experiment for statistically significant pairs can be achieved with 70 observers in Prolific. For 50 playlists in the experiment, 3500 participants were recruited where each participant evaluated only 33 stimuli. The average time spent per stimuli was 4.08 seconds in the experiment, indicating that the average time an observer spent for the experiment was 2 minutes 15 seconds. Each participant (including the rejected participants) was compensated for their time based on Prolific standards. At the beginning of their session, participants were informed that their data would be used in research and signed a consent form which the local ethical committee approved. Participants were also informed that they could stop the experiment at any given point without any consequences.

Participants were not limited based on demographics. The only limitations applied in the recruitment process were the 95% acceptance rate in previous studies to increase the reliability of the collected data and display device. We limited the experiment to participants with display devices of 1080p resolution and windows operating system in order to control the variety of the experimental conditions. Recruited participants were from over 20 different countries. This was ensured by publishing playlists with 6 hours intervals during the day. 2311 of the participants were male, with a mean age of 28.75 and a standard deviation of 9.47. 1154 participants were female, with a mean age of 31.54 and a standard deviation of 10.83. The remaining 35 observers preferred not to share their demographics with us.

4.2.3 Rejecting unreliable observers

In this section, we will only share the number of people rejected based on each screening methodology. Complete procedure of observer screening is explained in details in 5.3.3 and more details regarding the methodologies and their implementation can be found on Chapter 5. We rejected 49 observers based on golden units, 13 on the voting pattern, 56 on the voting speed, and 96 on the RT dissimilarity approach. Participants rejected

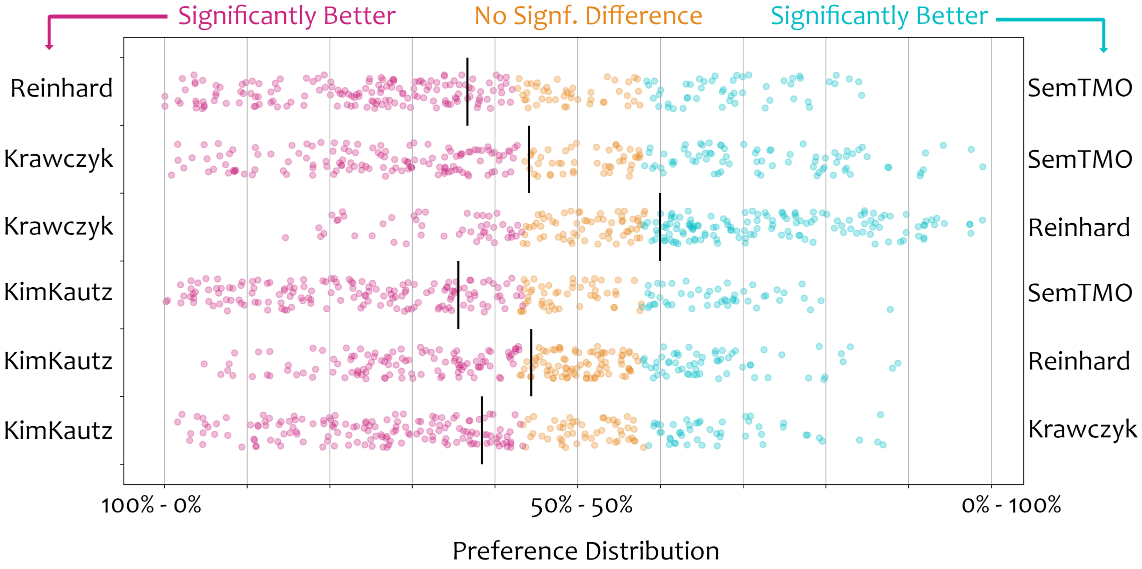


Figure 4.4 – Distribution of percentage of preferences for each pairwise comparison. Each data point represents a unique image pair in the dataset. Black lines indicates the mean values of the preference percentages represented on the horizontal axis.

based on golden units were identified during the experiment, and new participants were recruited as a replacement. However, the rest of the rejection methodologies were applied after conducting the experiment. Therefore the number of valid unique observers might be less than 70 in some playlists, 67 unique obs per playlist on average.

4.3 TMO Performance Evaluation

Although the primary goal of the dataset is to provide a challenging dataset for the benchmark of tone mapped IQA metrics and a representative dataset to develop new metrics, it is also an invaluable opportunity to evaluate TMO performances.

As previously described in Section 4.1, 250 SRC were tone mapped with four different TMOs, and each tone mapped image for a given SRC was compared in a pairwise fashion. This results in 6 pairs of TMO comparisons for each 250 SRCs. Percentage of preferences towards one of the tone mapped images in each pair can be used to assess the performance of TMOs. Moreover, we can determine the statistical significance of subjective preferences for each pair.

There are several ways to determine the statistical significance of the differences be-

Table 4.1 – The table reports comparative results of TMOs in terms of percentages of pairs where each TMO on the row is significantly better than the TMO on the column.

	KimKautz	Krawczyk	Reinhard	SemTMO
KimKautz	-	**60%**	**42%**	**62%**
Krawczyk	19%	-	19%	**52%**
Reinhard	24%	**56%**	-	**62%**
SemTMO	19%	30%	20%	-

tween different distributions[12, 27]. It has been shown that Barnard's exact test is more powerful than alternative statistical tests such as Fisher's exact test [27] on 2 × 2 contingency tables [75]. We use Barnard's exact test since pair comparison results can be represented by 2 × 2 matrices. For a given pair P_{AB}, preference of I_A and I_B can be arranged diagonally as $\left(\begin{smallmatrix} I_A & I_B \\ I_B & I_A \end{smallmatrix}\right)$ and Barnard's test can be used to determine the statistical significance of subjective preference differences between I_A and I_B.

Figure 4.4 presents the TMO performances in terms of distribution of pairwise preference percentages. Unique pairs of TMO (6 in total) are divided into individual rows for ease of reading. Each point in the plot represents a unique image pair from the dataset, which is tone mapped with TMOs indicated on the left and right sides of the plot. The figure displays the preference in terms of the percentage of observers on the horizontal axis. Points close to one side of the horizontal axis indicate a higher preference towards the corresponding TMO on that side. Additionally, the statistical significance of the pairwise preferences is color-coded as labeled in the figure.

As Fig. 4.4 indicates, KimKautzTMO has a superior performance compared to the rest of the TMOs evaluated in the experiment. Reinhard performs the second best, while Krawczyk is slightly better than SemTMO as the third-best TMO. Additionally, we can quantify the results based on the number of pairs where one TMO is better/worse than another as summarized in Table 4.1. Each cell on the table indicates the percentage of pairs that have a statistically significant preference towards the TMO on the row compared to the TMO on the corresponding column. Note that, the sum of percentages between the two TMO is not equal to 100% due to pairs with a statistically non-significant difference (points with yellow color in Fig. 4.4).

4.4 Objective Quality Metric Performance Evaluation

We start the section by introducing the evaluated IQA metrics. Later we introduce the evaluation scenario[59] followed by the pre-processing of the subjective preferences. Finally, we present the result of our evaluation and discuss IQA metric performances.

4.4.1 Selected IQA metrics

We selected two full-reference and two no-reference metrics from the literature dedicated to tone-mapped image quality assessment.

TMQI: is a full-reference image quality metric to assess the quality of tone mapped images [122]. Structural and naturalness measures are combined to evaluate the tone-mapped image's quality with respect to the HDR image. It is the state-of-the-art quality metric for tone mapped image quality assessment.

NIQMC: is a no-reference image quality metric that is developed to assess the quality of contrast distorted images [38]. It combines the local and global features to generate a quality score. Although it is not specifically developed for tone mapped image quality assessment, it shows a relatively high correlation with subjective opinions in aesthetic evaluation tasks.

BTMQI: is a no-reference image quality metric to assess the quality of tone mapped image by combining 11 features related to information entropy, statistical naturalness, and structural preservation[36].

FFTMI: is a full-reference tone mapped image quality metric[57]. It relies on structural similarity, feature naturalness, and feature similarity between the HDR and tone mapped images.

4.4.2 Evaluation criteria

Traditional correlation measurements rely on ground truth Mean Opinion Scores (MOS) obtained through rating experiments. Correlation between the MOS and predicted quality scores are computed to evaluate the objective IQA metrics. Although there is a strong linear correlation between pairwise preferences and MOS, it is not straightforward to map pairwise preferences into a global quality scale [126]. Additional precautions in experiment design are often required, such as cross-content comparisons, which may drastically

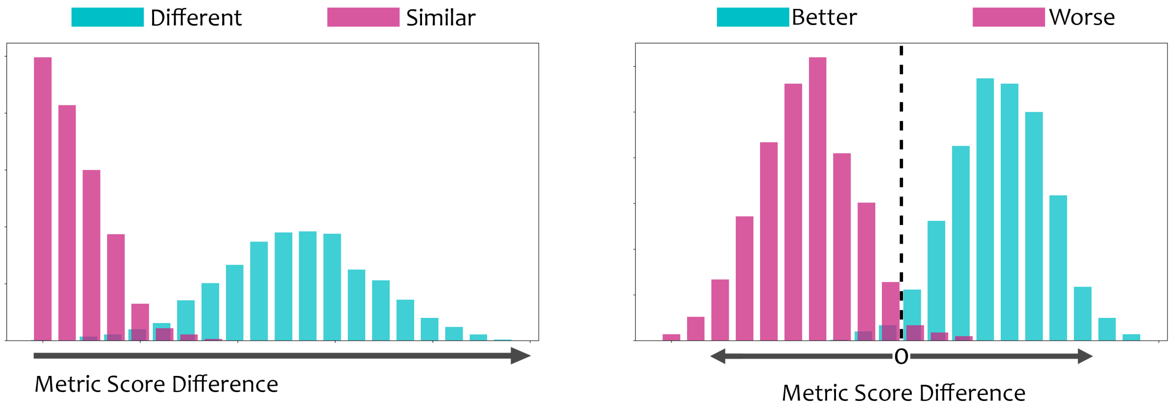


Figure 4.5 – Ideal distributions for different vs similar and better vs worse analyses on the left and right respectively.

increase the cost of the experiment and are often not beneficial for the purpose. Alternatively, Krasula et al. [59] propose an evaluation model which does not rely on mapping the collected preferences into a standard scale. It also enables the merging of multiple datasets while determining the statistical significance of the performance differences.

In the Krasula model, performance evaluation of the objective quality metrics is conducted in two different stages. The first stage focuses on how good the IQA metrics are at distinguishing between pairs with and without statistically significant differences. The second stage determines whether the metrics can recognize the image with higher preference in pairs with a statistically significant difference.

Numerical analysis of the metric performances for both stages is done with AUC values, while the distribution of metric score differences can be visualized in a histogram. Ideal distributions of the metric score differences are shown in Figure 4.5. A more detailed introduction of the Krasula model is given 6.2.2.

4.4.3 Pre-processing subjective preferences

As described in Section 4.4.2, the Krasula method requires the statistical significance of the differences for each pair of images. In order to do so, we use Barnard's exact test[12] to determine whether a pair contains a statistically significant difference between the two tone mapped images. Each pair of images (P_{AB}) are arranged in a 2×2 matrix as $\begin{pmatrix} I_A & I_B \\ I_B & I_A \end{pmatrix}$ where I_A and I_B are the number of observers who prefer images A and B, respectively. Among 1500 pairs in total, we determine that the 1154 pairs contain a statistically significant difference with 95% confidence.

Furthermore, for better and worse analysis, pairs with a statistically significant difference are divided into two groups as better and worse. We split the different pairs into two groups as better (736) and worse (418). Better pairs indicate the pairs where the image on the left is better than the image on the right, and conversely, worse pairs indicate the pairs where the image on the right is better than the image on the left. Although any pair can easily be categorized as better or worse by swapping the image positions, we used the initial positioning of the dataset as the random seed since the number of better and worse pairs are similar.

4.4.4 Pre-processing objective quality metric predictions

Objective quality metrics predict a quality score for each tone mapped image. For a given pair P_{AB}, we calculate the predicted quality score difference as $m_A - m_B$ where m is the objective quality metric. Once we gather all predicted score differences for the evaluated metrics, we move on to the evaluation scenario as described in Section 4.4.2.

4.4.5 Evaluation results

In this section, we present our analysis of the performances of selected objective quality metrics. As previously introduced in Section 4.4.2, evaluation is done in two steps.

Different vs similar analysis:

Firstly, we analyze the metrics in their ability to distinguish pairs with and without statistically significant differences. As discussed earlier, the ideal distribution of predicted quality score differences should be similar to the one depicted in Figure 4.5.

Figure 4.6 presents the results of the histogram of metric score differences for different and similar pairs. Blue represents the pairs with a statistically significant difference for each plot, whereas pink represents similar pairs. Visual analysis of the histograms reveals that none of the metrics provides a similar distribution to the ideal scenario. Although the difference between each metrics' distribution is subtle, we can see that the FFTMI metric score differences for similar pairs have higher occurrences for values closer to zero.

AUC values of each metric are provided at the corner of the corresponding plots in Figure 4.6. By comparing the AUC values, we can observe that the performances of TMQI, NIQMC, and BTMQI are close to each other. Statistical test results also suggest

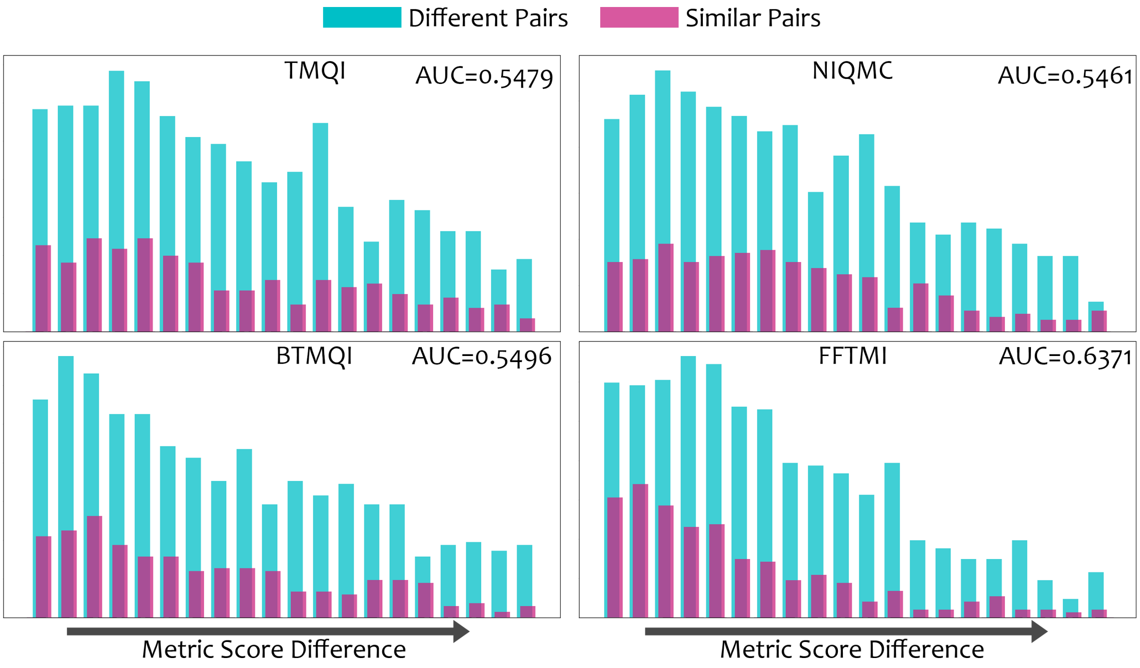


Figure 4.6 –

no significant difference between the performances of TMQI, NIQMC, and BTMQI. On another front, FFTMI outperforms the rest of the metrics, and the difference in performance with other tested metrics is statistically significant. Although FFTMI is the best performing among the selected metrics, its performance is far from ideal, indicating room for improvement in tone mapped image quality assessment.

Better vs worse analysis:

As explained in Section 4.4.2, better vs worse analysis aims to determine the accuracy of objective metrics at identifying the image with higher quality in a pair. The result of the analysis is presented as a histogram of metric score differences for better and worse pair categories in Figure 4.7. It can be observed that better vs worse analysis draws a similar conclusion different vs similar analysis. An example of an expected distribution for metric score differences was depicted on the right plot in Figure 4.5. As it can be observed, none of the metrics provides a similar distribution. Moreover, AUC values for each metric are reported on the corner of each corresponding plot. While it is far from the ideal distribution, FFTMI provides the highest AUC value when compared to the rest of the metrics.

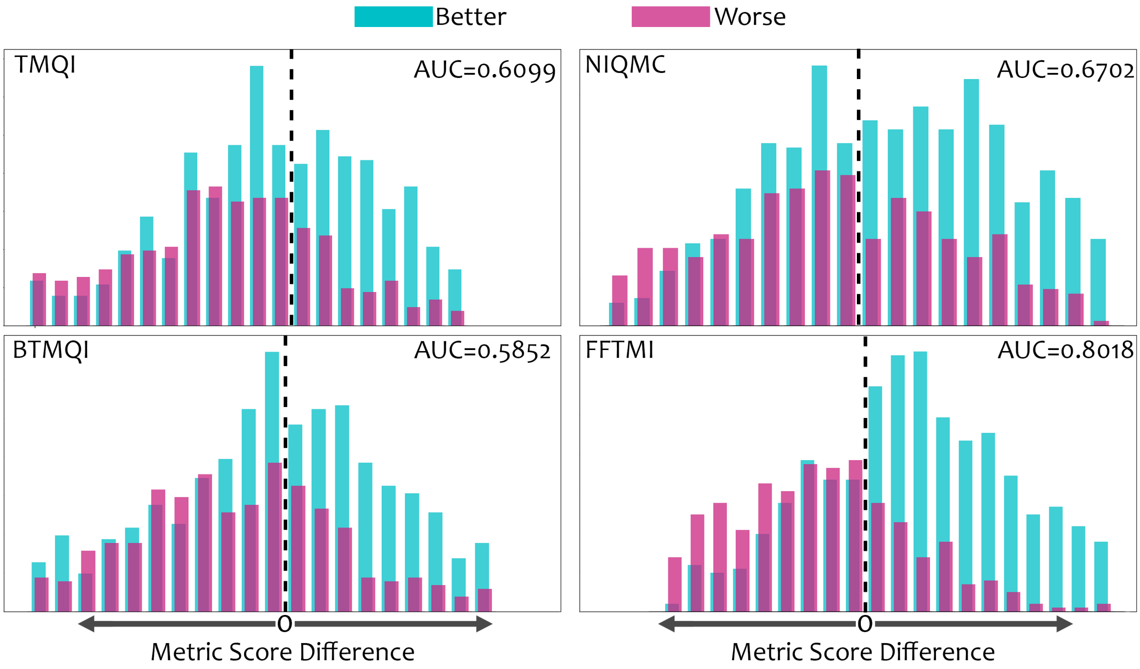


Figure 4.7 –

In addition to the AUC analysis, we evaluate the metric performances in the Better/Worse classification task with their percentage of correct classifications. Acquired correct classification percentages were 58%, 61%, 56%, 72% for TMQI, NIQMC, BTMQI and FFTMI respectively. Statistical significance results acquired by Fisher's exact test on correct classification rates indicate significantly better performance for FFTMI when compared to others. The performance of NIQMC is also significantly better than TMQI and BTMQI, whereas there is no statistically significant difference between TMQI and BTMQI performances.

4.5 Discussion

As discussed in Chapter 1, it is easier and more natural for participants to compare the quality of two images than to assign a quality score to each image individually. Despite the advantages of pairwise comparison over rating tasks, metric development often relies on MOS scores. A method has been proposed to acquire MOS from pairwise preferences [126]. The authors conduct a series of experiments to acquire MOS scores from pairwise preferences and suggest including cross-content comparisons into the experiment to scale each stimulus into a global quality scale. However, it is not valuable to include cross-content comparisons in many use cases such as ours.

In order to develop objective IQA models directly on pairwise preferences, alternative objective functions might be incorporated into training. Prashnani et al. used a modified Bradley Terry (BT) [14] model as an objective function to train a deep learning model on probabilistic pairwise preference data [86]. The model predicts quality scores for each stimulus during training, and pairwise preference probabilities are calculated from the predicted scores with a modified version of BT. After training, the model is able to predict quality scores for individual stimuli (in comparison to a pristine reference image).

In this chapter, we conducted a large-scale experiment on tone mapped image quality evaluation via crowdsourcing. To the best of our knowledge, this is the largest publicly available TMO evaluation dataset: 250 unique HDR images used to generate 1000 tone mapped images which provides 1500 pair comparisons. 3500 observers participated in the subjective experiment where approximately 70 unique observers evaluated each pair. 4 state-of-the-art TMO performances were evaluated, where KimKautzTMO [56] was most often preferred. ReinhardTMO [90] performed the second best while KrawczykTMO [60] came in third place, performing slightly better than the SemTMO [35] in fourth.

We utilized our content selection strategy proposed in Chapter 2 to select representative and challenging HDR crops from high-resolution HDR images. We further developed an objective quality metric based clustering method to balance the ambiguity of the pairs in the experiment. It is crucial to have such balance for developing metrics, specifically for learning-based models.

Finally, we provide a benchmark for well-known tone mapped image quality metrics based on the Krasula method [59]. We discussed how to utilize collected data to develop novel objective quality metrics and benchmark existing metrics. Collected pairwise preferences, stimuli used in the experiment, and scripts are publicly available to further research.

To the best of our knowledge, there is a lack of a well-established methodology for observer reliability in pairwise comparison experiments. In addition to behavioral tools, we used a novel approach to statistically evaluate the observer reliability and remove the outliers in our pairwise comparison experiment. Proposed novel methodology, as well as the behavioral tools for observer screening, are discussed in detail in Chapter 5.

Chapter 5

Screening Observers in Crowdsourced Pair Comparison Experiments

This chapter introduces a set of observer screening methodologies that aim to identify unreliable observers in pairwise comparison experiments. We can split the suggested methodologies into two categories as *behavioral* and *statistical* tools. Although behavioral tools introduced in this chapter are well known in the literature, we propose a novel statistical methodology for observer screening in pairwise comparison experiments. The proposed method was submitted to IEEE Transaction on Multimedia (TMM) journal, and it is currently under review. Part of the work is also published in 2021 IEEE International Conference on Multimedia Expo Workshops (ICME 2021) as a result of our collaborations with Mona Abid and Matthieu Perreira Da Silva from University of Nantes.

Analyses on our chapter mostly rely on our large-scale tone mapped image quality evaluation dataset, RV-TMO (see Chapter 4). Behavioral tools are introduced in Section 5.2 and Rogers-Tanimoto dissimilarity is introduced in Section 5.3.2. Complete process of detecting outliers and the number of rejected outliers in RV-TMO dataset are given in Section 5.3.3. Finally, we investigate the effect of QoE task subjectivity on RT dissimilarity measure in Section 5.4.

5.1 Challenges & Motivation

In the previous chapter, we discussed the design choices to increase the reliability of subjective annotations. We concluded that the Prolific [88] could be used for subjective quality evaluation of tone mapped images. On the other hand, this does not imply that there are no outliers or spammers among the crowd.

ITU standards [48] recommends outlier rejection methodologies which targets rating experiments. Simply the procedure relies on calculating correlation coefficients between the global MOS and individual participant opinions. Participants with low correlation are considered outliers. However, there is no well-established methodology to identify outliers in pairwise comparison experiments.

Due to the binary nature of pairwise preferences, it is challenging to detect outliers based solely on statistical measures. In rating experiments, detecting a participant who votes differently (e.g., voting 1 on a 1-5 scale for a stimulus with 4.3 MOS) can be relatively simple. However, when the choice is binary (either voting for image A or image B), the difference between an outlier and a valid opinion is small. Moreover, QoE tasks with high subjectivity further narrow this gap between a valid opinion and an outlier.

This chapter aims to provide a set of tools for identifying unreliable observers in a crowdsourced pairwise comparison experiment. We can categorize the provided tools into two groups as behavioral and statistical. Behavioral tools rely on detecting suspicious behavior with the data collected during the experiment, such as voting speed, voting position patterns, and golden units. Although they are powerful at identifying certain behaviors, not all types of outliers can be detected with just behavioral analysis. Furthermore, we propose a novel outlier detection strategy for pairwise comparison experiments which relies on Rogers-Tanimoto dissimilarity (RT dissimilarity) [91] measure.

5.2 Behavioral Screening Tools

Behavioral methodologies can be defined as tools that aim to identify irregular behaviors by building the expectations of a reliable observer. Prior knowledge regarding the dataset, subjective task and experiment design can be incorporated to develop such tools. Although they are powerful methodologies to detect certain unreliable behaviors, they may not be enough to capture every type of spammer profile. Nevertheless, their value in screening observers is prominent and widely adopted in the literature. This section investigates three screening methodologies that fall into this category: voting speed, voting pattern, and golden units.

5.2.1 Voting speed

Certain spammer profiles on crowdsourcing experiments tend to optimize their efforts by finishing more tasks and minimizing the time spent on each task, resulting in a lack of

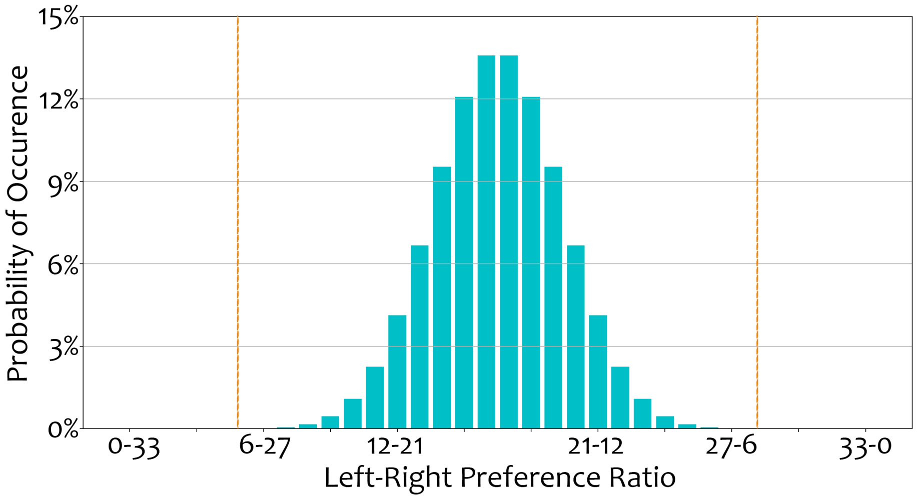


Figure 5.1 – Occurrence probability of each left-right ratio for 33 stimuli. Dashed lines represent the limit for rejection.

attention and consequently falsely submitted answers. Although sophisticated spammers can automate their efforts and avoid speed checks, this method aims to identify unreliable participants who fail to afford such techniques. Considering that the voting speed analysis has no additional cost other than recording time stamps of observer preference submissions, incorporating it in the experiment brings no disadvantages.

5.2.2 Voting pattern

Spammers can minimize their efforts by selecting the stimulus in the same position continuously. In rating experiments, this can occur by providing the same score over and over again without care. Obviously, this type of behavior is not appreciated when collecting subjective preferences. Thankfully, it is not difficult to identify such behaviors if necessary data is collected during the experiment.

Specifically for pairwise experiments, where two stimuli are displayed at a given time, observers might select the stimulus on the same position again and again. Observers were presented with 33 side-by-side stimuli during the experiment. Therefore, we made a probabilistic analysis of the voting patterns for the given experiment. Figure 5.1 depicts the probabilities of constantly voting on one side during the experiment. As indicated by the yellow vertical lines, the threshold is chosen as five or fewer votes on a single position (left/right). Probabilistically, it is unlikely (around once per 10000 observers) to vote on

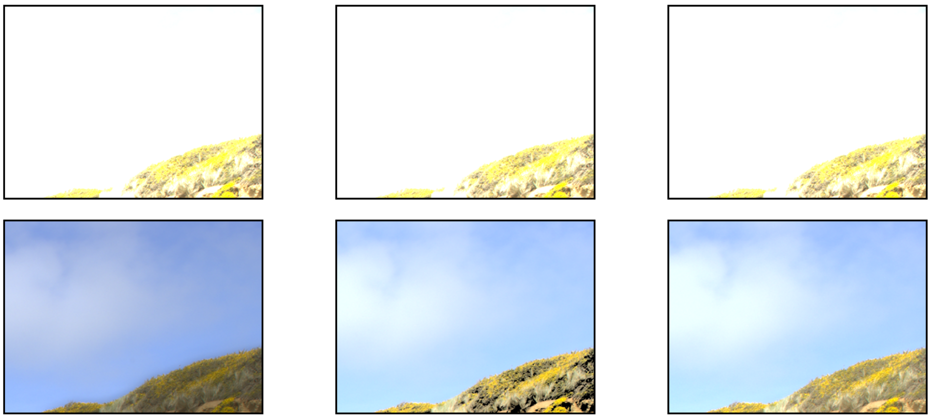


Figure 5.2 – Pair of images in each column were used as golden units.

one position less than 6 times among the presented 33 stimuli.

5.2.3 Golden units

Golden units are quite powerful thanks to the assumptions made prior to the experiments. Golden unit is a stimulus where the correct answer is given and expected from each participant, and those who cannot provide the expected answer can be flagged as unreliable. Prior to the experiment, a set of stimuli is selected as golden units. As expected, selected stimuli play a crucial role by providing a threshold for rejecting participants.

Selected stimuli as golden units for the subjective experiment are displayed in Figure 5.2. The selection of golden units was made by a pilot test, a controlled environment laboratory experiment. 40 participants were recruited for the pilot test, and all participants provided the same answer for the golden units without any specifications about the stimuli. As can be seen in the Figure, each column is presented the participants as golden units. Preference towards strongly over-exposed images, displayed on the top row, is considered an unreliable behavior indicator. Selected golden units were added to each playlist and shuffled to prevent position bias.

5.3 Estimating Observer Reliability from PC data

Certain spammer profiles may be targeted via behavioral tools. However, spammers who do not fit any of the profiles targeted above or spammers with sophisticated strategies cannot be identified with behavioral tools alone. Therefore, ITU standardization efforts [48] suggest statistical tools to reject observers based on estimated reliability. However, suggested tools target only the rating experiments, and there are no well-established methodologies of statistical evaluation of reliability for ranking experiments. In this section, we introduce a novel methodology for estimating observer reliability in pairwise comparison experiments. It relies on Rogers-Tanimoto (RT) dissimilarity [91] to estimate how much the two observers agree with each other. We randomly generate 1000 synthetic spammers to create an expected spammer RT dissimilarity distribution. Then we compare the RT dissimilarities of real observers with the expected distribution to identify unreliable observers.

5.3.1 Synthetic Spammer Profiles

Four different spammer profiles were introduced in this section. Each spammer profile is randomly generated based on the rules which define the given behavior. For each spammer profile, the intensity of its spammer behavior is controlled by a variable.

Random voter: An observer may randomly vote on I_A or on I_B during the experiment due to lack of attention and lack of motivation. We generate this behavior by randomly sampling binary preferences for each stimulus. We control the intensity of this behavior by selecting a real observer and replacing its pairwise preferences with random votes. The amount of votes to be changed is controlled with a variable.

Repeater: An observer might show a position bias, thus providing his/her pairwise preferences based on image position, *i.e., left/right, top/bottom.* We simulate this behavior by repeating a random position. Similarly, we control the intensity of this behavior by selecting a real observer and using a variable to control the number of repeated votes to be replaced by a variable.

Inverted voter: Due to misunderstanding of the task or simply for malicious motives, the observer may submit their preferences on the wrong stimuli, *i.e., left instead of right or right instead of left.* We generate this behavior based on a randomly selected real observer and inverting his/her pairwise preferences. The amount of votes to be inverted is controlled by a variable to adjust the spammer behavior intensity.

Mixed: Finally, we generate a mixed spammer profile based on the combination of behaviors described above. A single variable is used for the intensity of all spammer behavior mixed.

5.3.2 Rogers-Tanimoto dissimilarity

As discussed earlier, pairwise preferences of an observer are represented as a binary vector, *i.e., 0 for selecting stimulus on the left and 1 for selecting stimulus on the right.* Therefore each observer has a binary vector of length N where N is the number of stimuli in the experiment. This allows us to use binary distance metrics in order to measure the similarity between the common pairs of any given two observers. Although there are numerous binary distance metrics [16, 53, 91] available in the literature, RT dissimilarity is particularly attractive for pairwise preferences due to the weight factor, which can be utilized to prioritize certain stimuli over others. This allows us to penalize the observers more when they disagree with the majority on *easy* pairs, whereas penalizing less for pairs with high ambiguity.

RT dissimilarity (RT_{ij}) of two observers obs_i and obs_j is calculated as follows:

$$RT_{ij} = \frac{2 \times (v_d)}{v_a + 2 \times (v_d)} \tag{5.1}$$

$$RT_{ij} = \frac{2 \times (v_{di} + v_{dj})}{v_a + 2 \times (v_{di} + v_{dj})} \tag{5.2}$$

where v_d is the number of stimuli for which obs_i and obs_j disagree, while v_a is the number of stimuli for which obs_i and obs_j agree. Additionally, we incorporate a weight for each stimuli evaluated by obs_i and obs_j. It is calculated separately for each playlist. Weights are calculated with the following function:

$$w_{AB} = \frac{|p_A - p_B|}{N} \tag{5.3}$$

Where N is the number of observers who evaluated the image pair A, B, p_A is the number of an observer who selected image A (I_A) while p_B is the number of an observer who selected image B (I_B), for pairs with high ambiguity *(e.g.,* 50% *prefers I_A and other* 50% *prefers I_B)*, calculated weight is closer to 0 as ambiguity increases, whereas for pairs with high agreement on one image weight is closer to 1. This allows us to penalize the observers who disagree with the majority. On the other hand, high ambiguity pairs do not contribute to unreliability as much.

5.3.3 Rejecting unreliable observers from the experiment

For a given set of stimuli, a playlist of 30 pairwise comparisons in our case, the algorithm 25 summarizes the spammer detection procedure.

Algorithm 2: Detecting unreliable observers based on the expected RT dissimilarity range of randomly generated synthetic spammer profiles.

```
Gather pairwise preferences from N observers for m number of stimuli
Apply behavioral tools and filter out unreliable observers among the initial N.
Initialize a two-dimensional array $RT_S$ with size $\{1000, \hat{N}\}$ to store RT
  dissimilarities of synthetic spammers
while $i<1000$ do
    Select a random observer ($o_r$) among the filtered  observers.
    Generate a synthetic spammer ($s_i$) based on $o_r$ and a randomly selected
      spammer profile.
    while $j<\hat{N}$ do
        Calculate RT dissimilarity between $s_i$ and $o_j$
        Store it on $RT_S[i, j]$
    end
end
Create an expected threshold $RT_{low}$ for RT dissimilarity of spammers based on
  $RT_s$
Initialize a two-dimensional array $RT_O$ with size $\{\hat{N}, \hat{N}-1\}$ to store RT
  dissimilarities of real observers.
Loop over $\hat{N}$ observers.
while $k<\hat{N}$ do
    Calculate RT dissimilarities between $obs_k$ and the rest of the observers.
    Store it on $RT_O[k, :]$
    Calculate the $10_{th}(RT_{k-low})$ and $90_{th}(RT_{k-high})$ percentiles.
    Calculate the overlap between $[RT_{k-low}, RT_{k-high}]$ and the threshold $RT_{low}$
    if Overlap is higher than 80% then
        Mark observer as unreliable
    else
        Mark observer as reliable
    end
end
Result: Return unreliable observers
```

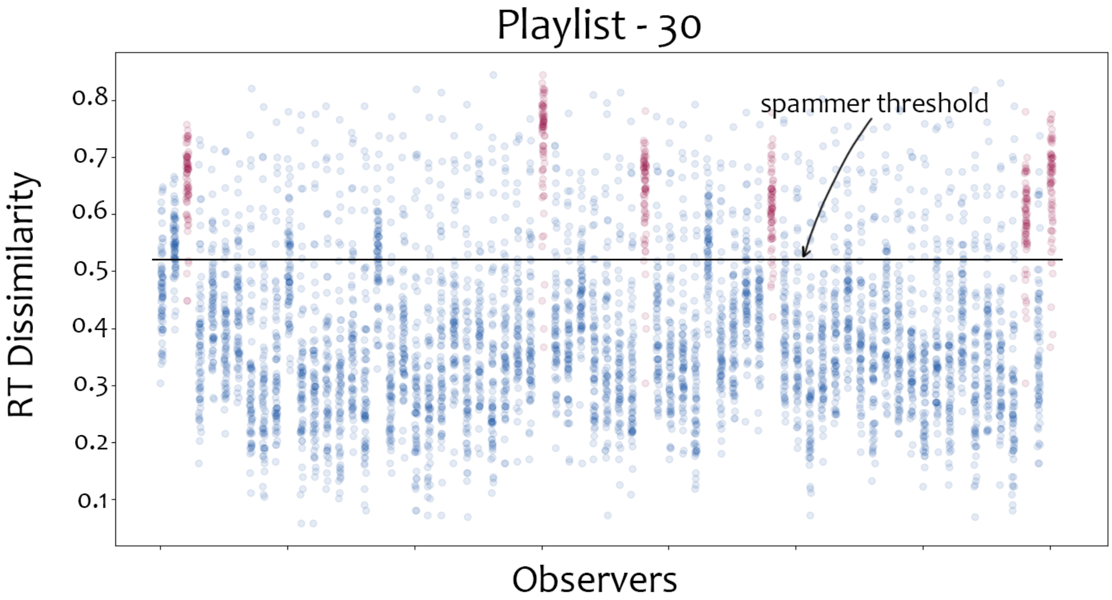


Figure 5.3 – RT dissimilarity distributions of observers in playlist-30. Observers are split into individual columns and observers who show similarity to synthetic spammers are displayed with magenta color.

As the first step, behavioral tools are applied in order to filter out observers with unreliable behaviors. Among 3500 participants, 49 were rejected due to golden unit check, 13 due to voting pattern check, and 56 were rejected due to voting speed.

Then, functions 5.1 and 5.3 are used together to calculate the RT dissimilarities. For each playlist in the experiment, 30 pairs in each playlist, observers' preferences are converted into binary form. With function 5.3, weight of each stimuli in each playlist is calculated. Since the ambiguity of each pair affects the RT dissimilarities, each playlist is treated separately. For each playlist, we generate 1000 synthetic spammers with the four spammer profiles introduced in Section 5.3.1. RT dissimilarity between each generated spammer and every other real observer in the corresponding playlist is calculated. After gathering $\hat{N}$ RT dissimilarity values for each generated spammer, 10 percentile of the $\hat{N} \times 1000$ RT dissimilarity values are used to create an expected unreliable behavior threshold.

Figure 5.3 presents the RT similarity of each observer with every other observer in playlist 30. The estimated spammer threshold (10_{th} percentile of the generated spammers' RT dissimilarity) is shown with a vertical black line over the plot. 6 observers were found to be unreliable by having 80% of their RT dissimilarities above the threshold, and thus

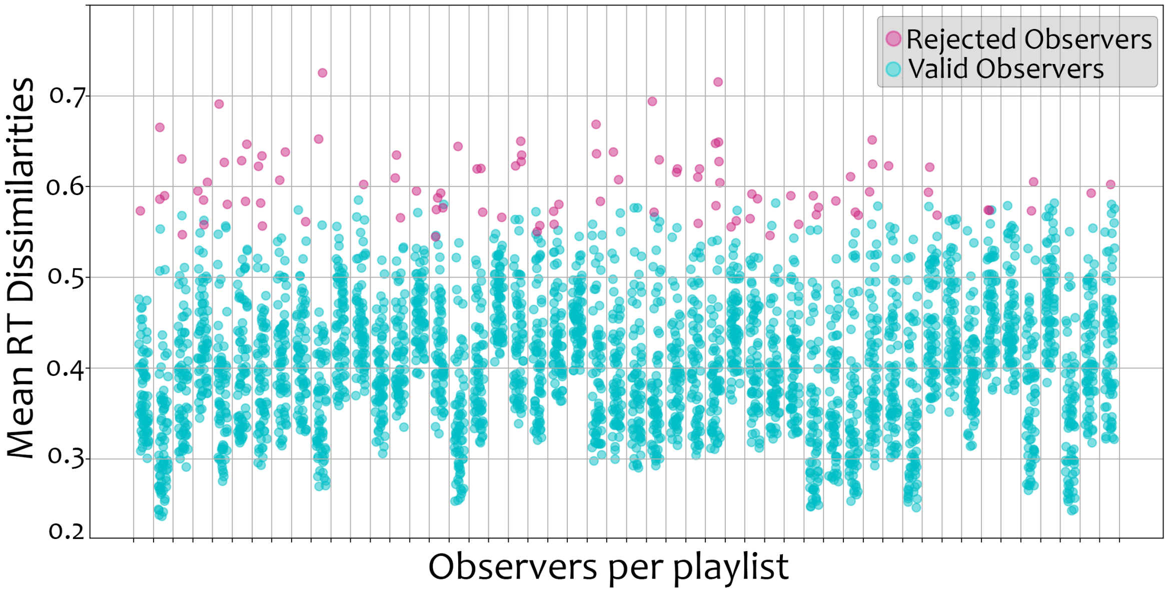
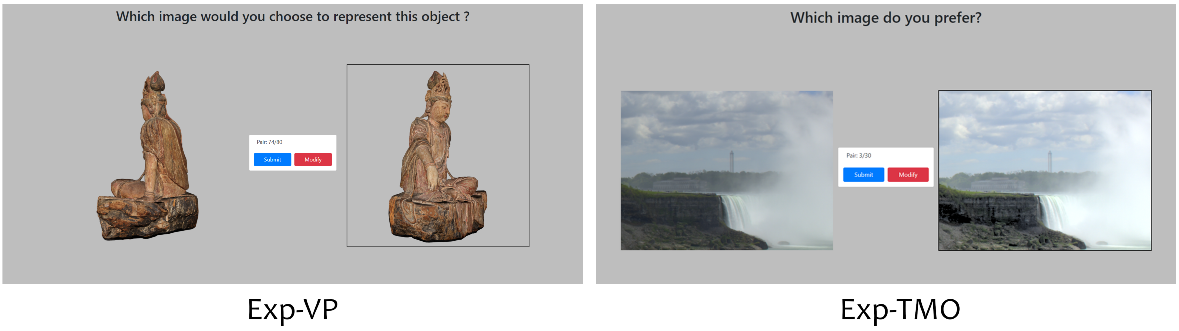


Figure 5.4 – Mean RT dissimilarity values of observers. Observers are grouped on horizontal axis by their corresponding playlists.

their subjective preferences are not included in the final dataset. These observers are color-coded with magenta color for ease of reading.

Similarly, all 50 playlists in the experiment were analyzed, and in total, 96 observers were rejected due to a higher similarity of the RT dissimilarities with the generated spammer profiles. Since individually displaying all playlists would be Figure 5.4 shows the mean RT dissimilarities of all observers. Observers are grouped into their corresponding playlists on each column. Rejected observers are displayed with magenta color. We can observe that each playlist has a different distribution of mean RT dissimilarities. This indicates that a fixed threshold for all types of content is not sufficient to define spammer behavior. In each individual playlist, we can see that the rejected observers are well separated from the rest of the observers in the playlists with few exceptions. Furthermore, the threshold of 80% similarity with the 10_{th} percentile range of synthetic spammer is intuitive and can be adjusted based on how strict the requirements are.


Which image would you choose to represent this object ?

Pair: 74/80

Which image do you prefer?

Pair: 3/30

Exp-VP

Exp-TMO

Figure 5.5 – Sample screenshots from Exp-VP and Exp-TMO experiments.

5.4 Synthetic Spammer Detection on 2 QoE Tasks

Further evaluation of the proposed methodology is done through synthetic spammer detection on two different QoE tasks. The main difference between the two experiments is the QoE task. The subjectivity of the task affects the inter-observer agreement. Since RT dissimilarity or any other statistical measure relies on the agreement between the observers, the comparison of two tasks provides a valuable insight. Details regarding the experiments are given in Section 5.4.1. After introducing the experiments, we analyze the effect of spammer proportion and intensity of the spammer behavior on the RT dissimilarities.

5.4.1 Subjective experiment designs

While sharing the same experimental methodology, the two experiments differ in terms of stimuli, research question, and QoE task. The first experiment, Exp-TMO, is conducted with traditional 2D images on a highly subjective task, *i.e., aesthetic quality assessment of tone mapped HDR images*. The second experiment, Exp-VP, is conducted on rendered views of 3D objects to select the most representative view of each object, *i.e., 3D viewpoint subjective preference.*

Conducted experiments share the following fundamental differences: the subjectivity of the questions directed to observers, source content being used, the purpose of the collected data. Both experiments were conducted through Prolific [88] crowdsourcing platform with the same recruitment pool. In each experiment, we followed a pairwise comparison methodology. Sample screenshots for both experiments are presented in Figure 5.5.

Dataset & Stimuli Generation

Exp-TMO: 20 HDR crops with $640 \times 480px$ from Fairchild HDR dataset [24] were used as source content. Similar to the large-scale dataset introduced in Chapter 4, we used 4 different TMOs from the literature to tone map the source content; namely, SemanticTMO[35], KimkautzTMO[56], KrawczykTMO[60], and ReinhardTMO[90]. Without any cross-content evaluation, with 20 SRCs and 4 HRCs, we generate 80 tone mapped stimuli which results in 120 image pairs for comparison.

Exp-VP: 21 high-resolution 3D meshes with color information were used to generate stimuli. 3D meshes belong to 4 different semantic categories: human, art, animals, and objects. Each 3D mesh was rendered from 4 different angles to generate 4 viewpoints with 90 degrees rotations. Each view was rendered to fit in a $600 \times 600px$ resolution window. 21 SRC with 4 viewpoints provides us with 84 rendered images and 126 image pairs for comparison.

Experiment Setup & Participant Recruitment:

Both experiments use a side-by-side formation to display stimuli as depicted in Figure 5.5. Both experiments were conducted on Prolific crowdsourcing platform [88] with observers recruited through Prolific participant pool. Display resolution was limited to 1080p to ensure a similar viewing condition for each observer. No time limit was set for both experiments. Due to the lower attention span of observers in crowdsourcing experiments, each experiment was split into smaller playlists to shorten the experiment duration. 100 unique observers evaluated each stimulus in both datasets.

5.4.2 Influence of spammer proportion

This analysis aims to measure the influence of the proportion of spammers on the RT dissimilarity measure. Synthetic spammers are generated based on the spammer profiles introduced in Section 5.3.1. The adjustable parameters in each spammer profile were fixed to 80%. We systematically increased the proportion of spammers inserted into each experiment and calculated the RT dissimilarity at each incremental. This allows us to analyze the discriminative power of RT dissimilarity with varying spammer proportions.

Figure 5.6 presents the results. Each experiment is plotted separately. The horizontal axis represents the spammer proportion in each plot, whereas the vertical axis represents the RT dissimilarity. Solid lines indicate the mean RT dissimilarity of real observers, while dashed lines represent synthetic spammers' mean RT dissimilarity.

We can observe that the spammer RT dissimilarities are similar between the two

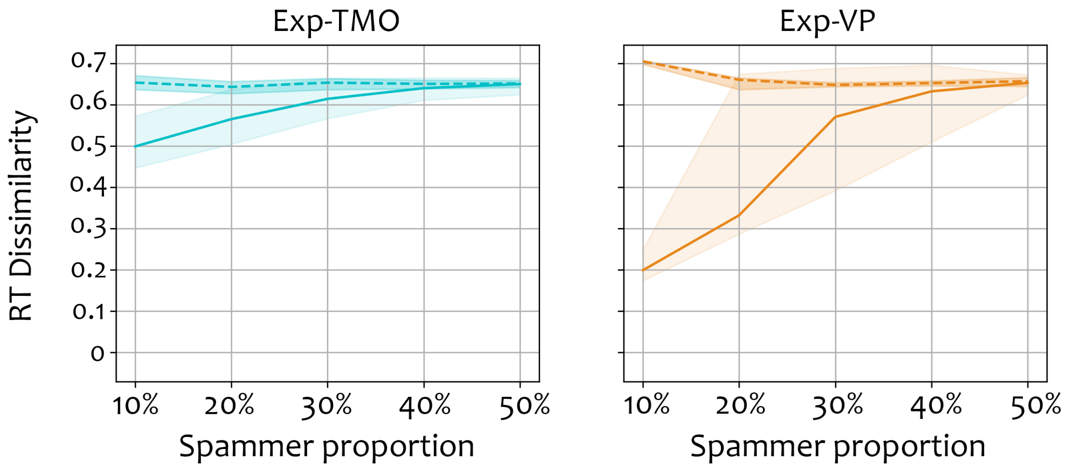


Figure 5.6 – Mean and 75% percentile range of RT dissimilarity values of real observers and spammers for varying proportion of spammers. Solid and dashed lines represent the real observers and spammers respectively for each experiment

experiments while real observer RT dissimilarities depict a different behavior. At lower spammer proportions, the RT dissimilarity of real observers in Exp-VP is much lower than real observers in Exp-TMO. This can be explained by the low subjectivity of the Exp-VP task when compared to Exp-TMO. With increasing spammer proportion, we observe that the mean RT dissimilarities of real observers and synthetic spammers are getting closer. At 40% spammers, the mean RT dissimilarity of real observers overlaps with synthetic spammers. It indicates that identifying spammers with RT dissimilarity measure becomes quite difficult where 40% of the observers are a spammer. Due to the higher subjectivity of the QoE task in Exp-TMO, 30% is sufficient to blend the synthetic spammers among real observers.

5.4.3 Influence of spammer behaviour intensity

A spammer may not have a malicious goal from the beginning of a subjective experiment, and she may provide honest opinions until he loses attention or gets bored. Therefore the intensity of spammer behavior may vary from person to person. As explained earlier in the Section 5.3.1, we control the intensity of each spammer profile with an adjustable parameter.

In order to analyze the effect of spammer behavior intensity on mean RT dissimilarity measures, we fixed the spammer proportion to 20% for each experiment. We systemati-

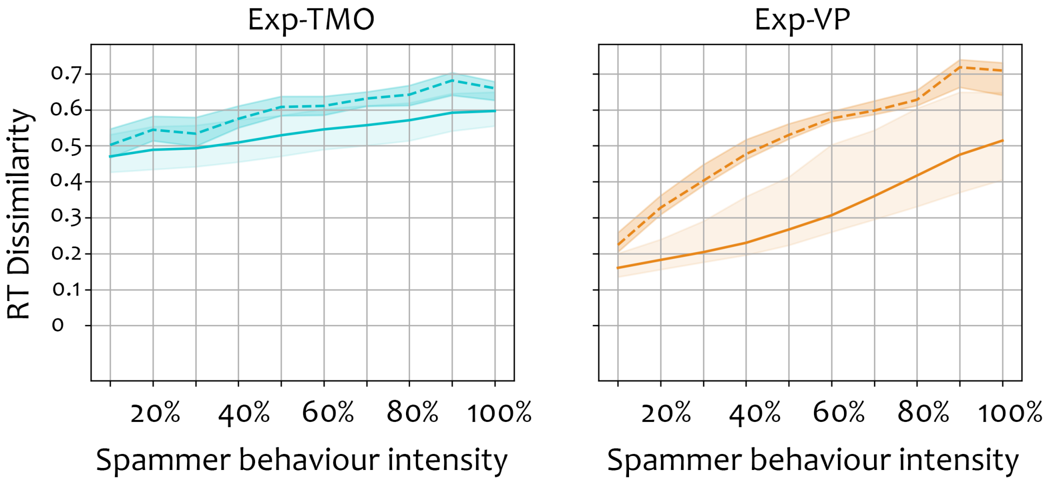


Figure 5.7 – Mean and 75% percentile range of RT dissimilarity values of real observers and spammers for varying spammer behavior intensity. Solid and dashed lines represent the real observers and spammers respectively for each experiment

cally increased the spammer behavior intensity and calculated the RT dissimilarities of real observers and synthetic spammers at each incremental. Figure 5.6 presents the result of the analysis. For each experiment, mean RT dissimilarities of real observers and synthetic spammers are plotted separately. In each plot, the solid line represents the mean RT dissimilarity of real observers, whereas the dashed line represents the mean RT dissimilarity of synthetic spammers. Also, 75% percentile ranges of the RT dissimilarities are displayed as a region around each line.

An important observation from the results is that the overlap between 75% percentile ranges between real observers and synthetic spammers are lower in Exp-VP with varying spammer behavior intensity. Similar to the spammer proportion analysis, this can be explained by the lower subjectivity of the QoE task in Exp-VP. We also observe that the overlap between 75% percentile ranges of real observers and synthetic spammers decreases with higher spammer behavior intensity in Exp-VP.

5.5 Discussion

In this chapter, we provided a set of behavioral tools and a novel methodology to identify unreliable observers. Furthermore we conducted two experiments with different QoE tasks and minimal experimental design differences to understand the impact of the task subjectivity on the RT dissimilarity measure.

Behavioral tools are powerful at identifying the targeted spammer profiles. However, spammers without a pre-defined behavior are practically undetectable with behavioral tools. Therefore, there is a need for statistical measures to identify unreliable observers without any categorization of the spammer profiles. This is especially challenging in PC experiments due to the binary nature of subjective preferences. To this end, we propose a novel methodology that relies on Rogers-Tanimoto dissimilarity measure to detect unreliable observers. By first applying the behavioral observer screening tools on the collected subjective preferences, we increase the overall inter-reliability of the remaining reliable observers. This allows us to isolate the observers with high dissimilarity to the rest of the crowd.

Our findings on comparison of the two QoE tasks indicate that the inter-observer agreement is highly task-dependent. Therefore, statistical measures indicating a general "good" or "bad" agreement level can only be used relatively between the tasks. In order to increase the robustness to task differences, simple thresholding of agreement measures should be avoided. Additionally, the subjectivity of the QoE assessment tasks influences the spammer tolerance of agreement measures. QoE assessment tasks with higher subjectivity should use additional precautions, such as golden units, to decrease the overlapping range of agreement values between spammers and real observers. Finally, while providing insight, using mean agreement value to detect spammers may not be sufficient. Methods should utilize approaches that can benefit from the measures between individual observers rather than relying single agreement value for each observer.

Part II

Objective Quality Evaluation of Multimedia Content

Chapter 6

Objective Quality Assessment

6.1 Objective Quality Metrics

As discussed in the first part of the thesis, specifically in Chapter 1, subjective evaluation of image quality is the most reliable option. On the other hand, it also has several disadvantages. Conducting subjective studies is time-consuming and costly. It is not possible to conduct subjective studies for real-time needs, such as optimizing image processing tools based on image quality. Therefore, there is an undeniable need for algorithms to assess image quality without conducting subjective experiments. Such algorithms are commonly known as quality metrics and can be referred to as quality indexes, measures, or models. This thesis will mainly refer to these algorithms as image quality metrics and accept other terminologies equal.

Objective quality metrics are often categorized into three groups based on the presence of the reference image as input. Full-reference (FR) metrics require access to the reference stimuli while measuring the quality of the distorted stimuli. Reduced-reference (RR) metrics only require a set of features from the reference image, whereas no-reference (NR) metrics do not require any access to the reference image to measure the distorted image quality.

Objective quality metrics are expected to provide a good correlation with the subjective quality evaluation results, and they are often developed for specific applications. For example, a metric that provides a high correlation with subjective opinions at measuring the quality of compressed images may not perform as well on evaluating the quality of depth-based image rendering (DIBR) algorithms.

This chapter aims not to review all of the existing quality metrics in the literature but to provide the necessary foundation for the experimental work in the following chapters. In order to do so, we introduce a set of metrics for the multimedia content covered in the following chapters. Moreover, we discuss the methodologies used for the performance evaluation of quality metrics.

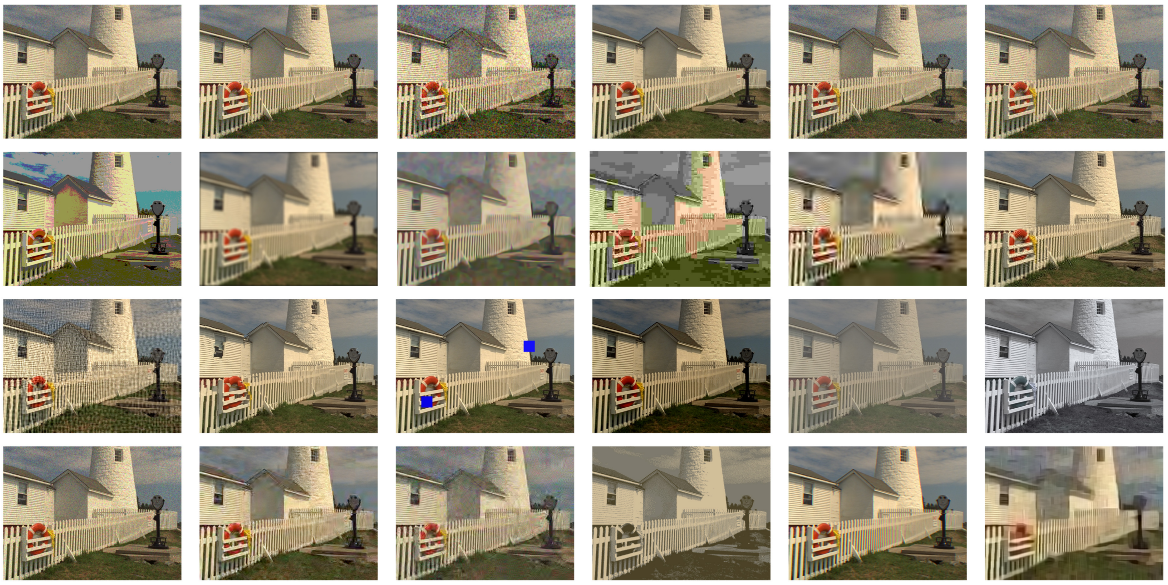


Figure 6.1 – Sample images with all the distortions available in TID-2013 dataset[85]

6.1.1 Image quality metrics

In this section, we will introduce general-purpose image quality metrics. These metrics are often expected to be able to handle various kinds of distortions such as JPEG compression artifacts, Gaussian blur, or simply noise. For example, TID-2013 dataset[85] contains 24 different distortions related to acquisition, transmission and compression as shown in Figure 6.1.

For the mentioned distortions, the oldest and the most developed type is FR metrics. They measure the similarity (fidelity) of the distorted images with respect to reference images. A measure of similarity between the two images is a good indicator of the perceived quality for many applications. A comprehensive overview of the image quality metrics can be found in [127].

Probably the most commonly used metrics are MSE and PSNR. Despite their wide acceptance as an FR metric, they often perform poorly on IQA due to a lack of consideration of image and HVS characteristics.

On another front, FR metrics often exploit our knowledge about HVS or image properties. Structural similarity index (SSIM) [119] is the most popular and one of the simplest of such metrics. It measures the image quality in terms of luminance, contrast, and structure. Prior to similarity estimation, the image is divided into small patches, and the final estimation is often indicated by the mean value of the similarity of patches in terms of all three factors (luminance, contrast, and structure).

Multi-scale structural similarity index (MS-SSIM) [117] later introduced to improve SSIM by calculating it over multiple scales. Each scale is obtained by filtering the previous scale with a low-pass filter and down-scaling with a factor of 2. Ideally, the number of scales can be estimated based on the viewing distance and image resolution, however typically set to 5.

Visual information fidelity (VIF) [101] is another metric utilizes a multi-scale decomposition. The natural scene statistic (NSS) model is used to describe the reference and distorted images. In addition, an HVS channel is used, and a proposed VIF measure is used to quantify the quality difference between reference and distorted image pair. Note that VIF scores can exceed the expected upper limit 1, indicating a higher quality for the "distorted" image.

Feature similarity index (FSIM) [129] uses two low-level features (phase congruency, gradient magnitude) to measure the difference between reference and distorted images. Also, FSIMc was proposed as an extension to FSIM, which utilizes the color information in YIQ color space.

HDR-VDP 2.2 [74] is a metric that uses a complicated HVS model to predict the physical difference between the reference and distorted image. The HVS model in HDR-VDP 2.2 includes intra-ocular light scatter, photoreceptor spectral sensitivity, luminance masking, and achromatic response estimation to model the optical and retinal pathway of the HVS. Furthermore, it uses a multi-scale decomposition and contrast sensitivity function (CSF) and takes several masking effects into account. Thanks to the detailed HVS model, it can account for various display parameters and viewing conditions. The metric finally outputs an error visibility map, a simple visibility score, and a quality score prediction for the distorted image. Later, the metric further improved to its third generation HDR-VDP 3[1]; however, the study has not been published yet.

6.1.2 Light field quality metrics

Despite the recent advancements in light field acquisition, processing, and display technologies, quality assessment of light field content is still not fully explored. While existing 2D image quality metrics can be used to assess the spatial quality of the light

[1] available at: https://sourceforge.net/projects/hdrvdp/files/hdrvdp/

field content, assessing the quality of the light field in angular domain and in a cumulative way remains challenging. To this end, several objective quality metrics were proposed in the last decade. In this section, we will introduce a few selected objective quality metrics for light field quality evaluation. Note that this is not an exhaustive list of state of the art but rather an overview to provide necessary background on light field quality metrics.

BELIF [103] is a no-reference light field quality metric that relies on tensor spatial characteristic features for spatial quality and tensor structure variation index for angular consistency. Tucker decomposition is utilized prior to feature extraction in order to reduce the redundancy within a light field content. It is developed for dense light field images, and its performance has not been reported for sparse light field content.

SDFM [111] is proposed as a full-reference light field quality metric that utilizes sub-aperture views to extract symmetry and depth features to quantify light field quality. Similarly, the model is developed and evaluated on dense light field datasets.

Fang et al. proposed a full-reference light field quality metric that relies on gradient magnitude similarity of reference and distorted epipolar plane images[26]. Two directions as horizontal and vertical, are used to quantify the gradient similarity.

6.1.3 Point cloud quality metrics

Existing commonly used point-based metrics can be categorized based on three main approaches: point-to-point [76], point-to-plane [108], and plane-to-plane [9] differences in 3D space. The point refers to each point in the point cloud, whereas the term “plane” refers to the plane of a point defined by its normal vector. The missing point normals were estimated using Matlab’s `pcnormals` function. The geometry metrics are computed using either root mean square (RMS) distance, mean square error (MSE), or Hausdorff distance measures. Minimum, mean, and median are also used to pool the difference scores.

In addition to geometry differences, color differences are also calculated using point-to-point correspondence. MSE or PSNR is calculated from the differences between the corresponding points’ assigned color values. These color metrics are calculated for Y, U, and V channels.

Alternatively, image quality metrics can be used on the rendered point clouds for quality assessment. For volumetric videos, temporal pooling methodologies can be used to estimate the final quality score from the estimated quality of individual frames.

6.2 Measuring quality metric performances

Although objective quality metrics provide a low-cost and fast way for quality evaluation of multimedia content, they cannot be an alternative to subjective evaluation unless their performance is validated. Performance evaluation of the objective quality metrics needs to be done on a representative dataset for the given QoE scenario.

ITU standardization group provides a set of tools for performance evaluation of objective quality metrics[49]. VQEG group also provides a report with recommendations[52]. In this section, we will provide some of the recommended methodologies as well as an alternative method proposed by Krasula et al.[59].

6.2.1 Recommended measures

ITU-T P.1401 [49] recommendations provide a set of guidelines for performance evaluation of quality metrics. Before measuring the performance, it is recommended to apply a mapping between predicted quality scores and ground truth MOS values. Monotonic mapping procedures, such as linear mapping, third order polynomial, or logistic mapping, are often used. In order to fit the predicted scores into the same range with MOS values, root mean squared error (RMSE) is minimized between the mapped values and MOS scores.

Since the following measures are commonly used in the domain and explained by recommendation documents, we introduce the measures without going into detail. We recommend interested readers to refer to the recommendation documents [49, 52].

Pearson's correlation coefficient (PCC): measures the linear relation between the objective quality metric scores and MOS values. The ideal relation between the mapped objective quality metric predictions and MOS is expected to be linear. PCC values range $[0, 1]$ where higher values indicate a better correlation.

Root mean squared error (RMSE): is used to measure the objective quality metric accuracy. It is calculated as the mean square root of the difference between MOS and predicted quality scores. Lower RMSE values indicate better performance, and the range of RMSE values depends on the range of MOS values.

Outlier Ratio (OR): is another alternative to measure the objective quality metric accuracy. It is defined as the ratio of the number of mapped scores outside the confidence interval, and lower OR indicates a higher accuracy.

Spearman's rank-order correlation coefficient (SROCC): is a non-parametric measure of objective quality metric performance. It does not require mapping the predicted objective quality metric scores onto the MOS range since it relies on ranking the stimuli rather than the numeric values.

Kendall's rank-order correlation coefficient (KROCC): is another alternative where the rank order of the stimuli is used for determining objective quality metric performances. For KROCC, all possible pairs of stimuli are checked whether MOS values and objective quality scores agree on the rank of the stimuli in the pair.

6.2.2 Krasula model

In the Krasula model, performance evaluation of the objective quality metrics is conducted in two different stages. The first stage focuses on the ability of how good the IQA metrics are at distinguishing between pairs with and without statistically significant difference. The second stage aims to determine whether the metrics are able to recognize the image with higher preference in pairs with a statistically significant difference.

In the first stage, *different vs. similar analysis*, image pairs are split into two groups as different and similar based on the statistical significance of subjective preferences. Metric score differences of pairs are used to evaluate the metric performance. Ideally, the difference between the predicted quality scores should be higher for the image pairs with a statistically significant difference. Inversely, metric score differences for pairs without a significant difference are expected to be low. An example of the ideal distribution is visualized on the left plot in Figure 6.2. To determine the abilities of metrics in different-similar binary classification scenario, Receiver Operating Characteristic (ROC) [106] is used in the Krasula model. The performance of the classifier can then be quantified by Area Under the ROC Curve (AUC).

Similarly, the better and worse analysis evaluates the metric performances based on the differences of estimated quality scores. In this analysis, the aim is to determine whether the metrics are able to correctly recognize the higher quality image in a pair. An example of the ideal distribution of metric score differences is depicted on the right plot of Figure 6.2. Stimuli are split into two groups as better and worse in the pre-processing stage. Alternatively, stimuli orders can be swapped, and all significantly different pairs can be used in both categories. Based on the metric score differences, AUC analysis can be carried out. Another way to analyze is the correct classification percentages of better and worse categories.

When comparing multiple objective quality metrics, it is essential to determine if the

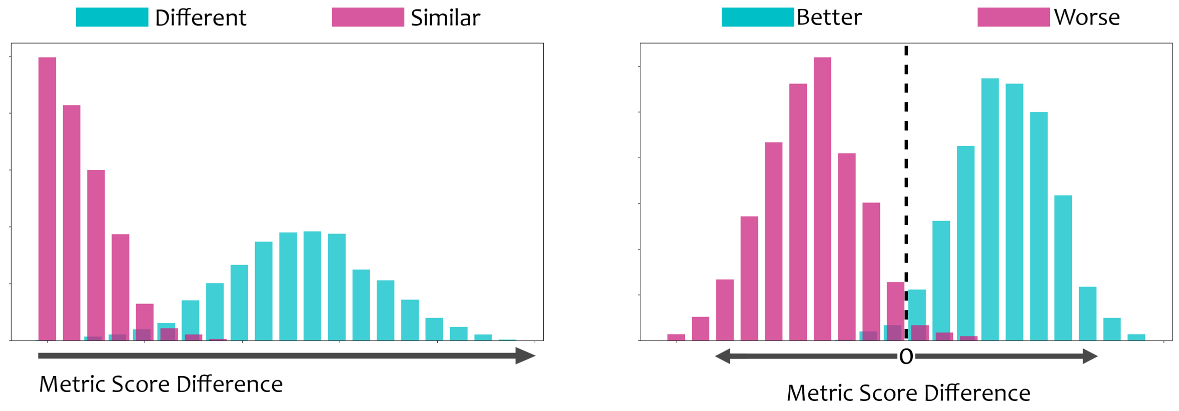


Figure 6.2 – Examples of the ideal distributions of metric score differences for the two evaluation scenario. Different vs similar analysis on the left, better vs worse analysis on the right.

differences in metric performances are statistically significant. Krasula model relies on the method proposed by Hanley and McNeil[41]. A critical ratio c_{mn} is calculated between the AUC and the standard error (SE) of AUC of the metrics m and n with the following function:

$$c_{mn} = \frac{AUC_m - AUC_n}{\sqrt{{SE_m}^2 + {SE_n}^2 - 2rSE_mSEn}}, \tag{6.1}$$

where standard error [42] of each metric's AUC are as follows:

$$\sqrt{\frac{AUC(1 - AUC) + (N_1 - 1)(AUC/Q_1 - AUC^2) + (N_2 - 1)(Q_2 - AUC^2)}{N_1N_2}}, \tag{6.2}$$

where N_1 and N_2 are the number of stimuli in each group in the ROC analysis (different/similar or better/worse). And Q_1 and Q_2 are

$$\begin{aligned} Q_1 &= AUC/(2 - AUC), \\ Q_2 &= 2AUC^2/(1 + AUC). \end{aligned} \tag{6.3}$$

Finally, the probability of the difference between the AUC of two metrics m and n can be determined as the cumulative distribution function of the $cdb(c_{mn})$.

To determine the difference between correct classification percentages, the Krasula model utilizes Fisher's exact test[27]. Furthermore, it relies on Benjamini-Hochberg model [13] in the cases where more than two metrics are being evaluated to compensate for the type I error propagation.

6.3 Conclusion

This chapter introduced the concepts and prior work related to the contributions made in the following chapters regarding the objective quality evaluation of immersive multimedia content. We first discussed the existing approaches on traditional image quality assessment that allow us to build a relation between just noticeable differences and image quality in Chapter 7. Moreover, we provided a brief overview of the previous work on light field and point cloud quality assessment to set the ground for the proposed work on the following chapters.

On another front, we discussed the existing methodologies for performance evaluation of objective quality metrics. First, we introduced the commonly used correlation measures. Furthermore, we introduced the Krasula method as a more suitable alternative methodology for real life image quality assessment scenarios. Additionally, it has the benefit of the ability to combine multiple datasets for evaluation which reduces the bias and increase the reliability of evaluation. Ideally, raw subjective opinion scores are needed to utilize Krasula method. If raw scores are not available, at least the standard deviation and number of observers information are needed along the MOS. In the following chapters, we will utilize the Krasula method when it is possible. In other words, when the required information is available in the dataset. In other cases we will use the recommended correlation measures.

Chapter 7

Can Just Noticeable Differences be Used to Understand Quality Range

7.1 What is Just Noticeable Difference?

As discussed in the earlier parts of the thesis, subjective quality evaluation of images commonly relies on collecting opinion scores from a set of observers. Then MOS can be used to represent the image quality on a continuous scale. HVS, on the other hand, does not perceive quality in a continuous fashion but rather as a staircase function. In other words, small shifts on a continuous quality scale may not be perceived by the observer.

To this end, just noticeable difference (JND) provides a binary measurement to quantify the perceptual differences between a given image pair. JND defines the minimum amount of degradation required to be perceived by the observers, and it is constant for a given content in a given viewing condition for a given observer.

We hypothesize that, for a given content, distortion type, viewing condition, and level of distortion, the proportion of individuals with a JND threshold greater than the distortion level is an indicator of the quality of the distorted image compared to the pristine image.

7.2 JND-based Image Quality Datasets

The last decade brought a surge into JND-based image quality datasets. A non-exhaustive list of publicly available datasets is given in Table 7.1. Although listed datasets are collected via laboratory experiments, they provide a large amount of PVSs thanks to the high number of QP levels of compression algorithms. All of the datasets use PC

Table 7.1 – Non-exhaustive list of JND-based image/video quality datasets in the literature.

	nb of SRC	nb of PVS	nb of Observations
MCL-JCI [54]	50	5000	30
VVC-JND [102]	202	7878	20
MCL-JCV [115]	30	1530	50
VideoSet [116]	220	44800	30
SIAT-JSSI [25]	10	3510	36
JND-Pano [72]	40	4000	25
QAD-HEVC [46]	40	2040	30

methodology to collect subjective annotations on the visibility of the compression artefacts of still images or videos.

MCL-JCI [54] dataset collects a set of JND points for 50 still images with JPEG compression artefacts. The number of JND points collected per stimuli varies among observers. A statistical approach later proposed to merge multiple JND levels into a common JND step per SRC. The authors also provided raw subjective JND steps alongside estimated JND steps.

VVC-JND [102] dataset relies on Versatile Video Coding (VVC) with QP values ranging from 13 to 51 on 202 still images. PC methodology is used with a side-by-side presentation of the stimuli to collect JND step information.

MCL-JCV [115] dataset collects multiple JND steps for 30 videos on compression artefacts of H.264/AVC encoding algorithm with 51 QP levels.

VideoSet [116] dataset collected on 220 videos of 5 seconds length with varying spatial resolutions (*i.e.,* 4096 × 2160, 4096 × 1714, 3840 × 2160), frame rates (*i.e.,* 60, 30, 24 fps.) and color formats (*i.e.,* YUV444p, YUV422p, YUV420p). Although variations were lowered in pre-processing stage, dataset provides a large number of SRC with wide range of properties. H.264/AVC encoding is used to compress images with QP values ranging between 8 and 47. 3 JND points were provided for each SRC.

JND-Pano [72] dataset contains panoramic still images with first JND levels of the JPEG compression artefacts. Observers have equipped head-mounted displays with the freedom of changing the field of view during the experiment.

QAD-HEVC [46] dataset contains 40 videos of 5 seconds length compressed with HEVC encoder (51 QP levels). First, JND points were collected with binary-search fashion proposed in MCL-JCI dataset[54].

Figure 7.1 – 50 SRCs used in MCL-JCI [54] dataset.

7.2.1 MCL-JCI dataset

The proposed model (D-JNDQ, see section 7.3) is developed based on the first JND step information provided in MCL-JCI [54] dataset. Therefore, before introducing the model, we investigate the dataset in detail.

The dataset aims to measure the number of distinguishable quality levels among the JPEG compression intensities (*i.e.,* QP levels ranging from 1 to 100.). 50 SRCs (with $1920 \times 1080px$ spatial resolution) were collected for this purpose from 10 different semantic categories. Data collection was done in a laboratory environment with a 65" display with a native resolution of $3840 \times 2160px$. The viewing distance was set to 2 meters (1.6 times the picture height). 30 unique observers provided a set of JND points for each SRC. In total, more than 150 volunteers participated in the experiment, of which 10 of them were experts in the field of image quality assessment.

JND step search is conducted consecutively. For a given SRC and a given observer, a pristine image is used as the first anchor point in search of the first JND point. Once the first JND step was determined, it was used as an anchor point for the second JND step. This process continued until the search was terminated. To generalize, to find the n^{th} JND step (JND_n), $(n-1)^{th}$ JND step (JND_{n-1}) was used as an anchor point.

JND step search was terminated in two cases. The first case is when the difference in QP levels between two neighboring JND steps is equal to one. In other words, when there is one QP level difference between JND_{n-1} and JND_n. The second case is when the observer finds a noticeable difference between the anchor point and k^{th} QP level (QP_k)

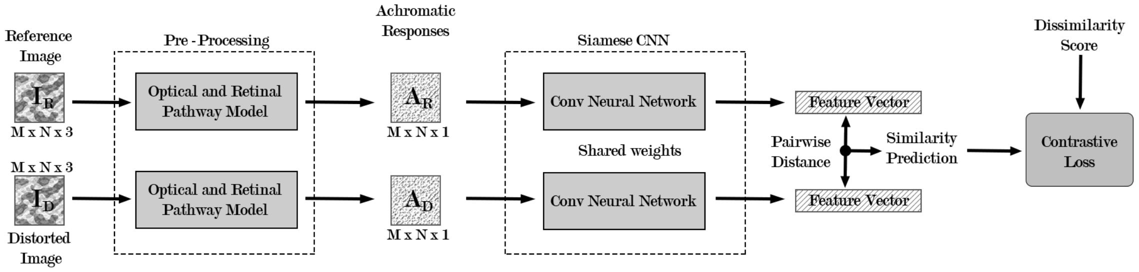


Figure 7.2 – Diagram of the proposed model where $I_{R/D}$ indicates the input images, and $A_{R/D}$ denotes the achromatic responses.

whereas cannot find a noticeable difference between the anchor point and the $(k+1)^{th}$ QP level (QP_{k+1}).

Although the dataset provides a statistical methodology to merge multiple sets of JND steps (acquired from 30 unique observers for each SRC) into a single JND staircase function, we rely only on the raw first JND step information (first JND step per SRC per observer) from the dataset. Therefore, we will not go into detail about the suggested post-processing stage in the paper. Interested readers are recommended to refer to the original paper[54]. Information regarding our proposed processing methodology of the first JND steps is given in Section 7.3.4.

7.3 D-JNDQ : Learning Image Quality from JND

7.3.1 Model overview

HVS can mainly be split into four broad parts as the optical, retinal, lateral geniculate nucleus, and visual cortex processing[29]. In the proposed framework, we simplify our approach by dividing this complex process into two. We first use an existing Optical and Retinal Pathway model to pre-process input images, *i.e.*, the Optical and Retinal Pathway proposed in HDR-VDP 2[74]. This module provides an estimation of the achromatic responses for displayed images. Optical and Retinal processing of HVS highly affects the visibility of distortions. Hence, including this module as a pre-processing tool simplifies the similarity prediction. After acquiring achromatic responses of both the reference and distorted images, the remaining task is to predict the similarity between the achromatic responses inputs.

Regarding its proven success in visual similarity and pairwise ranking prediction tasks

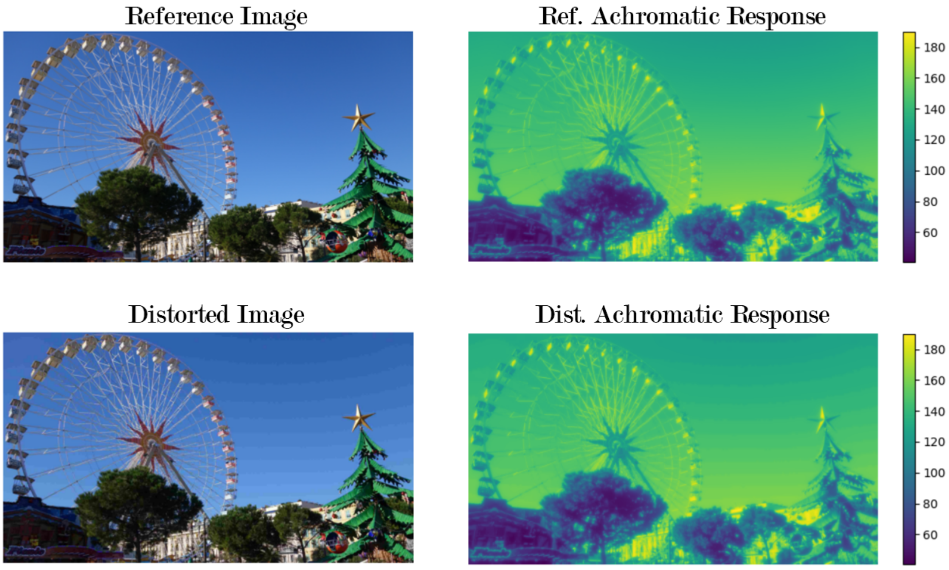


Figure 7.3 – Reference and distorted image (QP=17) with corresponding achromatic responses for SRC-7 in MCL-JCI[54]

.

[93], Siamese CNN was employed to predict the similarity between input pairs. In general, Siamese networks are equipped with two or more identical networks with shared weights to learn the embedding between a pair or triplet of input data.

The overall structure of the proposed model [1] is shown in Figure 7.2. All the achromatic responses are acquired by pre-processing input RGB images with the optical and retinal pathway model from HDR-VDP 2. Then, they are fed into the Siamese CNN to extract their latent representation, *i.e., feature vectors*. Afterward, the pairwise distance between outputted feature vectors is calculated to compute a similarity score. During training, contrastive loss [40] is used between the predicted quality scores and the ground truth dissimilarity scores acquired from MCL-JCI dataset [54].

7.3.2 Optical and retinal pathway model

Optical and Retinal Pathway is modeled as a combination of 4 sub-modules in the HDR-VDP 2 [74]. The first module accounts for the light scattering that occurs in the cornea, lens, and retina. It is defined by a modulation transfer function (MTF) estimated via psychophysical studies. The second module calculates the probability of a

[1] model available on: https:/github.com/kyillene/D-JNDQ

photo-receptor sensing a photon at a corresponding wavelength. It outputs cone and rod responses of the input image. The third module mimics the non-linear response to light of the photo-receptors. It is modeled as a non-linear transducer function. The final module converts the non-linear responses into joint cone and rod achromatic responses by simple summation. An example of achromatic responses for a reference and distorted image from the MCL-JCI dataset [54] is presented in Figure 7.3.

By incorporating Optical and Retinal Pathway into the pre-processing stage, the masking effects occurring at this stage of the visual pipeline could be well considered. By enhancing or masking the distortions visibility with existing knowledge in the domain, the training complexity of the similarity network could be well simplified and accelerated. Nevertheless, it enhances the generalization of the model for tackling unseen distortion types and supra-threshold distortion values.

7.3.3 Siamese-Net for quality prediction

The Siamese network is utilized as a feature extractor without any fully connected layers. On top of this backbone, we directly compute the pairwise distances. This architecture facilitates arbitrary input resolutions. We design our Siamese network from scratch. It is consists of 5 convolutional layers with batch normalization, ReLu activation layers. To reduce the spatial resolution, a stride of 2 was adapted for the first 4 convolutional layers.

For the last layer of the network, a sigmoid activation function is employed without a stride. After flattening the output feature vector, they are then used to calculate the similarity score between the reference and distorted images.

7.3.4 Dataset and training details

After experimenting on the MCL-JCI [54] dataset, it was observed that the task of detecting the first JND steps (JND_1) and following JND steps (JND_n, where $n > 1$) are different. While identifying the JND_1, an observer tried to identify the difference between the reference and distorted image. However, for the later JND steps, this task gradually turned into a preference task, *i.e., which stimulus is preferred compared to the other*. More specifically, instead of "at which QP level the distortion becomes visible", the question evolved into "at which QP level the distortion becomes more disturbing". This observation encourages us to utilize only the first JND point for labeling the training dataset.

For training the proposed model, we used MCL-JCI [54] dataset (for the introduction

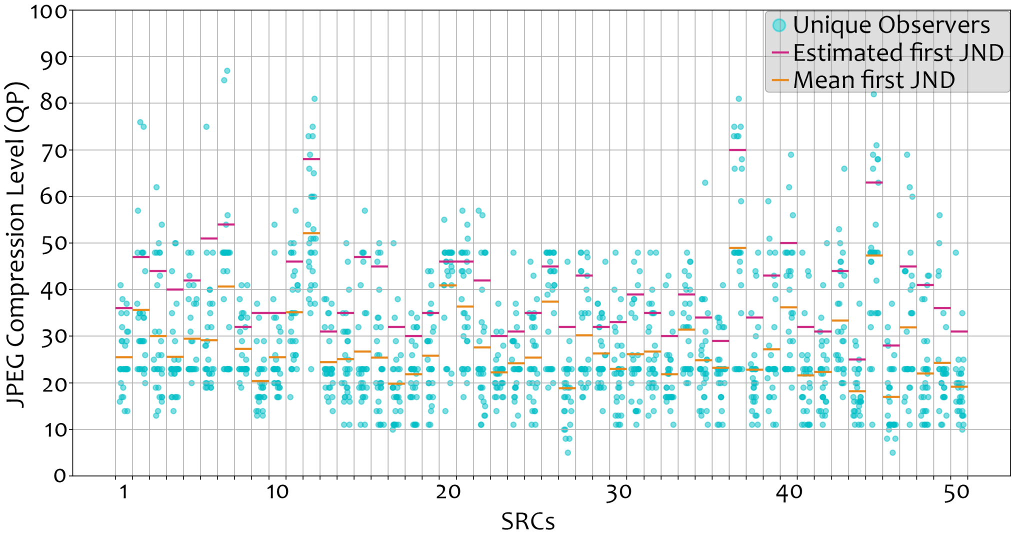


Figure 7.4 – Distribution of first JND steps for each SRC in the MCL-JCI dataset [54].

of the dataset, see Section 7.2.1). For a given SRC in the dataset, 30 unique observers provided this information. Figure 7.4 shows the distribution of *JND₁* for each observer for any given SRC in the dataset. Each point in the plot represents the QP levels (indicated on the vertical axis) of the *JND₁* for a unique observer for a given SRC (indicated on the horizontal axis). Yellow markers indicate the mean QP level for the *JND₁* for the given SRC, whereas magenta markers indicate the estimated QP level for the *JND₁* based on the methodology proposed in [54].

As can be observed from the figure, the perception of observers vary greatly. In other words, QP levels corresponding to *JND₁* vary greatly among observers. Therefore, instead of accumulating QP levels for *JND₁* from all observers into a single value, we merged individual opinions into a staircase function defining the *JND₁* similar to [3, 67]. A given SRC_k and compressed image with QP level QP_i were paired as P_{k-i}. A dissimilarity score (d_{k-i}, ranges in $[0, 1]$) was assigned to each pair P_{k-i} based on the percentage of observers with QP levels beyond QP_i corresponding to *JND₁*. In Figure 7.5, we present the acquired dissimilarity scores as a heat map. Each SRC is represented in a row with QP levels decreasing from left to right. Dissimilarity scores increase as the QP levels decrease.

After acquiring the dissimilarity score for each pair, as described in Section 7.3.2, each SRC (SRC_k) and compressed images are converted into achromatic responses using the Optical and Retinal Pathway model from HDR-VDP 2 [74]. The obtained achromatic

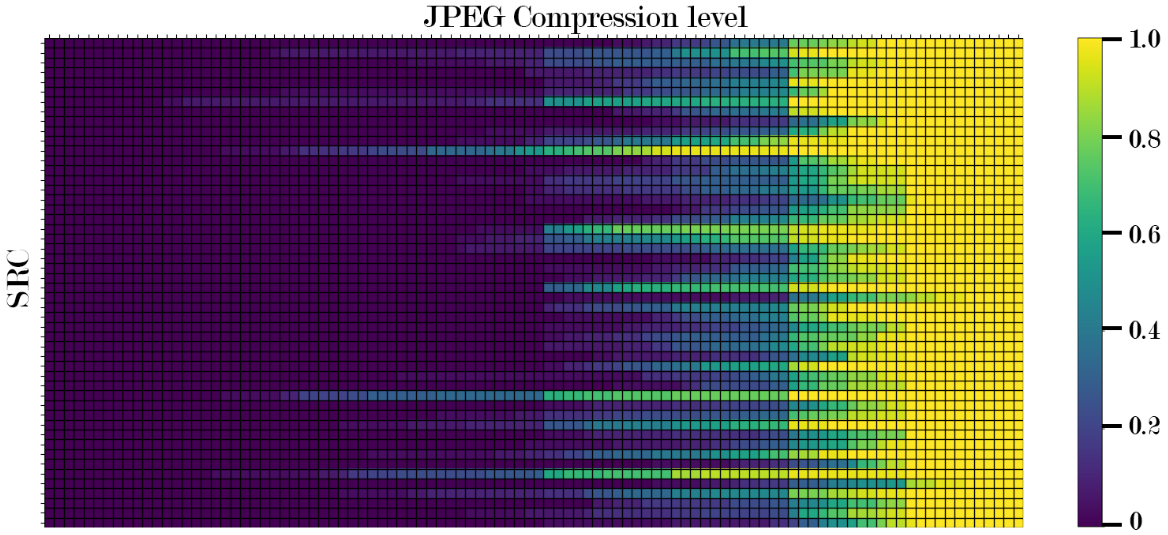


Figure 7.5 – Dissimilarity scores acquired by using first JND steps of each observer in MCL-JCI dataset[54]. Each row represents an SRC. Columns are ordered from the highest QP level to the lowest, left to right.

responses share the same spatial resolution with input images. However, pixel values are represented in a single channel, resulting in an array of size 1920 × 1080 × 1.

After pre-processing the dataset as described above, we conducted hyperparameter tuning for the Siamese network. Contrastive loss [40] function was used with a batch size of 32 during training. We found out that 0.03 learning rate with Adam optimizer provides us the best convergence speed and lowest validation loss with the final network structure. Finally, the Siamese network was trained for 100 epochs over the training dataset with the optimal hyperparameters. We experimented with weight decay and regularization terms during hyperparameter search; however, we observed no improvement in training convergence or model accuracy.

7.4 Performance Evaluation

7.4.1 Evaluation on TID-2013 dataset

It is worth mentioning that our model was trained only on JPEG distortions with the first JND. To prove the generalization of the proposed model on unseen distortions and novel supra-threshold distortion levels, we conducted a cross-dataset evaluation on the TID-2013 dataset [85]. TID-2013 dataset contains 24 different distortions, including but not limited to noise, blur, transmission error, compression distortions. They are categorized into 6 overlapping groups. In total, there are 3000 distorted images with varying distortion intensity and distortion types.

Correlation with MOS

We tested the model on all 3000 images without any pre-training. We used the scripts provided by the authors to calculate the correlation between the predicted results and the MOS. As such, correlation results are directly comparable with other metric correlations acquired by the authors. Table 7.4.1 reports the SROCC values of the proposed model and the other methodologies provided by [85]. The proposed model, *i.e.,* D-JNDQ, provides competitive results with the compared metrics in Noise, Actual and Simple categories and provides better results in the New and Color category of distortions compared to other evaluated metrics. The proposed model achieved the lowest performance on the subset of the Exotic category. This is mainly due to the preferential nature of the distortions in this category. Detecting the distortion plays a minimal role for distortions, such as local block-wise distortion, since the distortions are visible at all levels with different variations rather than different intensities. Therefore, we expected a poor prediction performance in this category, which also reduces the overall correlation results.

Table 7.2 – SROCC values for selected metrics in TID-2013

	Noise	Actual	Simple	Exotic	New	Color	Full
D-JNDQ	0.851	0.881	0.894	0.315	**0.842**	**0.813**	0.589
HDR-VDP 3	0.829	0.847	0.929	0.822	0.679	0.635	0.772
FSIM	0.897	0.911	0.949	**0.844**	0.649	0.565	0.801
FSIMc	0.902	0.915	0.947	0.841	0.788	0.755	0.851
PSNR	0.822	0.825	0.913	0.597	0.618	0.535	0.640
PSNRc	0.769	0.803	0.876	0.562	0.777	0.734	0.687
PSNRHA	**0.923**	**0.938**	**0.953**	0.825	0.701	0.632	0.819
SSIM	0.757	0.788	0.837	0.632	0.579	0.505	0.637
MSSSIM	0.873	0,887	0.905	0.841	0.631	0.566	0.787
VIFP	0.784	0.815	0.897	0.557	0.589	0.506	0.608

Evaluation with Krasula model

In addition to Spearman correlation evaluation, we also analyzed the performance of identifying significant pairs. In this analysis, we have excluded the 4 distortion types

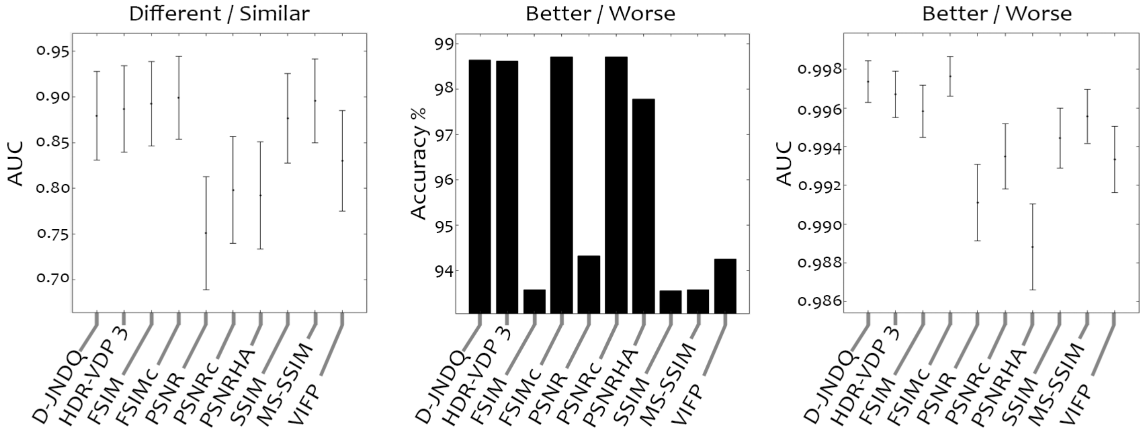


Figure 7.6 – Metric performances on TID-2013 dataset [85] excluding part of the "Exotic" category.

mentioned earlier (out of total 24 types) from the “Exotic“ category. We followed the strategy proposed in [59] to stress out the performances of considered models. In Figure 7.6, the left sub-figure presents the area under curve (AUC) values for each metric at identifying significant and non-significant pairs. Similarly, the right figure shows AUC values for each metric in identifying better or worse image pairs, while the central figure indicates the metric's accuracy in distinguishing better or worse images in significant pairs. Although there is no significant difference in many metric performances, the proposed metric (D-JNDQ) has a competitive performance in identifying significant versus similar pairs. For better/worse analysis, all metrics seem to perform well overall. D-JDNQ, HDR-VDP 3, FSIM, and FSIMc have a significantly better performance than the rest of the evaluated metrics in terms of AUC values. D-JNDQ, HDR-VDP 3, FSIMc, and PSNRc have more than 98% accuracy on identifying whether a stimulus within a significant pair is significantly better or worse than another.

7.4.2 Ablation study

Table 7.3 – SROCC values with and without pre-processing.

	Noise	Actual	Simple	Exotic	New	Color	Full
A.R. Input	0.851	0.881	0.894	0.315	0.842	0.813	0.589
RGB Input	0.742	0.750	0.801	0.141	0.703	0.734	0.446

Table. 7.3 depicts the ablation study results. The best model parameters for each input type were trained for the same amount of iterations. Results show that the model with achromatic response input has a higher correlation with the MOS compared to the one using RGB inputs.

7.5 Conclusion

We propose a learning-based metric, D-JNDQ trained using the first JND point information. The optical and retinal pathway model from HDR-VDP 2 is used as a preprocessing module to improve the performance of the metric. Our experimental results show that the metric is well generalized in quality assessment of various types of distortions in both sub and suprathreshold intensities. It is demonstrated that the first JND points provide rich information for image quality assessment. Additionally, the proposed metric shows poor performance for certain distortion types, where the image quality task is related to distortion preference rather than distortion visibility. Since we utilized a distortion visibility database to develop the metric, this is not a surprising outcome. We also believe that the proposed approach can be extended on video quality evaluation tasks following a similar recipe.

Chapter 8

Objective Quality Assessment of Light Field Content

This chapter is dedicated to the objective quality assessment of light field content. Contributions in this chapter were previously published in two different peer-reviewed papers, at 2019 8th European Workshop on Visual Information Processing (EUVIP) [4] and 2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW) [5]. The paper titled "Investigating Epipolar Plane Image Representations for Objective Quality Evaluation of Light Field Images" [4] received the Best Student Paper Award. Moreover, we contributed to the IEEE P3333.1.4 standardization document for the quality evaluation of light field content (currently in progress) with our findings.

Fig. 8.1 summarizes the organization of the chapter. After introducing the related theoretical concepts, we answer the following questions:

— What are the characteristics of light field specific distortions?

— What makes Epipolar Plane Image representation suitable for objective quality assessment of light field content?

Building on the acquired answers, a no-reference objective light field IQA metric is proposed in the final section.

8.1 Theoretical Introduction

8.1.1 Light Field Representation and Visualization

Michael Faraday introduced the concept of the light field over a century ago in his lecture titled "Thoughts on Ray Vibrations" [28]. In the last century, a 7D plenoptic function was introduced to define the modern Light Field. It is described as below:

$$L_{7d} = P(x, y, z, \theta, \phi, t, \lambda) \tag{8.1}$$

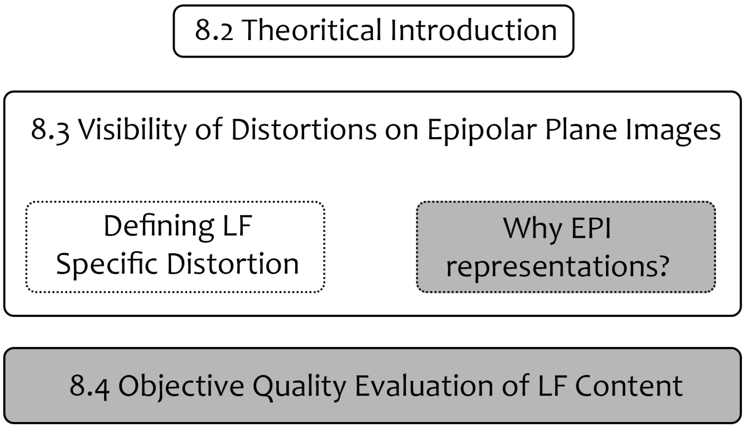

8.2 Theoritical Introduction
8.3 Visibility of Distortions on Epipolar Plane Images
Defining LF Specific Distortion
Why EPI representations?
8.4 Objective Quality Evaluation of LF Content

Figure 8.1 – Organization of the chapter. Gray boxes indicate the contributions corresponding to represented sections.

which represents the light ray from any given point (x, y, z) in 3D space, to any direction (θ, ϕ) in 3D space for any given time t and wavelength λ. Although the 7D plenoptic function has a comprehensive definition, it is not fully utilized in practical applications due to the high dimensionality of the data. A practically more desirable version, 4D plenoptic function, is introduced as a result. It represents each ray with 4 points defined on two parallel planes. The coordinates are denoted with (u, v) for the image plane and (s, t) for the camera plane.

Although the 4D plenoptic function simplifies the light field representation, it is still difficult to imagine in a natural way. Considering that the 4D plenoptic function represents the light field over two parallel planes, we can visualize it from the perspective of each plane. *"Lenslet array"* visualizes the light field from the point of view of the image plane (u, v). This allows us to represent the whole light field as a 2-dimensional image as a collection of rays from different viewpoints from the camera plane (s, t) approaching to image plane (u, v). Fig. 8.2.a provides an example of such visualization. Secondly, we can imagine the camera plane (s, t) collecting the light rays emitted from the image plane (u, v). This results in a 2-dimensional array of images. In this visualization, each image in the 2-dimensional array is called a sub-aperture view. 9 sub-aperture views of a sample light field are visualized in Fig. 8.2.b. In addition to the two perspectives described above, the light field can be visualized as slices over sub-aperture views. By choosing one dimension from the camera plane (s, t) and another dimension from the image plane (u, v), we can represent the slices on (u, s) or (v, t) coordinates. Each slice is called an Epipolar

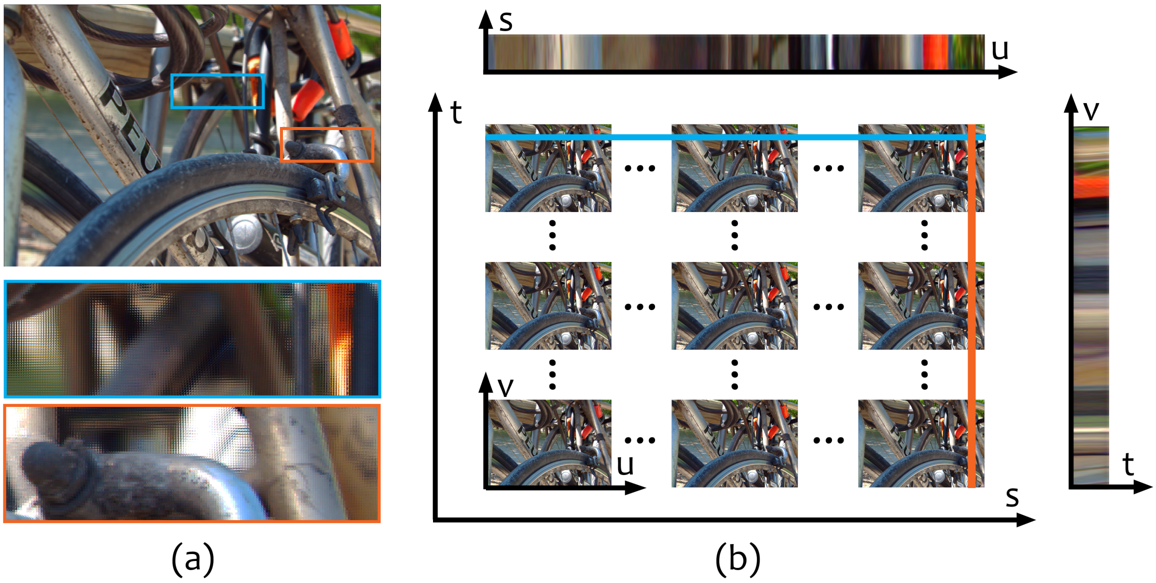


Figure 8.2 – Alternative visualizations of a sample Light Field.

Plane Image (EPI). 2 EPI of a sample light field is visualized on Fig. 8.2.b with their corresponding slices drawn over the sub-aperture views.

8.1.2 Light Field Processing

Acquiring a light field image is not straightforward as traditional photography since the light field captures the light distribution at each location on the sensor. This can be achieved by using multiple sensors simultaneously (i.e., camera array), using a single sensor in a sequential manner (i.e., robotic arm), or specialized light field cameras which rely on lenslet arrays located between the aperture and the camera sensor (i.e., plenoptic cameras). Generally, selecting the desired capturing methodology comes down to a resolution trade-off between spatial, angular, and time dimension. In order to overcome the limitations due to the trade-off between light field dimensions, super-resolution models are proposed. Although spatial and temporal super-resolution is widely researched for traditional 2D images and videos, they have higher importance for light field imaging. Additionally, angular super-resolution (i.e., view synthesis) models are developed specifically for light field content.

The high dimensionality of the light field content also amplifies the importance of compression/transmission of content. While maintaining perceptual quality in spatial and temporal domains like traditional 2D content, light field compression needs to ensure consistency on the angular domain.

Processing tools from different stages of the light field imaging pipeline often alter the perceptual quality of light field content. It makes the quality assessment of light field content necessary to develop models which provide the most benefit while preserving the perceptual quality of the processed light field.

8.2 Visibility of Light Field Specific Distortions on Epipolar Plane Images

In order to correctly assess the impact of distortions on the perceived quality, one needs to understand the nature of distortions. Without knowing what to measure, it is not possible to learn how to measure. This section first explores the light field specific distortions before discussing EPI representation and its benefits in revealing such distortions.

8.2.1 Characteristics of light field specific distortions

Due to the immature stage of light field technology, domain specific distortions are not fully understood and well established. A logical approach would be identifying the type of distortions is considering its potential sources. Therefore, we can consider various light field imaging pipeline stages, such as acquisition, transmission, reconstruction, and display.

Transmission related distortions in light field imaging pipeline generally occur due to lossy compression algorithms. Various light field quality datasets considered state of the art compression algorithms to generate stimuli for quality evaluation.

Reconstruction related distortions occur due to algorithms used to increase the density of sub-aperture views of light field content. While simple algorithms like linear or nearest-neighbor interpolation can be used, sophisticated algorithms based on optical flow or depth maps also exist. Recent light field quality datasets include such distortions for stimuli generation.

Display related distortions are generally challenging to replicate since it requires specific use cases with a variety of sophisticated display options in hand. In the literature, only display related distortion considered in the light field quality dataset is related to multi-

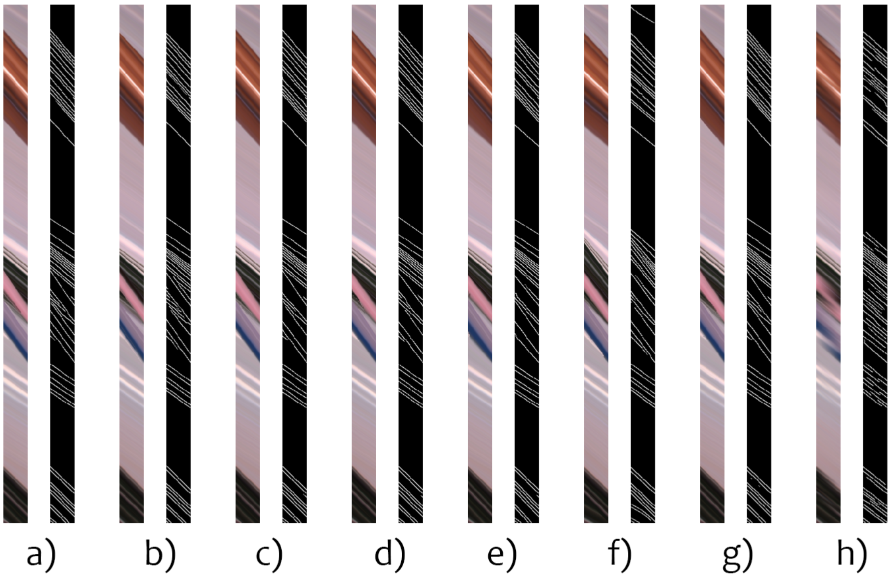


Figure 8.3 – Sample EPI slices and their corresponding edge maps from FVV dataset [97] a)Reference, b)3D-HEVC, c)Multi-View Video, d)HEVC Test Model, e)JPEG200, f)Lossless edge depth coding, g) Color channel corrolation based coding, h)Z-LAR-RP

view auto-stereoscopic displays. To create such distortions, stimuli used in the dataset were altered with varying levels of Gaussian blur in the angular domain.

Additionally, distortions related to the 2D image and video content might as well appear in the light field imaging pipeline. Acquisition related distortions generally fall into this category. On top of this, light field content captured with plenoptic cameras also introduces vignetting around the surrounding sub-aperture views.

8.2.2 Visibility of Distortions on EPI Representations

A surface on a given object is called Lambertian if the light scatters the same from any angle. This indicates a non-reflective and non-refractive property for the object. Such objects appear as straight lines on EPI slices. Non-Lambertian surfaces such as glass, shiny metal surfaces, etc., can overlap other lines in EPI slices. Therefore, even though we expect a set of perfectly straight lines on EPI representations, we might end up with overlapping lines, which may appear as distortions for simple algorithms. Therefore it is crucial to understand the characteristics of light field related distortions and their appearance on EPI representations.

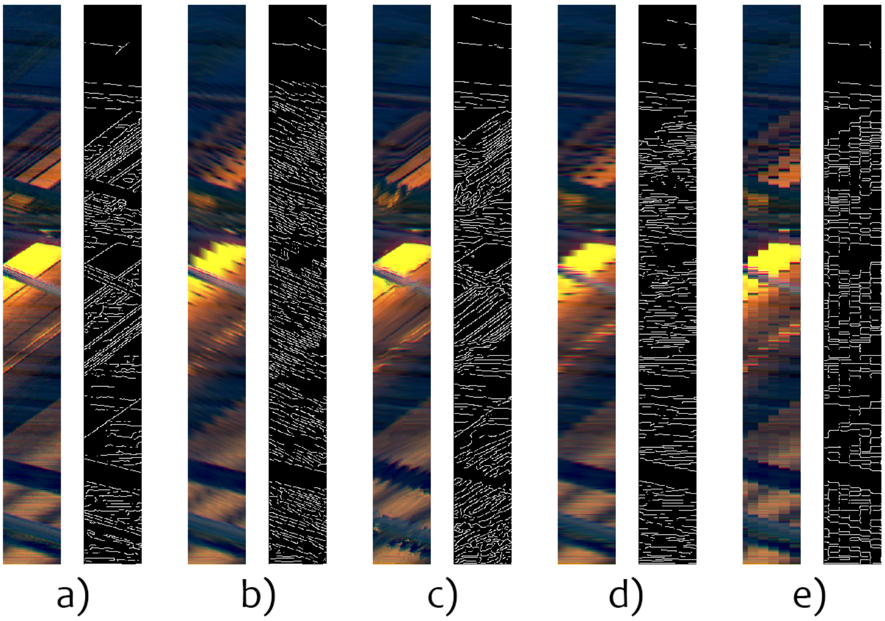


Figure 8.4 – Sample EPI slices and their corresponding edge maps from the MPI-LFA dataset[1]. a)Reference, b)Depth map interpolation, c)Optical flow interpolation, d)Linear interpolation, e)Nearest neighborhood interpolation

Fig. 8.3 presents the sample EPI slices from the FVV dataset [97] with their corresponding edge maps acquired with Canny edge detection algorithm[15]. Fig. 8.3.a is the sample EPI slice from the pristine reference image and its edge map. The rest of the EPI slices corresponds to a particular distorted stimulus available in the dataset. Although some are more visible on the sample EPI slices, we can observe a clear difference for each distortion in comparison to the reference.

Similarly, EPI slices and their corresponding edge maps from the MPI-LFA dataset [1] are presented in Fig. 8.4. Only 4 distortion types out of 7 from the dataset are visualized along with the pristine reference. An immediate observation is that the nature of each interpolation method is different, and they are all quite visible on the presented EPI slices. Most importantly, EPI slices of pristine light field content are well structured and mainly contain straight lines without breaks throughout the angular dimension. Most distortions disturb such straight lines by shifting pixel values arbitrarily. Therefore quality metrics that rely on structural features are promising candidates for light field quality evaluation on EPIs.

Experiment 1 - Edge Detection on EPIs:

Table 8.1 – NICE [92] metric performance with various edge detector algorithms in terms of PCC on the MPI-LFA dataset[1].

	PCC	SROCC
Canny [15]	**0.7145**	**0.6552**
Sobel [104]	0.6457	0.5951
Prewitt [87]	0.6450	0.6038

Table 8.2 – Selected metric performances on EPI representations and sub-aperture views in terms of PCC.

	EPI	View
MW-PSNR [94]	0.7698	0.7921
GMSD [121]	0.7410	0.6715
NICE [92]	0.5122	0.4310

For various distortion types in both datasets, we observe that the simple edge detection algorithms emphasize the visibility of distortions on EPI slices. Three edge detection algorithms have been investigated to select the best performing at revealing the structural differences between the pristine reference and distorted stimuli. Natural Image Counter Evaluation(NICE) metric is used for the experiment [92]. NICE predict the image quality based on the difference between the edge map of the distorted and pristine reference. Edge maps are first dilated with a plus-sign kernel, and then non-zero elements in the XOR maps between the dilated edge maps are used to predict the final quality score. 2 source images were chosen from the MPI-LFA dataset [1] which provides 800 comparisons. Even though it is far from the complete dataset, it provides insight into the edge detector performances.

Table 8.1 presents the PCC and SROCC values between MOS values and predicted quality scores for each edge detector when incorporated with NICE. Among the commonly used edge detectors in the literature, we see that the Canny edge detector provides the highest correlation with the subjective scores.

Experiment 2 - Sub-Aperture Views vs EPIs:

Based on the observations above, three different quality metrics are selected for the experiment on FVV dataset [97]. Predicted quality scores are acquired by averaging over sub-aperture views and EPI representations. Performance of each quality metric is evaluated with PCC values between the predicted quality scores and provided MOS. Results are presented in Table 8.2. Both GMSD [121] and NICE [92] performs significantly better

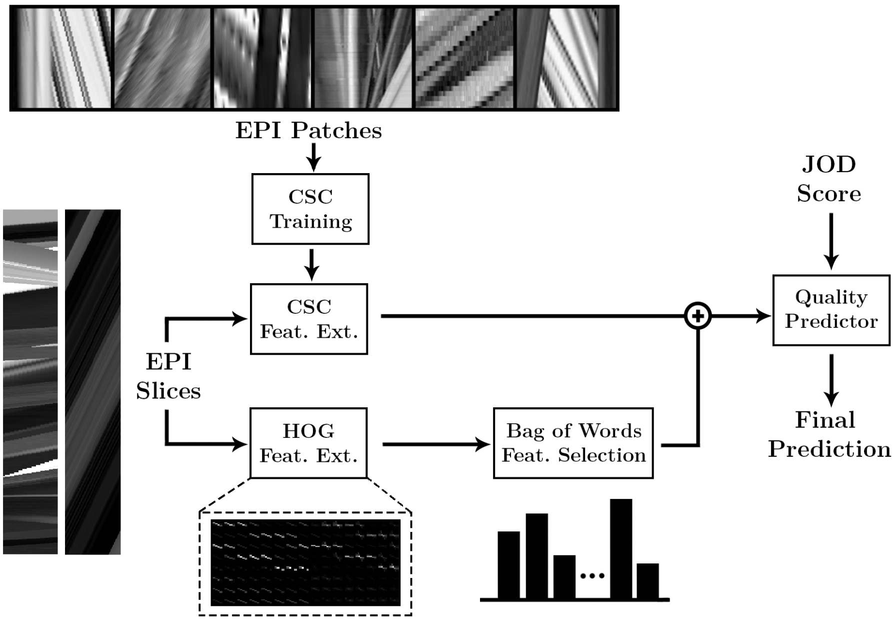


Figure 8.5 – General scheme of the proposed NR light field image quality metric

on EPI representations in comparison to sub-aperture views. The lower performance of MW-PSNR [94] on EPI representations can be explained by the low spatial resolution of the EPI slices since MW-PSNR can only use 2 scales of the morphological wavelet decomposition instead 4 on sub-aperture views.

8.3 No-Reference Objective Quality Assessment Metric for Light Field Images

Based on the previously discussed points above, a no-reference light field image quality metric is proposed. It relies on quantifying the structure-related distortions within EPI representations. On the one hand, Histogram of Oriented Gradient (HOG) descriptors are used with a bag-of-words codebook to represent the overall structural statistics of EPI slices. On the other hand, a Convolutional Sparse Coding (CDC) codebook is trained on a carefully selected set of EPI patches with significant light field related distortions. Fig. 8.5 depicts the general structure of the proposed model.

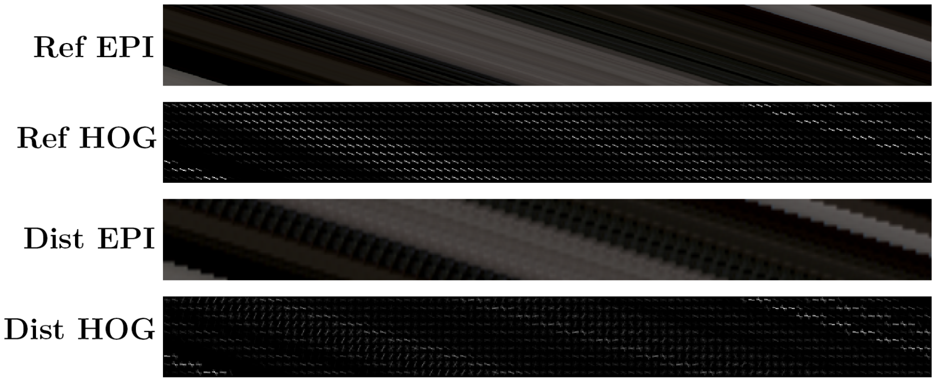


Figure 8.6 – Reference and Distorted EPI slices with corresponding HOG feature maps.

8.3.1 Histogram of Oriented Gradients based Bag of Words Model

2 EPI slices and their corresponding HOG maps are visualized in Fig. 8.6. Hog maps contain feature blocks that indicate the direction of the magnitude with $\pi/16$ resolution. The intensity of the white color in each direction indicates the magnitude of the gradient for the corresponding direction. We can observe a clear difference between the HOG feature blocks by comparing the HOG maps of pristine reference and distorted EPI slices. While HOG feature blocks in the reference HOG map are consistent in direction and magnitude along the diagonal lines, feature blocks in distorted HOG maps vary in direction and magnitude.

Furthermore, EPI patches with size 24 × 24 px are analyzed. Fig. 8.7 visualized the output of the analysis. Each column contains an EPI patch and its corresponding circular histogram. Note that each histogram normalized within and magnitude of the gradients are indicated with relative values on the left side of the circular plot. We can further confirm that gradients in distorted EPI patches vary greatly when compared to reference EPI patches.

Based on the discussion above, HOG is utilized to quantify distorted EPI patches in the proposed model. EPI slices divided into 10 × 10 px blocks and HOG features are calculated block-wise with 16 orientations that cover the $[0, 360]$ degrees.

During training, extracted HOG features are used to train a BoW dictionary to acquire a global HOG representation for a given EPI slice. With the BoW approach, image patches are being clustered into different groups where each group belongs to a particular type of structure.

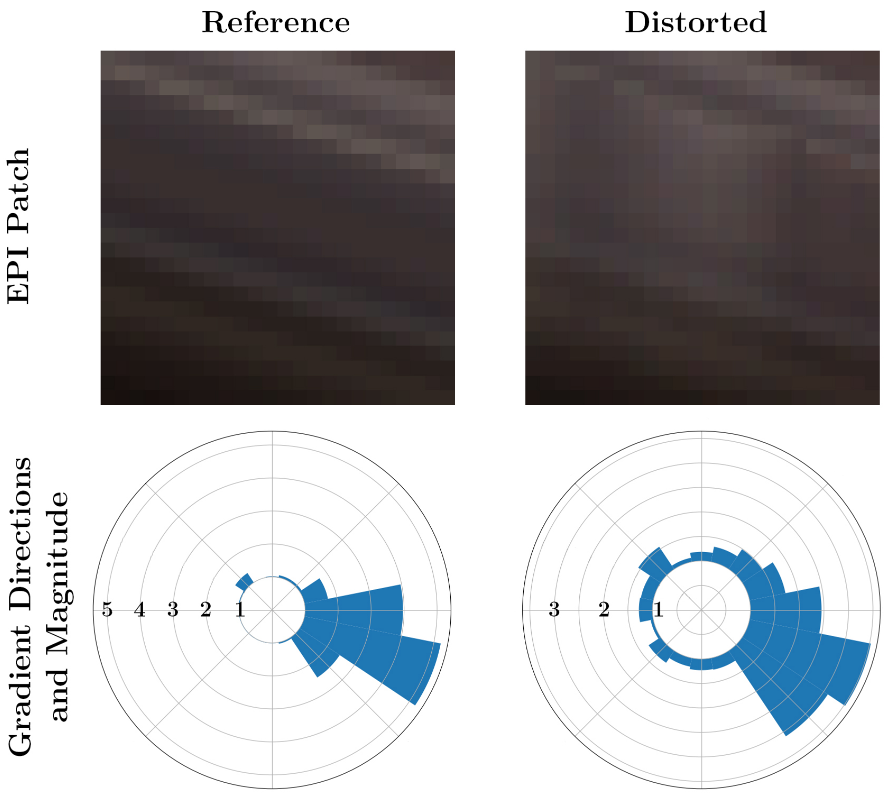


Figure 8.7 – Above, reference and distorted light field EPI patches. Below HOG directions and magnitudes for each patch are plotted on circular histograms.

8.3.2 Convolutional Sparse Coding

Convolutional Sparse Coding (CSC) learns a sparse representation of an image by convolutional filters. It has been shown that image quality assessment in the HVS also adheres to the strategy of sparse coding [2, 80]. Based on this, CSC was utilized in the proposed model for quantifying structure-related distortions.

For training, 1k candidate EPI patches (100×100 px) were collected from the MPI-LFA dataset [1] which covers all six types of distortions of five levels of intensity. Furthermore, 360 patches among the initially selected 1k patches are selected by two experts based on the agreement with visible structural distortions.

The following objective function is used to learn the dictionary on the collected EPI

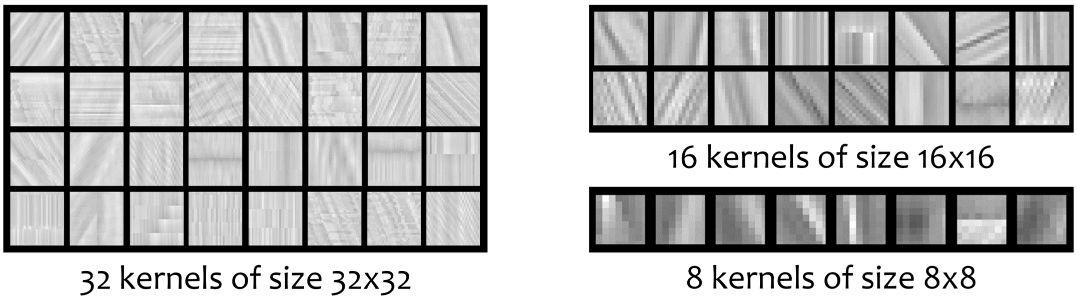


Figure 8.8 – Kernels learned using CSC.

patches:

$$\begin{aligned} &\min_{D,Z} \frac{1}{2} \|y - \sum_{e=1}^{E} D_e Z_e\|_2^2 + \alpha \sum_{e=1}^{E} \|Z_e\|_1 \\ &\text{s.t.} \quad \|D_e\|_2^2 1 \quad \forall e \in \{1, ..., E\} \end{aligned} \tag{8.2}$$

Where y is an input image, D_e is the e_{th} element of the CSC dictionary, Z_e is the feature map with respect to the kernel D_e, α is a parameter that balances the reconstruction loss and the sparsity, E is the number of elements in the dictionary, and indicates the convolution operation.

Collected EPI patches were divided into 30 batches, and one of the out-of-the-shelf was applied to speed up the optimization calculations [114]. After training, 293 kernels were acquired in total. Noisy kernels are discarded based on the energy [114]. In the end, 56 kernels were kept with 3 different sizes, as shown in Figure 8.8. It can be observed that the kernels are well representative of various distortion characteristics.

Using the learned dictionary, for a given $m \times n$ EPI slice, it could then be represented by a $m \times n \times E$ tensor of feature maps $Z_{EPI} = [Z_1; \ldots; Z_e; \ldots; Z_E]$, where each map Z_e is the response of using kernel D_e. With the feature maps, a CSC-based feature descriptor f_{csc} could be then computed using as done in [68, 69]:

$$f_{csc} = (f_{act}(Z_1), \ldots, f_{act}(Z_E)), \tag{8.3}$$

where f_{act} is defined as

$$f_{act}(Z_e) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{1}({}_e(i,j) > \varepsilon)}{m \times n}, \tag{8.4}$$

Table 8.3 – Overall Performances on the MPI-LFA dataset [1]. Median values on the validation set are reported for each measure across 1000 runs.

Metrics	PCC_m	$SROCC_m$	$RMSE_m$
PSNR	0.7830	0.8078	1.2697
SSIM [119]	0.7123	0.7027	1.4327
VIF [101]	0.7861	0.7843	1.2618
FSIM [129]	0.7679	0.7776	1.3075
MS-SSIM [117]	0.7518	0.7675	1.3461
IW-SSIM [118]	0.7966	0.8124	1.2340
BRISQUE [77]	0.7597	0.6724	1.1317
NFERM [39]	0.7451	0.6454	1.1036
MW-PSNR-reduced [95]	0.6757	0.7217	1.5048
MW-PSNR-full [94]	0.6770	0.7232	1.5023
BELIF [103]	0.9096	0.8854	0.7877
Proposed [5]	0.9005	0.8942	0.8916

ϵ is a threshold for selecting activated pixels. Function $f_{act}(\cdot)$ aggregates the number of pixels which are above the threshold ϵ in each sparse feature map Z_e corresponding to each kernel D_e. Intuitively, this function counts the number of pixels that are activated by the corresponding kernel. In other words, since the kernels are trained to capture light field specific artefacts, this process can be interpreted as the computation of certain types of artifacts in the entire image and thus can be used to indicate perceived quality.

8.3.3 Quality Prediction

Learned kernels are then used as feature extractors along with the HOG feature descriptors. Support Vector Regression (SVR) model is trained to predict the final quality scores from extracted features with 1000-fold cross-validation. This procedure helps to eliminate bias toward the training set.

8.3.4 Performance Evaluation

For the MPI-LFA dataset, the authors used the just objectionable differences(JOD) as the scale. Zero(0) JOD score means having no quality difference, while negative values indicate an observable quality difference. There are robust evaluation methods for image quality metrics, such as the model proposed by Krasula et al[**lucas**]. Global correlation

Table 8.4 – Performances of the ablative models using only HOG or CSC features in comparison to the full model.

	PCC	SROCC	RMSE
HOG	0.7845	0.7782	1.2690
CSC	0.8143	0.8088	1.1740
Both	0.9005	0.8942	0.8916

measures such as PCC, SCC do not consider the uncertainty in the subjective scores, and objective metrics need to be mapped to the subjective quality experiment range. The proposed approach by Krasula et al. resolves these problems. However, in order to use the proposed evaluation methodology, we need to have access to statistical information about the subjective test. if MOS scores, and not individual scores, are the only reported results of subjective experiments, it is not enough to run such full comprehensive evaluation. Unfortunately, MPI-LFA dataset does not provide enough statistical information about the results rather than the JOD scores, we utilized the cross-validation methodology for the evaluation of the proposed model

The model's performance was evaluated with global correlation measures such as PCC, SROCC, and RMSE with 1000-fold cross-validation. The dataset was randomly split into two disjoint sets at each fold as 80% for training and 20% for validation, and PCC, SROCC, and RMSE are calculated between the predicted quality scores and the subjective opinions.

Median values of PCC (PCC_m), SROCC ($SROCC_m$) and RMSE ($RMSE_m$) across 1000 folds are reported in Table 8.3. Performance of the several 2D FR image quality metrics, 2D NR image quality metrics, multi-view, and light field metrics were compared to the proposed metric. As observed, the proposed metric outperforms the traditional 2D and 3D metrics and achieves competitive performance compared to the state-of-the-art light field quality metric [103].

Ablation Study:

is necessary to ensure the contribution of each set of features (i.e., CSC and HOG) in the model. Results are shown in Table 8.4 in terms of PCC, SROCC and RMSE. The model trained with only CSC features outperforms the model with HOG features alone, but the difference in performance is not significant. When both feature extractors are used, a significant improvement in the model performance is observed.

8.4 Conclusion

In this chapter, we investigated the characteristics of light field related distortions. We further demonstrated the visibility of such distortions on the EPI representations. Simple structural metrics were shown to perform better on EPI representations compared to sub-aperture views. Based on our findings, we developed a no-reference image quality metric that relies on two different structural measures to quantify the distortions on the EPIs. The proposed metric has competitive performance on the popular MPI-LFA dataset.

Chapter 9

Quality Evaluation of Volumetric Video Content

This chapter is dedicated to the quality evaluation of volumetric video (VV) content. We investigate the influence of the temporal sub-sampling rate and sampling methodologies on the performance of objective VV quality metrics. Our findings presented in this chapter were previously published [8] at Picture Coding Symposium 2021 (PCS 2021) as part of our collaboration with Emin Zerman and Aljosa Smolic from Trinity College Dublin.

After a brief introduction of VV content and its applications in Section 9.1, we describe the vsenseVVDB2 volumetric video quality dataset [125] in Section 9.2. Influence of temporal sub-sampling rates on the objective quality evaluation of VVs are investigated in Section 9.3.4 whereas the impact of temporal pooling methodologies is investigated in Section 9.3.5. Combined effect off temporal sub-sampling rate and sampling methodologies is investigated in Section 9.3.6 before concluding the chapter in Section 9.4.

9.1 Volumetric Video

Volumetric video is a relatively new form of the immersive media type. The difference between static 3D models and VV content is rather trivial. VV contents are dynamic, i.e., they include the temporal changes of the 3D model within the captured time frame. VVs are usually represented as meshes or point clouds. Meshes contain vertices, edges that connect vertices, and texture maps for color and transparency information. On the other hand, point clouds contain vertices but do not contain any connectivity information between vertices. Color related information is stored individually for each point. The absence of connectivity relation between points in point cloud representations allows more straightforward storage and compression scheme. An example of structural differences between meshes and point clouds is given in Figure 9.1. Although there are structural

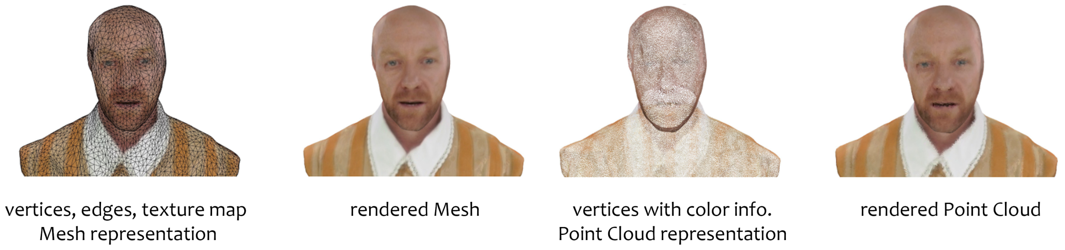

vertices, edges, texture map
Mesh representation
rendered Mesh
vertices with color info.
Point Cloud representation
rendered Point Cloud

Figure 9.1 – Mesh and Point Cloud representation examples. Images are taken from vsenseVVDB2 dataset[125]

differences between the two representations, renderings can be quite similar.

Volumetric video is often captured with a set of cameras surrounding the content being captured. Captured video sequences were used to reconstruct the 3D model. The acquired dynamic model is ideal for integrating into augmented reality (AR) or virtual reality (VR) applications. Possible usage covers a wide range of applications from entertainment to education. With the advance of capturing technologies and increased interest in research and industry, VV attracts undeniable attention from researchers. Point cloud compression became a hot topic for researchers in communities such as MPEG [99]. Subjective quality evaluation of VV also gained attention in the community[125].

9.2 Volumetric Video Quality Dataset: vsenseVVDB2

In this section, we introduce the V-SENSE volumetric video quality dataset (vsenseVVDB2)[125] which we rely on for the remaining of the chapter. Therefore, the explanations regarding the dataset are as detailed as possible to ensure the clarity of the contributions in the rest of the chapter. Interested readers are recommended to refer to the original paper[125] for more details.

vsenseVVDB2 dataset contains eight different volumetric videos in total. Four of these VVs are acquired by the authors and referred to as V-SENSE in the Table **??**. The remaining four VVs were taken from the publicly available 8i voxelized point cloud volumetric video dataset (8iVFB v2)[20]. VVs from V-SENSE include both mesh and point cloud representations, whereas VVs from 8i only contain point clouds. Both datasets contain only full-body human contents. Initial frames of each VV can be seen in Figure 9.2. The number of vertices for meshes and the number of points for point clouds are displayed

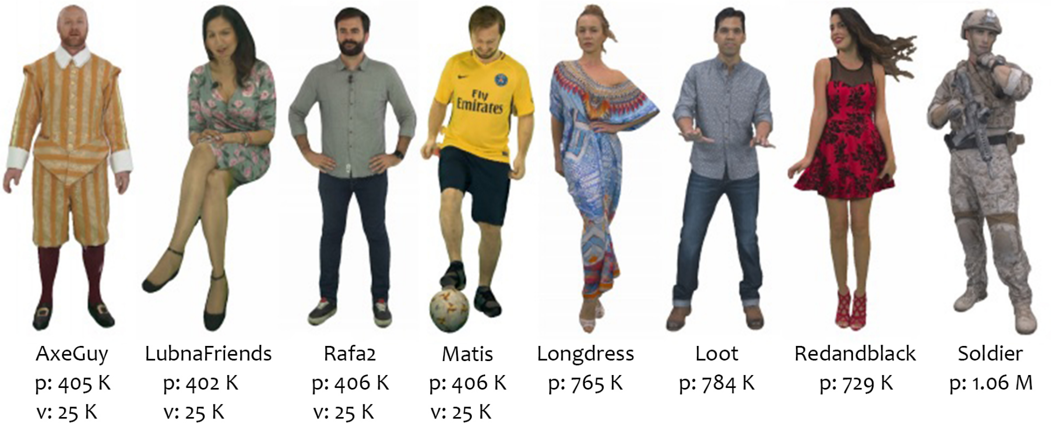


Figure 9.2 – Visualization of the initial frames from vsenseVVDB2 dataset. Number of vertices (v) for meshes and number of points (p) for point clouds are given below each stimuli.

below the name of each stimulus. All VVs contains 300 frames and are 10 second long (30 fps).

Perceived quality of the compression artefacts are evaluated on the collected VVs. Google Draco [32] encoder is used for meshes. Mesh textures are compressed with JPEG. From MPEG standardisation efforts [99], G-PCC and V-PCC is used for point cloud compression. G-PCC encoder is used with region-adaptive hierarchical transform whereas V-PCC encoder is used with all-intra and random-access modes. Details regarding to parameters of each encoder are given in Table 9.2. *QP* and *QT* are the quantisation parameters for Draco encoder and *JPEG* indicates the QP level of JPEG compression. *depth*, *level* and *colSt*(colorstepsize) are parameters for G-PCC encoder whereas *geoQP* and *texQP* are the V-PCC quantisation parameters for geometry and texture respectively.

Experiment conditions were set according to ITU recommendations[51]. 23 participants were recruited for the study. A unique VV (different than the eight stimuli included in the dataset) is used with V-PCC compression artefacts for the training session. ACR methodology (see Chapter 1.1.2 for more details) was adopted as experiment procedure. Stimuli were presented to observers as 10 seconds video renderings on a 24 " LCD display. Collected ratings, then converted to MOS/DMOS scores as needed.

Table 9.1 – Summary of encoder parameters utilized in vsenseVVDB2 dataset [125].

Compression		**Quality Levels**					
Methods		**R1**	**R2**	**R3**	**R4**	**R5**	**R6**
Draco +JPEG	**QP**	8	10	10	12	12	12
	QT	6	10	10	10	12	12
	JPEG	0	0	5	10	30	55
G-PCC	**depth**	10	10	10	10	10	10
	level	6	7	7	7	8	10
	colSt	64	32	16	8	4	1
V-PCC	**GeoOP**	32	28	24	29	16	-
	texQP	42	37	32	27	22	-

9.3 Influence of Temporal Sampling on Objective Quality Evaluation

One of the challenges to the quality evaluation of VV content is the high dimensionality of the data. Although there are not many metrics designed particularly for VV content, point, and point&color based metrics are often used for objective quality evaluation of VV content. Additionally, VV content can be rendered as videos, and traditional image-based quality metrics can be utilized on rendered views. However, the high dimensionality of the VV content makes these approaches time-consuming to utilize for computationally demanding tasks (e.g., optimization of compression algorithms). Therefore in this work, we seek to answer the following question:

"Can we speed up metric computation for VV quality assessment without sacrificing the accuracy?"

In particular, we explore the possibility of reducing the temporal dimensionality of the VV content. To do so, we uniformly sub-sample (reducing the frame rate) of the VV content and investigate the effect on objective quality metric performances. An example of a uniform sub-sampling is visualized on Figure 9.3.

Moreover, we also analyze the effect of temporal pooling methods on the accuracy of objective quality metrics. Various temporal pooling methods have been proposed in the video quality domain to increase the prediction accuracy and speed up calculations[113, 100]. Benchmark study of pooling methods on blind video quality evaluation concluded that pooling on video quality evaluation is content dependent, and an ensemble approach

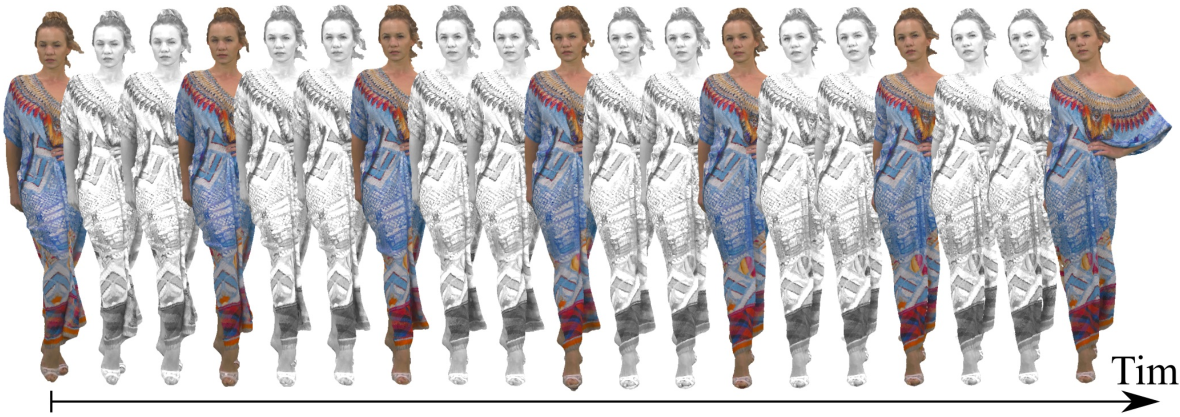


Figure 9.3 – Visualization of an example uniform temporal sub-sampling scheme, where the sampled frames are shown in color while others are shown in black& white

could improve the results[113]. Another study suggests that arithmetic pooling (i.e., taking the mean of quality predictions of individual frames) works better than sophisticated pooling methods for sequences of length in the order of minutes[100]. To the best of our knowledge, these questions are unanswered for VV content.

Finally, we explore the combined effect of temporal sub-sampling and pooling methods on the objective quality metrics. The results of these experiments provide insight for speeding up the objective quality evaluation of VV content.

9.3.1 Temporal pooling methods

Temporal pooling methods considered in our experimentation are summarized in Table 9.2. Given formulas use common notations. q_i is the estimated quality score of i_{th} frame in the video. i ranges from 1 to N, where N indicates the last frame of the video. Finally, the quality score of the video is denoted as Q.

Arithmetic mean is calculated as the mean value of quality scores across frames within the VV. Harmonic mean uses a similar definition with a negative exponent to have a higher impact on frames with lower quality. Minkowski mean is a generalized version of the arithmetic and harmonic mean with an adjustable parameter. When $p = 1$ and $p = -1$, Minkowski mean provides the same results with arithmetic and harmonic means, respectively.

VQ pooling is proposed as an adaptive spatio-temporal pooling strategy[81]. We only use the temporal pooling part as suggested in [113]. Concretely, quality scores of all frames

Table 9.2 – Definitions and selected parameters for pooling methods.

Pooling method	Formula	Parameter
Arithmetic mean	$Q = \frac{1}{N}\sum_{i=1}^{N} q_i$	-
Harmonic mean	$Q = \left(\frac{1}{N}\sum_{i=1}^{N} q_i^{-1}\right)^{-1}$	-
Minkowski mean	$Q = \left(\frac{1}{N}\sum_{i=1}^{N} q_i^{p}\right)^{1/p}$	$p = 2$
VQ pooling [81]	$Q = \frac{\sum_{i\in G_L} q_i + w \cdot \sum_{i\in G_H} q_i}{\lvert G_L\rvert + w \cdot \lvert G_H\rvert},\ w = \left(1 - \frac{M_L}{M_H}\right)^2$	-
Percentile pooling [113]	$Q = \frac{1}{\lvert P_{low}\rvert}\sum_{i\in P_{low}} q_i$	Percentile = 10%
Primacy pooling [79]	$Q = \sum_{i=1}^{N} w_i q_i,\ w_i = \frac{\exp(-\alpha i)}{\sum_{j=1}^{L}\exp(-\alpha j)}, 0 \leq i \leq L$	$L = 360, \alpha = 0.01$
Recency pooling [79]	$Q = \sum_{i=1}^{N} w_i q_i,\ w_i = \frac{\exp(-\alpha(L-i))}{\sum_{j=1}^{L}\exp(-\alpha(L-j))}, 0 \leq i \leq L$	$L = 360, \alpha = 0.01$

are clustered with the K-means clustering algorithm into two groups: lower quality (G_L) and higher quality(G_H) frames. Afterwards, final quality score is calculated with the formula given in Table 9.2 where $|G_L|$ and $|G_H|$ is the cardinality of respective clusters. M_L and M_H is the mean value of clustered scores.

Percentile pooling is proposed based on the phenomenon that the observer opinions are affected more by the worse frames of the video content[113]. Table 9.3 expresses the formula used for percentile pooling where P_{low} indicates the frames that are in the lower 10% percentile.

Primacy pooling [79] takes advantage of the tendency of observers, where the beginning of the video has a higher impact on the final quality. Conversely, recency pooling [79] captures the opposite behavior, where observers tend to remember the last part of the video while evaluating the video quality. An adjustable parameter α can be used to increase the intensity of these phenomenons.

9.3.2 Temporal sub-sampling rates

Each VV sequence in vsenseVVDB2 is of 10 seconds in length with 300 frames. In other words frame rate of the videos is 30 (fps). In order to sub-sample the videos, for a frame-rate k, we took the first frame among $30/k$ frames and skipped the rest. To ensure a uniform sub-sample with first and last frame of each VV content in evaluation, we choose the divisor $k \in K = \{1, 2, 3, 5, 6, 10, 15, 30\}$.

9.3.3 Selected objective quality metrics

In total, 30 quality metrics are used in this work, where 11 are image-based, and the remaining 19 are point-based metrics. We can further divide the point-based metrics into two categories as point and point&color based metrics.

Image-based metrics

As widely used metrics, peak signal to noise ratio (PSNR) and structural similarity index (SSIM) [119] are included in the experiments. They are both FR metrics. While PSNR is based on pixel value differences between the two compared images, SSIM compares the two images in terms of luminance, contrast, and structure.

MP-PSNR[96] is another FR metric included in our experimentation. It is based on multi-scale pyramid decomposition. Mean square error (MSE) quantifies the intensity of the distortions between the reference and distorted images. Similarly, MW-PSNR [94] was proposed based on morphological wavelet decomposition. MSE between multi-scale wavelet bands of reference and the distorted image is used to calculate the final image quality. RR versions of both metrics later introduced in [95] which utilizes only detailed features of the reference image from higher scales of the decomposition pyramids.

EM-IQM [70] was proposed to evaluate the depth-based image rendering (DIBR) related distortions. It measures the structural deformations between the reference and distorted images based on an elastic metric working on the curves. Analogously, SI-IQM[71] was proposed to evaluate the structural distortions between reference and distorted images from a higher semantic level.

NIQSV [109] estimates the image quality by quantifying the non-smooth regions via morphological operations. It was later extended to NIQSV+[110] by incorporating an indicator for dis-occluded areas. Finally, a learning-based NR image quality metric APT [37] is also included in the experiment.

Point-based metrics

Point-based metrics considered in this work are based on three main approaches from the literature: point-to-point [76], point-to-plane [108] and plane-to-plane [9]. The plane is defined by the normal vector of the point. In the cases where point normals are not

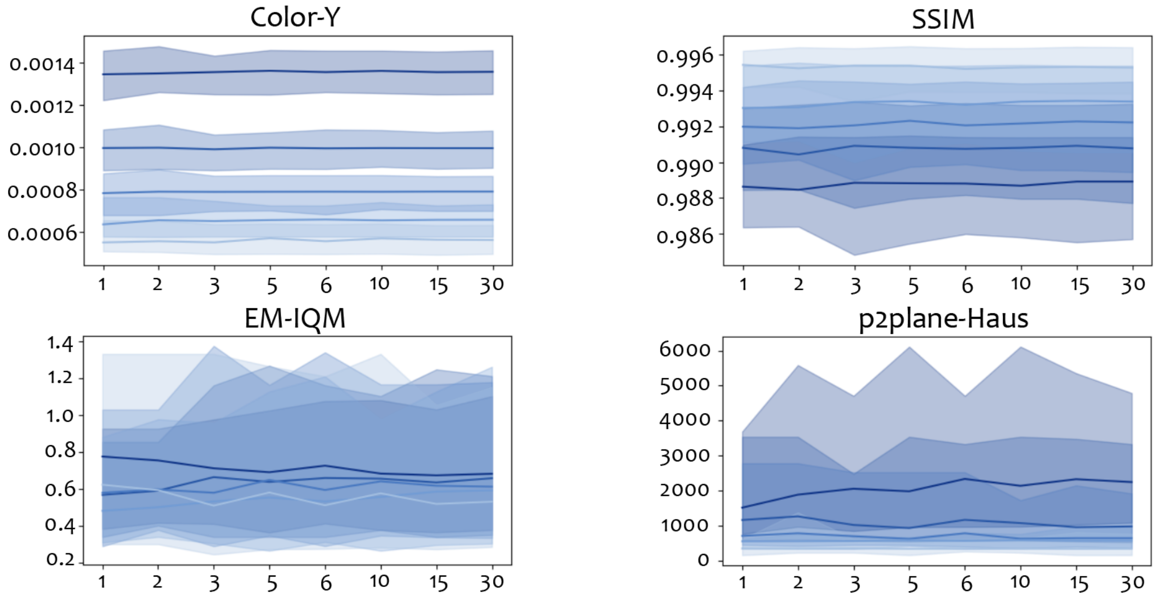


Figure 9.4 – 95% percentile range and the median values of the selected metric scores for 5 levels of V-PCC coding at different frame rates for AxeGuy stimuli

known, Matlab's "pcnormals" function was used. Geometry metrics are computed using either root mean square (RMS), mean square error (MSE), or Hausdorff distances to quantify the differences between reference and compressed point clouds. Pooling of the quantitative differences between points is done with either minimum, mean or median functions.

In addition, differences in color information can be utilized to quantify visual quality. For this, MSE or PSNR is calculated from the differences between corresponding points' assigned color values. These color metrics may operate on Y, U, and V color channels.

9.3.4 Impact of temporal sub-sampling rate

A straightforward way of speeding up the computation of VV quality assessment is by reducing temporal sampling frequency. In order to understand the effect of reducing temporal sampling frequency on the accuracy of objective quality metrics, we uniformly sub-sampled the volumetric videos from vsenseVVDB2 [125] dataset (see Section 9.2) with 8 different temporal frequencies (see Section 9.3.2).

We initially examined the effect of the sub-sampling rate on the estimated quality scores. Figure 9.4 visualizes the result of this analysis. 4 metrics were selected for this

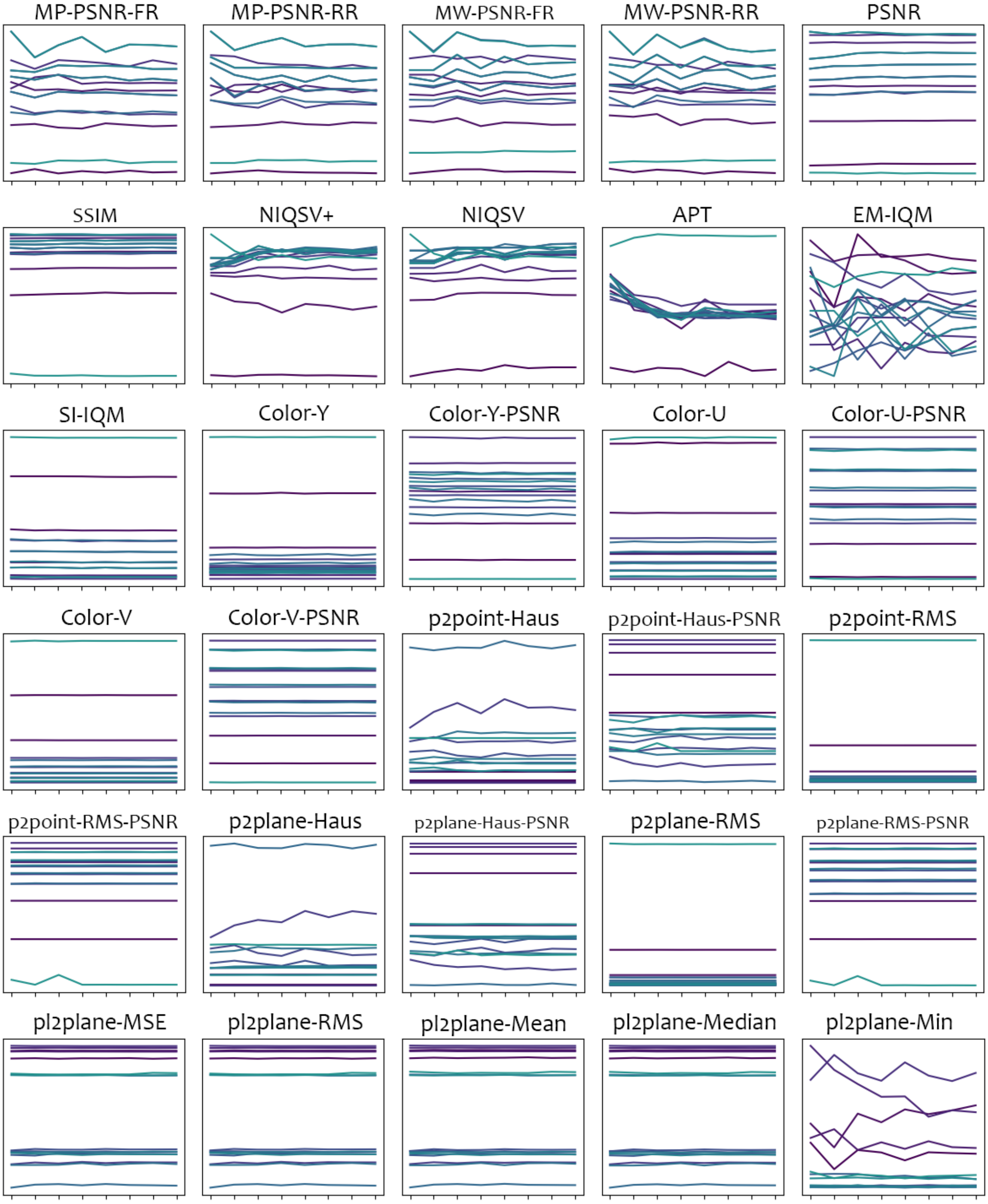


Figure 9.5 – Each line represents the median metric score that changes over 8 temporal sampling frequencies for a compression type/level over the “AxeGuy" source content. X axis is the fps value for each temporal sampling frequency. Y axis is the metric scores normalized for each metric individually.

analysis, 2 higher-performing metrics: Color-Y and SSIM, and 2 lower-performing metrics: EM-IQM and p2plane-Haus. The horizontal axis in each plot is the temporal sampling frequency (fps), and the vertical axis is the predicted quality score for each metric. Each line corresponds to the stimuli “AxeGuy” compressed with V-PCC coding [99] at a certain level. 95% percentile ranges for each stimulus are also indicated around the corresponding lines. We observe that for higher-performing metrics (Color-Y, SSIM), the range of the metric score does not change along with the temporal sampling frequency. On the other hand, lower-performing metrics (EM-IQM, p2plane-Haus) fluctuate with the varied temporal sampling frequency.

Table 9.3 – SROCC values between metric scores and DMOS for different temporal sampling rates with arithmetic mean.

	1-fps	2-fps	3-fps	5-fps	6-fps	10-fps	15-fps	30-fps
MP-PSNR-FR	0.5767	0.4569	0.5127	0.4611	0.4368	0.4051	0.4308	0.4740
MP-PSNR-RR	0.6552	0.6392	0.6337	0.6285	0.6289	0.6323	0.6290	0.6275
MW-PSNR-FR	0.6055	0.6125	0.6132	0.6073	0.6128	0.6130	0.6095	0.6084
MW-PSNR-RR	0.6325	0.6319	0.6392	0.6303	0.6412	0.6341	0.6304	0.6317
PSNR	0.7404	0.7404	0.7422	0.7379	0.7384	0.7407	0.7405	0.7415
SSIM	**0.8544**	**0.8538**	**0.8533**	**0.8509**	**0.8535**	**0.8529**	**0.8518**	**0.8531**
NIQSV	0.0610	0.2051	0.1222	0.1080	0.2360	0.2451	0.1130	0.0728
NIQSV+	0.1934	0.1975	0.2038	0.2167	0.2066	0.2028	0.2129	0.2021
APT	0.1156	0.1253	0.1366	0.1336	0.1471	0.0500	0.0634	0.1000
EM-IQM	0.3259	0.3678	0.4231	0.4366	0.4793	0.4462	0.4519	0.4436
SI-IQM	0.8459	0.8386	0.8374	0.8385	0.8380	0.8378	0.8394	0.8374
Color-Y	**0.7900**	**0.7891**	**0.7815**	**0.7803**	**0.7809**	0.7799	**0.7810**	**0.7818**
Color-Y-PSNR	0.7884	0.7882	0.7799	0.7790	0.7798	**0.7799**	0.7801	0.7811
Color-U	0.4804	0.4780	0.4799	0.4798	0.4761	0.4772	0.4799	0.4772
Color-U-PSNR	0.4834	0.4800	0.4829	0.4842	0.4802	0.4808	0.4837	0.4807
Color-V	0.4718	0.4664	0.4714	0.4713	0.4661	0.4658	0.4713	0.4658
Color-V-PSNR	0.4717	0.4654	0.4711	0.4710	0.4600	0.4643	0.4712	0.4511
p2point-Haus	0.2793	0.2771	0.2733	0.3316	0.2814	0.2554	0.2935	0.2793
p2point-Haus-PSNR	0.5641	0.5819	0.6000	0.6027	0.6056	0.6045	0.6066	0.6099
p2point-RMS	**0.8673**	**0.8692**	**0.8694**	**0.8683**	**0.8692**	**0.8671**	**0.8655**	**0.8665**
p2point-RMS-PSNR	0.7731	0.7991	0.7798	0.7834	0.7814	0.7826	0.7800	0.7704
p2plane-Haus	0.0593	0.1152	0.1220	0.0759	0.1292	0.0781	0.1856	0.1682
p2plane-Haus-PSNR	0.1406	0.1664	0.1403	0.1382	0.1296	0.1460	0.1307	0.1453
p2plane-RMS	0.3677	0.3648	0.3655	0.3666	0.3653	0.3665	0.3620	0.3647
p2plane-RMS-PSNR	0.8426	0.8433	0.8367	0.8363	0.4515	0.8392	0.8376	0.8363
pl2plane-MSE	0.3928	0.3723	0.3717	0.3722	0.3730	0.3651	0.3718	0.3655
pl2plane-RMS	0.3928	0.3739	0.3718	0.3726	0.3730	0.3725	0.3715	0.3726
pl2plane-Mean	0.3340	0.3315	0.3295	0.3297	0.3302	0.3294	0.3290	0.3339
pl2plane-Median	0.3340	0.3315	0.3295	0.3297	0.3302	0.3294	0.3290	0.3339
pl2plane-Min	0.1494	-0.0333	0.1010	0.1089	0.1025	0.0354	0.0177	-0.0009

Similarly, Figure 9.5 presents all the median metric scores for 16 different compressed versions of “AxeGuy” content. It can be observed that the majority of the metric scores are not affected by temporal sampling frequency. With these observations, we evaluated the metric performances expecting a non-significant difference in higher performing metric performances for various temporal sampling frequency. Table 9.3 presents the PCC values for each metric under different temporal sampling frequencies. It could be observed that metrics with higher performance (with SROCC values higher than 0.5) have insignificant performance differences with varied temporal sampling frequencies. Although the conclusion is the same as in terms of SROCC, we additionally provide the PCC and RMSE values in the supplementary material [1] of our publication [8] or in Annexes.

9.3.5 Impact of temporal pooling method

Similarly, to analyze the impact of temporal pooling methods on objective quality metrics, we first plot the estimated metric score differences at 30 fps versus the DMOS scores in Figure 9.6. The horizontal axis in each plot is the metric score, while the vertical axis is the DMOS for each stimulus. Again 4 metrics were selected for this analysis as 2 higher-performing (SSIM, Color-Y) and 2 lower-performing (EM-IQM, p2plane-Haus). Each row of scatter plots corresponds to a certain metric indicated on the left, while each column of scatter plots corresponds to a certain pooling method indicated above. As shown in the 1^{st} and 2^{nd} rows (*i.e.,* higher-performing metrics: SSIM and Color-Y), the distributions of the data points do not differ from each other significantly. On the contrary, 3^{rd} and 4^{th} rows (*i.e.,* lower-performing metrics: EM-IQM and p2plane-Haus) shows a high variance across different temporal pooling methods.

Considering the minimal change in metric scores with varying temporal pooling methods, we do not expect a significant difference in metric performances. Table 9.4 presents the performances of each metric with different pooling methods in terms of SROCC. In the table, it can be clearly observed that the change of temporal pooling method does not have a significant impact on higher-performing metrics (metrics with SROCC values higher than 0.5). Although the conclusion is the same as in terms of SROCC, we additionally provide the PCC and RMSE values in the supplementary material [2] of our publication [8] or in Annexes.

[1] "https://hal.archives-ouvertes.fr/hal-03206240"

[2] "https://hal.archives-ouvertes.fr/hal-03206240"

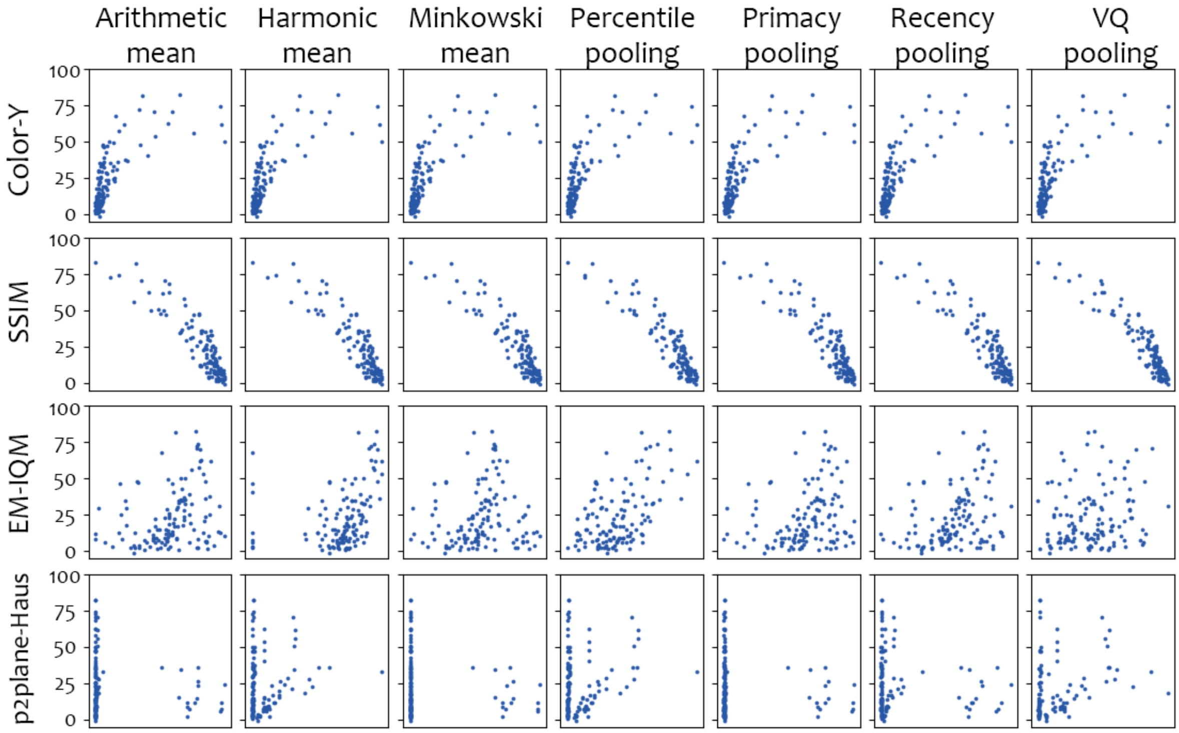


Figure 9.6 – Scatter plots of objective scores predicted by selected quality metrics versus the DMOS scores. Each column corresponds to a certain pooling method indicated above.

Table 9.4 – SROCC values between metric scores and DMOS for different pooling methods with 30 fps.

	Arithmetic mean	Harmonic mean	Minkowski mean	Percentile pooling	Primacy pooling	Recency pooling	VQ Pooling
MP-PSNR-FR	0.4740	0.6489	0.4502	0.4962	0.4876	0.4504	0.7392
MP-PSNR-RR	0.6275	0.6194	0.6336	0.4967	0.6394	0.6337	0.6978
MW-PSNR-FR	0.6084	0.6045	0.6116	0.5079	0.6182	0.6153	0.6640
MW-PSNR-RR	0.6317	0.6236	0.6347	0.5098	0.6440	0.6322	0.6701
PSNR	0.7415	0.7404	0.7416	0.7253	0.7376	0.7476	0.7370
SSIM	**0.8531**	**0.8532**	**0.8531**	**0.8649**	**0.8511**	**0.8522**	**0.8955**
NIQSV	0.0728	0.1011	0.2371	0.1577	0.2517	0.1064	0.1018
NIQSV+	0.2021	0.2021	0.2075	0.2136	0.2387	0.2006	0.2720
APT	0.1000	0.0568	0.1220	0.0950	0.1567	0.1445	0.0114
EM-IQM	0.4436	0.5293	0.4731	0.4233	0.4272	0.3956	0.2045
SI-IQM	0.8374	0.8384	0.8372	0.8377	0.8422	0.8343	0.8291
Color-Y	**0.7818**	0.7789	0.7791	**0.7807**	0.7821	**0.7800**	0.7607
Color-Y-PSNR	0.7811	**0.7817**	**0.7807**	0.7654	**0.7842**	0.7786	**0.7649**
Color-U	0.4772	0.4806	0.4760	0.4910	0.4763	0.4805	0.5007
Color-U-PSNR	0.4807	0.4792	0.4808	0.4588	0.4781	0.4799	0.4855
Color-V	0.4658	0.4684	0.4649	0.4727	0.4690	0.4658	0.4813
Color-V-PSNR	0.4511	0.4672	0.4678	0.4512	0.4702	0.4663	0.4711
p2point-Haus	0.2793	0.3461	-0.1189	0.3091	0.2566	0.2012	-
p2point-Haus-PSNR	0.6099	0.5377	0.3821	0.5123	0.5890	0.6158	-
p2point-RMS	**0.8665**	**0.8635**	**0.8575**	**0.8579**	**0.8633**	**0.8651**	0.8615
p2point-RMS-PSNR	0.7704	0.7903	0.7606	0.8397	0.7773	0.7658	0.6765
p2plane-Haus	0.1682	0.0091	-0.0307	0.2386	0.1820	0.1148	-
p2plane-Haus-PSNR	0.1453	0.1360	0.1735	0.0992	0.1527	0.1358	0.1797
p2plane-RMS	0.3647	0.3632	0.3579	0.3655	0.3603	0.3639	0.3695
p2plane-RMS-PSNR	0.8363	0.8321	0.5636	0.7949	0.8428	0.8285	**0.8644**
pl2plane-MSE	0.3655	0.3726	0.3728	0.2799	0.3613	0.3752	0.3388
pl2plane-RMS	0.3726	0.3726	0.3726	0.3294	0.3670	0.3755	0.3116
pl2plane-Mean	0.3339	0.3286	0.3285	0.3082	0.3392	0.3283	0.3080
pl2plane-Median	0.3339	0.3286	0.3285	0.3082	0.3392	0.3283	0.3080
pl2plane-Min	-0.0009	0.1073	0.0169	0.1460	0.0420	0.0447	0.0245

9.3.6 Combined effect of temporal sub-sampling and pooling methods

This section analyzes the combined effect of temporal sub-sampling rate and pooling methods on VV objective quality assessment. Figure 9.7 presents the performance of objective quality metrics in terms of SROCC. Each row corresponds to a different metric, and the order of the metrics is the same as the order in rows of Table 9.3 and Table 9.4. Each column shows a different combination of temporal sub-sampling rate and temporal pooling method. Columns are divided into groups of 8, with increasing fps from left to right as indicated at the top. Each group of columns corresponds to a certain temporal

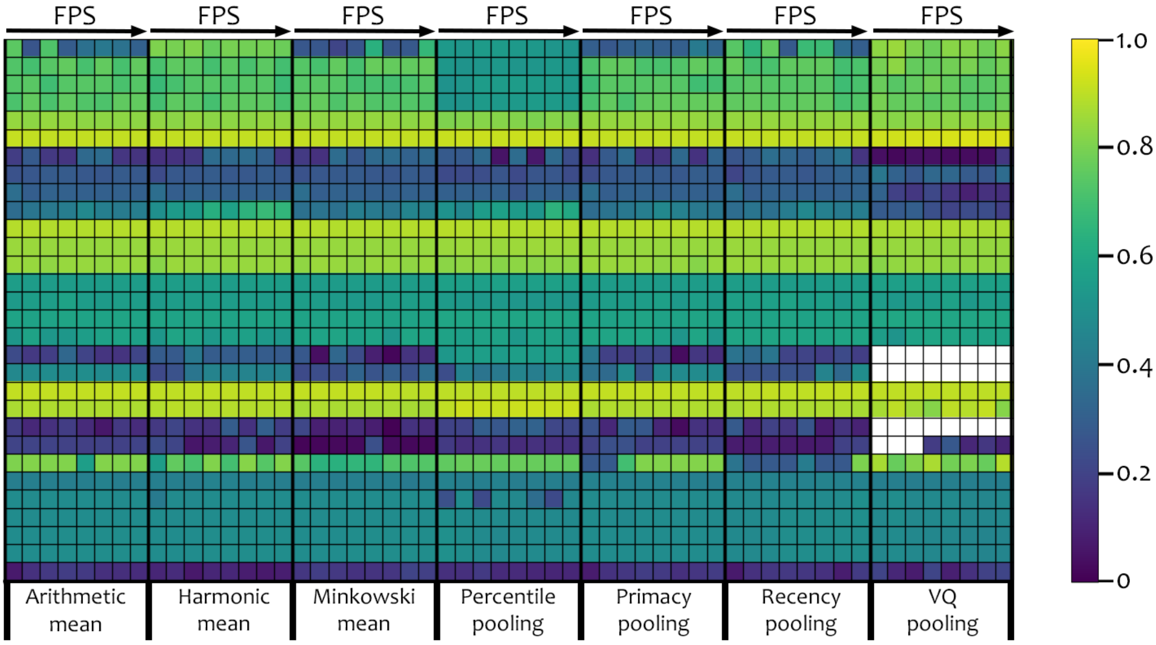


Figure 9.7 – The effect of both temporal sub-sampling rate and temporal pooling on the performance of the metric in terms of SROCC.

pooling method indicated below. The color of each cell depends on the SROCC values, as shown on the right.

It can be observed that no patterns are emerging from the combined analysis. Although some pooling methods help obtain better correlation results for some metrics, we do not observe a categorical preference among pooling methods. Similarly, the effect of temporal sub-sampling rate on the pooling methods is somewhat arbitrary, and increasing sub-sampling rate does not necessarily increase the metric accuracy for a given pooling method.

9.4 Conclusion

In this study, we conducted comprehensive experiments with 30 different metrics to investigate the effect of temporal sub-sampling and temporal pooling methods on the accuracy of volumetric video quality assessment. First, we investigated the effect of the temporal sampling rate. Our findings indicate that, even by sub-sampling the frame rate to 1 fps, metric scores and the metrics' performances do not show a significant difference compared to the full frame rate, i.e., 30 fps. In our experiment with different temporal pooling methods, we observed that better performances were achieved for image-based metrics by using the VQ-Pooling. We did not observe any categorical preference for color and point-based metrics among the tested temporal pooling methods.

Results show the temporal sub-sampling has minimal effect on metrics' correlations with ground truth subjective scores. This observation indicates that compression artifacts affect the perceived quality of the volumetric video uniformly in time. Our findings suggest that with no significant loss in the accuracy of both types of objective quality metrics, calculations can be sped up to 30 times for stimuli with point cloud compression artifacts. It should be noted that further research is required to further extend current conclusions for other types of distortions.

Each considered pooling method has a different priority for the temporal dimension. In our experiments, we observed minimal changes in metric performances with different pooling methods. Similar to the sub-sampling experiments, this indicates the uniform impact of the point cloud compression artifacts on perceived quality.

Our results provide insight regarding performances of various objective metrics for quality evaluation of point cloud compression algorithms on volumetric videos. Additionally, we provide statistical analysis for temporal pooling method selection for each metric. Finally, we show that the objective evaluation of the point cloud compression is minimally affected by the temporal sub-sampling rate, which allows the community to increase the computation efficiency of objective quality evaluation without sacrificing accuracy.

Conclusion & Perspective

As we see in the thesis, a number of contributions has been made covering both subjective and objective image quality assessment. To conclude the thesis, we will summarize our contributions around two main perspectives.

Subjective experiments to assess image quality is crucial for delivering a hyper-realistic user experience. Although standards and recommendations for subjective image quality assessment have been well established over the last few decades, new media formats bring additional concerns that are not well covered by the existing standards. Furthermore, for an extended period, pandemic conditions made it impossible to conduct such experiments within laboratory environments as it was traditionally done. Consequently, transferring laboratory experiments to crowdsourcing platforms gained urgent attention. This transition creates many challenges which previously were not a concern. Ultimately, like many other, we found ourselves asking the question:

Is crowdsourcing a viable solution for image quality assessment?

We investigated this question in detail at the first part of the thesis. Note that despite the use-case being tone mapped image quality evaluation, our findings shall be extended to a number of QoE scenarios. To this end, we conducted several studies including a large-scale experiment with 3500 participants. Our findings indicate that through appropriate experimental design and proper screening tools, crowdsourcing provides immense value for image quality domain. It brings marginal gains to the data-driven research.

An important factor in the success of transitioning from the laboratory to crowd was surely the experiment design. Reducing the load on the observer by splitting experiment into smaller chunks increased the reliability of the subjective annotations. Furthermore, relying on pairwise comparison rather than a rating task surely simplified the task for naive viewers with lower attention span.

Simplifying the task is not the only reason why we utilized pairwise comparison in our subjective studies. As discussed earlier, pairwise comparison is a more natural task for observers in terms of image quality evaluation. It does not rely on the understanding of a quality range (i.e., the expectations of good/bad image quality) for the observer by simply asking for a comparison between two alternatives. This brings us to the next main point we emphasised during the thesis.

Image quality does not lie on a continuous scale

HVS surely does not perceive the image quality as continuous but rather as a staircase function. In other words, we cannot perceive every small change in quality. This fact can be utilized in many aspects of image quality assessment. In experiment design this makes us better at pairwise comparison rather than direct rating. In the case of objective quality metrics, A-B comparison paradigm creates a valuable assessment scenario for real-life applications. In evaluation, alternative methodologies rely on significant difference rather than traditional correlation measures.

In this thesis, we utilized this knowledge and we made our contributions considering this fact. We demonstrated that an objective quality metric developed on first JND step information are capable of generalization to unseen distortions and supra-threshold level impairments. For transferring laboratory experiments to crowdsourcing, pairwise comparison design provides a more reliable data collection. Due to limited information shared in publicly available datasets, robust evaluation scenarios, such Krasula model, may not be utilized. In RV-TMO dataset, as an effort to enable and promote robust evaluation of objective quality metric performances, we released the raw scores as well as required scripts for the evaluation.

ANNEXES

Table 9.5 – PCC values between metric scores and DMOS for different temporal sampling rates with arithmetic mean.

	1-fps	2-fps	3-fps	5-fps	6-fps	10-fps	15-fps	30-fps
MP-PSNR-FR	0.7473	0.2655	0.7249	0.2905	0.3680	0.4008	0.3759	0.3088
MP-PSNR-RR	0.7287	0.7063	0.7394	0.7580	0.7098	0.7595	0.7089	0.7594
MW-PSNR-FR	0.7326	0.7432	0.6929	0.7357	0.7413	0.7404	0.7118	0.7378
MW-PSNR-RR	0.7155	0.7596	0.7180	0.7541	0.7269	0.7219	0.7559	0.7591
PSNR	0.8413	0.8298	0.8425	0.8406	0.8305	0.8298	0.8290	0.8286
SSIM	**0.9109**	**0.9093**	**0.9088**	**0.9076**	**0.9083**	**0.9082**	**0.9078**	**0.9081**
NIQSV	0.1806	0.2836	0.1668	0.1548	0.3423	0.3519	0.1596	0.1496
NIQSV+	0.2526	0.2724	0.2857	0.2884	0.2747	0.2727	0.2884	0.2734
APT	0.3624	0.2989	0.3152	0.3125	0.3079	0.3082	0.3088	0.3073
EM-IQM	0.3543	0.3882	0.4035	0.4253	0.4696	0.4309	0.4158	0.4283
SI-IQM	0.8897	0.8872	0.8855	0.8865	0.8861	0.8868	0.8875	0.8871
Color-Y	**0.8498**	**0.8485**	**0.8474**	**0.8464**	**0.8449**	**0.8450**	**0.8472**	**0.8453**
Color-Y-PSNR	0.8495	0.8442	0.8348	0.8460	0.8446	0.8377	0.8464	0.8447
Color-U	0.5552	0.5523	0.5534	0.5545	0.5504	0.5519	0.5548	0.5520
Color-U-PSNR	0.5463	0.5548	0.5572	0.5773	0.5525	0.5457	0.5564	0.5536
Color-V	0.5811	0.5775	0.5819	0.5828	0.5782	0.5781	0.5829	0.5783
Color-V-PSNR	0.5285	0.5771	0.5281	0.5284	0.5559	0.5706	0.5789	0.5688
p2point-Haus	0.2289	0.1649	0.1846	0.3340	0.2059	0.1539	0.1495	0.2092
p2point-Haus-PSNR	0.4670	0.4792	0.4871	0.4796	0.4918	0.4797	0.4875	0.4867
p2point-RMS	**0.9079**	**0.9073**	**0.9081**	**0.9081**	**0.9077**	**0.9068**	**0.9066**	**0.9068**
p2point-RMS-PSNR	0.8743	0.8865	0.8817	0.8795	0.8775	0.8801	0.8789	0.8750
p2plane-Haus	0.1834	0.1150	0.1438	0.1821	0.1224	0.0663	0.1650	0.1274
p2plane-Haus-PSNR	0.2041	0.2041	0.2001	0.1966	0.1898	0.1972	0.1948	0.1994
p2plane-RMS	0.4283	0.4316	0.4314	0.4294	0.4329	0.4433	0.4262	0.4315
p2plane-RMS-PSNR	0.8085	0.8144	0.8123	0.8123	0.5608	0.8130	0.8118	0.8088
pl2plane-MSE	0.4862	0.4813	0.4850	0.4852	0.4862	0.4849	0.4839	0.4863
pl2plane-RMS	0.4870	0.4835	0.4859	0.4873	0.4867	0.4861	0.4870	0.4870
pl2plane-Mean	0.4876	0.4852	0.4879	0.4891	0.4892	0.4884	0.4891	0.4893
pl2plane-Median	0.4876	0.4852	0.4879	0.4891	0.4892	0.4884	0.4891	0.4893
pl2plane-Min	0.0737	0.1434	0.1674	0.1800	0.1690	0.1370	0.1347	0.1282

Table 9.6 – RMSE values between metric scores and DMOS for different temporal sampling rates with arithmetic mean.

	1-fps	2-fps	3-fps	5-fps	6-fps	10-fps	15-fps	30-fps
MP-PSNR-FR	13.2316	19.1961	13.7214	19.0517	18.5129	18.2415	18.4535	18.9374
MP-PSNR-RR	13.6348	14.0949	13.4371	12.9869	14.0262	12.9527	14.0424	12.9692
MW-PSNR-FR	13.5513	13.3216	14.3563	13.4860	13.3643	13.3840	13.9899	13.4405
MW-PSNR-RR	13.9099	12.9490	13.8577	13.0759	13.6740	13.7775	13.0525	12.9618
PSNR	10.7628	11.1110	10.7241	10.7845	11.0891	11.1124	11.1351	11.1476
SSIM	**8.2153**	**8.2846**	**8.3087**	**8.3582**	**8.3279**	**8.3329**	**8.3515**	**8.3390**
NIQSV	19.5910	19.0928	19.6315	19.6706	18.7075	18.6376	19.6554	19.6864
NIQSV+	19.2647	19.1575	19.0806	19.0645	19.1443	19.1556	19.0641	19.1515
APT	18.5571	19.0000	18.8956	18.9132	18.9429	18.9408	18.9370	18.9471
EM-IQM	18.6184	18.3488	18.2175	18.0197	17.5784	17.9672	18.1072	17.9914
SI-IQM	9.0906	9.1860	9.2493	9.2113	9.2276	9.2026	9.1745	9.1916
Color-Y	**10.4972**	**10.5358**	**10.5724**	**10.6020**	**10.6497**	**10.6476**	**10.5766**	**10.6395**
Color-Y-PSNR	10.5029	10.6718	10.9625	10.6170	10.6608	10.8819	10.6022	10.6556
Color-U	16.5599	16.5983	16.5832	16.5687	16.6232	16.6036	16.5650	16.6021
Color-U-PSNR	16.6767	16.5648	16.5329	16.2584	16.5961	16.6856	16.5436	16.5810
Color-V	16.2037	16.2550	16.1920	16.1795	16.2454	16.2457	16.1784	16.2440
Color-V-PSNR	16.9023	16.2606	16.9077	16.9037	16.5558	16.3510	16.2348	16.3752
p2point-Haus	19.3829	19.6378	19.5682	18.7670	19.4838	19.8117	19.6865	19.4697
p2point-Haus-PSNR	17.6056	17.4758	17.3888	17.4712	17.3366	17.4702	17.3843	17.3933
p2point-RMS	**8.3471**	**8.3718**	**8.3395**	**8.3358**	**8.3561**	**8.3928**	**8.4019**	**8.3956**
p2point-RMS-PSNR	9.6631	9.2117	9.3933	9.4769	9.5471	9.4525	9.4968	9.6374
p2plane-Haus	19.5869	19.7785	19.7047	19.5867	19.7610	19.8899	19.6393	19.7480
p2plane-Haus-PSNR	19.4916	19.4912	19.5075	19.5217	19.5501	19.5198	19.5290	19.5107
p2plane-RMS	17.9921	17.9604	17.9619	17.9828	17.9495	17.8619	18.0118	17.9618
p2plane-RMS-PSNR	11.7182	11.5536	11.6126	11.6129	16.4991	11.5933	11.6250	11.7279
pl2plane-MSE	17.3989	17.4521	17.4117	17.4102	17.3989	17.4135	17.4241	17.3980
pl2plane-RMS	17.3899	17.4286	17.4024	17.3869	17.3932	17.3992	17.3896	17.3897
pl2plane-Mean	17.3828	17.4098	17.3797	17.3662	17.3652	17.3746	17.3665	17.3646
pl2plane-Median	17.3828	17.4098	17.3797	17.3662	17.3652	17.3746	17.3665	17.3646
pl2plane-Min	19.8562	19.7044	19.6293	19.5857	19.6241	19.7227	19.7287	19.7460

Table 9.7 – PCC values between metric scores and DMOS for different pooling methods with 30 fps.

	Arithmetic	Harmonic	Minkowski	Percentile	Primacy	Recency	VQ
	mean	mean	mean	pooling	pooling	pooling	pooling
MP-PSNR-FR	0.3088	0.7738	0.6641	0.5177	0.3282	0.2996	0.8260
MP-PSNR-RR	0.7594	0.7000	0.7669	0.5283	0.7355	0.7176	0.7932
MW-PSNR-FR	0.7378	0.6817	0.7401	0.5415	0.7010	0.7121	0.7631
MW-PSNR-RR	0.7591	0.7084	0.7251	0.5409	0.7668	0.7291	0.7698
PSNR	0.8286	0.8332	0.8426	0.8103	0.8258	0.8466	0.8428
SSIM	**0.9081**	**0.9081**	**0.9081**	**0.9190**	**0.9077**	**0.9115**	**0.9427**
NIQSV	0.1496	0.1564	0.3436	0.2367	0.3207	0.1561	0.1639
NIQSV+	0.2734	0.2754	0.2745	0.2323	0.2892	0.3154	0.3527
APT	0.3073	0.3073	0.3073	0.3086	0.3052	0.3183	0.1405
EM-IQM	0.4283	0.6565	0.4463	0.6035	0.3913	0.4249	0.1877
SI-IQM	0.8871	0.8874	0.8868	0.8870	0.8892	0.8853	0.8736
Color-Y	**0.8453**	**0.8460**	0.8443	**0.8531**	**0.8445**	**0.8461**	**0.8408**
Color-Y-PSNR	0.8447	0.8448	**0.8449**	0.8296	0.8443	0.8345	0.8352
Color-U	0.5520	0.5556	0.5482	0.5645	0.5520	0.5537	0.5694
Color-U-PSNR	0.5536	0.5533	0.5583	0.5270	0.5432	0.5668	0.5545
Color-V	0.5783	0.5803	0.5766	0.5889	0.5779	0.5777	0.5940
Color-V-PSNR	0.5688	0.5282	0.5800	0.5567	0.5245	0.5842	0.5776
p2point-Haus	0.2092	0.2888	0.1356	0.5604	0.1367	0.1866	-
p2point-Haus-PSNR	0.4867	0.4465	0.3558	0.4806	0.4745	0.4917	-
p2point-RMS	**0.9068**	**0.9050**	**0.9009**	0.9005	**0.9043**	**0.9060**	**0.9021**
p2point-RMS-PSNR	0.8750	0.8848	0.8667	**0.9202**	0.8777	0.8709	0.8184
p2plane-Haus	0.1274	0.1802	0.1234	0.2114	0.1388	0.0992	-
p2plane-Haus-PSNR	0.1994	0.1976	0.0241	0.1455	0.2017	0.1976	0.1028
p2plane-RMS	0.4315	0.4273	0.4300	0.4285	0.4258	0.4447	0.4318
p2plane-RMS-PSNR	0.8088	0.8154	0.7320	0.7648	0.8119	0.8039	0.8886
pl2plane-MSE	0.4863	0.4863	0.4861	0.4613	0.4821	0.4962	0.4670
pl2plane-RMS	0.4870	0.4870	0.4870	0.4751	0.4825	0.4985	0.4678
pl2plane-Mean	0.4893	0.4893	0.4895	0.4774	0.4843	0.4892	0.4539
pl2plane-Median	0.4893	0.4893	0.4895	0.4774	0.4843	0.4892	0.4539
pl2plane-Min	0.1282	0.0842	0.1896	0.1102	0.1792	0.1750	0.2079

Table 9.8 – RMSE values between metric scores and DMOS for different pooling methods with 30 fps.

	Arithmetic mean	Harmonic mean	Minkowski mean	Percentile pooling	Primacy pooling	Recency pooling	VQ Pooling
MP-PSNR-FR	18.9374	12.6152	14.8858	17.0350	18.8089	18.9961	11.2251
MP-PSNR-RR	12.9692	14.2188	12.7770	16.9044	13.4923	13.8670	12.1240
MW-PSNR-FR	13.4405	14.5667	13.3902	16.7387	14.1999	13.9793	12.8735
MW-PSNR-RR	12.9618	14.0529	13.7104	16.7467	12.8077	13.6272	12.7204
PSNR	11.1476	11.0099	10.7230	11.6741	11.2301	10.5974	10.7176
SSIM	**8.3390**	**8.3364**	**8.3356**	**7.8494**	**8.3556**	**8.1911**	**6.6402**
NIQSV	19.6864	19.6654	18.6981	19.3453	18.8591	19.6662	19.6458
NIQSV+	19.1515	19.1404	19.1456	19.3655	19.0594	18.8938	18.6312
APT	18.9471	18.9471	18.9471	18.9386	18.9605	18.8747	19.7128
EM-IQM	17.9914	15.0186	17.8175	15.8760	18.3230	18.0233	19.5564
SI-IQM	9.1916	9.1791	9.2005	9.1951	9.1112	9.2587	9.6910
Color-Y	**10.6395**	**10.6152**	10.6727	**10.3891**	**10.6634**	**10.6161**	**10.7774**
Color-Y-PSNR	10.6556	10.6531	**10.6512**	11.1160	10.7138	10.9713	10.9497
Color-U	16.6021	16.5550	16.6526	16.4346	16.6026	16.5797	16.3672
Color-U-PSNR	16.5810	16.5844	16.5196	16.9215	16.7203	16.4109	16.5690
Color-V	16.2440	16.2144	16.2674	16.0916	16.2491	16.2513	16.0180
Color-V-PSNR	16.3752	16.9061	16.2192	16.5404	16.9524	16.1588	16.2543
p2point-Haus	19.4697	19.0620	19.7264	16.4905	19.7234	19.5651	-
p2point-Haus-PSNR	17.3933	17.8151	18.6076	17.4600	17.5263	17.3368	-
p2point-RMS	**8.3956**	**8.4702**	**8.6405**	**8.6567**	**8.4994**	**8.4261**	**8.5917**
p2point-RMS-PSNR	9.6374	9.2760	9.9324	7.7939	9.5404	9.7849	11.4426
p2plane-Haus	19.7480	19.6215	19.7582	19.4605	19.7183	19.8122	-
p2plane-Haus-PSNR	19.5107	19.5891	19.9045	19.7141	19.5012	19.5180	19.8053
p2plane-RMS	17.9618	18.0015	17.9753	17.9897	18.0164	17.8330	17.9590
p2plane-RMS-PSNR	11.7279	11.5266	13.5670	12.8267	11.6232	11.8421	9.1364
pl2plane-MSE	17.3980	17.3971	17.3998	17.6650	17.4432	17.2871	17.6062
pl2plane-RMS	17.3897	17.3898	17.3901	17.5201	17.4393	17.2597	17.5970
pl2plane-Mean	17.3646	17.3644	17.3621	17.4948	17.4193	17.3658	17.7410
pl2plane-Median	17.3646	17.3644	17.3621	17.4948	17.4193	17.3658	17.7410
pl2plane-Min	19.7460	19.8395	19.5491	19.7890	19.5881	19.6031	19.4752

Bibliography

[1] Vamsi Kiran Adhikarla et al., « Towards a Quality Metric for Dense Light Fields », *in*: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 3720–3729 (cit. on pp. 116, 117, 120, 122).

[2] Ayyoub Ahar, Adriaan Barri, and Peter Schelkens, « From sparse coding significance to perceptual quality: A new approach for image quality assessment », *in*: *IEEE Transactions on Image Processing* 27.2 (2017), pp. 879–893 (cit. on p. 120).

[3] Ali Ak and Patrick Le Callet, « Towards Perceptually Plausible Training of Image Restoration Neural Networks », *in*: *Ninth International Conference on Image Processing Theory, Tools and Applications, IPTA 2019, Istanbul, Turkey, November 6-9, 2019*, IEEE, 2019, pp. 1–5, DOI: `10.1109/IPTA.2019.8936096`, URL: `https://doi.org/10.1109/IPTA.2019.8936096` (cit. on p. 105).

[4] Ali Ak and Patrick Le-Callet, « Investigating Epipolar Plane Image Representations for Objective Quality Evaluation of Light Field Images », *in*: *2019 8th European Workshop on Visual Information Processing (EUVIP)*, 2019, pp. 135–139, DOI: `10.1109/EUVIP47703.2019.8946194` (cit. on p. 111).

[5] Ali Ak, Suiyi Ling, and Patrick Le Callet, « No-Reference Quality Evaluation of Light Field Content Based on Structural Representation of The Epipolar Plane Image », *in*: *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2020, pp. 1–6, DOI: `10.1109/ICMEW46912.2020.9105975` (cit. on pp. 111, 122).

[6] Ali Ak et al., « A Comprehensive Analysis of Crowdsourcing for Subjective Evaluation of Tone Mapping Operators », *in*: *Image Quality and System Performance, IS T International Symposium on Electronic Imaging (EI 2021)*, San Francisco, United States, Jan. 2021, URL: `https://hal.archives-ouvertes.fr/hal-03020972` (cit. on pp. 26, 33, 43).

[7] Ali Ak et al., « On Spammer Detection In Crowdsourcing Pairwise Comparison Tasks: Case Study On Two Multimedia Qoe Assessment Scenarios », *in*: *2021 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2021, pp. 1–6, DOI: `10.1109/ICMEW53276.2021.9455992` (cit. on p. 25).

[8] Ali Ak et al., « The Effect of Temporal Sub-sampling on the Accuracy of Volumetric Video Quality Assessment », *in*: *2021 Picture Coding Symposium (PCS)* (2021), pp. 1–5 (cit. on pp. 125, 135).

[9] Evangelos Alexiou and Touradj Ebrahimi, « Point cloud quality assessment metric based on angular similarity », *in*: *International Conference on Multimedia & Expo (ICME)*, 2018 (cit. on pp. 94, 131).

[10] *Amazon Mechanical Turk*, `https://www.mturk.com`, Accessed: Aug 2021. [Online] (cit. on p. 31).

[11] Alessandro Artusi et al., « Overview and evaluation of the JPEG XT HDR image compression standard », *in*: *Journal of Real-Time Image Processing* 16.*2* (2019), pp. 413–428 (cit. on pp. 35, 60).

[12] George Alfred Barnard, « A new test for 2× 2 tables », *in*: *Natur* 156.*3954* (1945), p. 177 (cit. on pp. 23, 49, 67, 69).

[13] Yoav Benjamini and Yosef Hochberg, « Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing », *in*: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.*1* (1995), pp. 289–300, DOI: `https://doi.org/10.1111/j.2517-6161.1995.tb02031.x`, eprint: `https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1995.tb02031.x`, URL: `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.x` (cit. on p. 97).

[14] RALPH ALLAN BRADLEY, « RANK ANALYSIS OF INCOMPLETE BLOCK DESIGNS: III. SOME LARGE-SAMPLE RESULTS ON ESTIMATION AND POWER FOR A METHOD OF PAIRED COMPARISONS* », *in*: *Biometrika* 42.*3-4* (Dec. 1955), pp. 450–470, ISSN: 0006-3444, DOI: `10.1093/biomet/42.3-4.450`, eprint: `https://academic.oup.com/biomet/article-pdf/42/3-4/450/838690/42-3-4-450.pdf`, URL: `https://doi.org/10.1093/biomet/42.3-4.450` (cit. on pp. 23, 73).

[15] J. Canny, « A Computational Approach to Edge Detection », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8.*6* (1986), pp. 679–698, ISSN: 0162-8828, DOI: `10.1109/TPAMI.1986.4767851` (cit. on pp. 116, 117).

[16] Aaron Carass et al., « Evaluating White Matter Lesion Segmentations with Refined Sørensen-Dice Analysis », English (US), *in*: *Scientific Reports* 10.*1* (Dec. 2020), ISSN: 2045-2322, DOI: `10.1038/s41598-020-64803-w` (cit. on p. 80).

[17] Xim Cerdá-Company, C. Alejandro Párraga, and Xavier Otazu, « Which tone-mapping operator is the best? A comparative study of perceptual quality », *in*: *CoRR* abs/1601.04450 (2016), arXiv: `1601.04450`, URL: `http://arxiv.org/abs/1601.04450` (cit. on p. 60).

[18] Justin Cheng, Jaime Teevan, and Michael S. Bernstein, « Measuring Crowdsourcing Effort with Error-Time Curves », *in*: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, Seoul, Republic of Korea: Association for Computing Machinery, 2015, 1365–1374, ISBN: 9781450331456, DOI: `10.1145/2702123.2702145`, URL: `https://doi.org/10.1145/2702123.2702145` (cit. on p. 27).

[19] Jacob Cohen, « A Coefficient of Agreement for Nominal Scales », *in*: *Educational and Psychological Measurement* 20.*1* (1960), pp. 37–46, DOI: `10.1177/001316446002000104`, eprint: `https://doi.org/10.1177/001316446002000104`, URL: `https://doi.org/10.1177/001316446002000104` (cit. on p. 25).

[20] Eugene d'Eon et al., *8i Voxelized Full Bodies - A Voxelized Point Cloud Dataset, ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document WG11M40059/WG Geneva*, 2017 (cit. on p. 126).

[21] F. Drago et al., « Adaptive Logarithmic Mapping For Displaying High Contrast Scenes », *in*: *Computer Graphics Forum* 22.*3* (2003), pp. 419–426, DOI: `https://doi.org/10.1111/1467-8659.00689`, eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-8659.00689`, URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-8659.00689` (cit. on p. 29).

[22] Sebastian Egger-Lampl et al., « Crowdsourcing Quality of Experience Experiments », *in*: *Crowdsourcing and Human-Centered Experiments*, 2015 (cit. on p. 26).

[23] G. Eilertsen, R. K. Mantiuk, and J. Unger, « A comparative review of tone-mapping algorithms for high dynamic range video », *in*: *Computer Graphics Forum* 36.*2* (2017), pp. 565–592, DOI: https://doi.org/10.1111/cgf.13148, eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13148, URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13148 (cit. on p. 60).

[24] M.D. Fairchild, « The HDR photographic survey », *in*: (2007), pp. 233–238 (cit. on pp. 35, 46, 60, 86).

[25] Chunling Fan et al., « Interactive Subjective Study on Picture-level Just Noticeable Difference of Compressed Stereoscopic Images », *in*: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), pp. 8548–8552 (cit. on p. 100).

[26] Yuming Fang et al., « Light Filed Image Quality Assessment by Local and Global Features of Epipolar Plane Image », *in*: *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, 2018, pp. 1–6, DOI: 10.1109/BigMM.2018.8499086 (cit. on p. 94).

[27] Ronald A Fisher, « On the interpretation of χ 2 from contingency tables, and the calculation of P », *in*: *Journal of the Royal Statistical Society* 85.*1* (1922), pp. 87–94 (cit. on pp. 23, 67, 97).

[28] Michael Faraday Esq. D.C.L. F.R.S., « LIV. Thoughts on ray-vibrations », *in*: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 28.*188* (1846), pp. 345–350, DOI: 10.1080/14786444608645431 (cit. on p. 111).

[29] Xinbo Gao et al., « Image quality assessment and human visual system », *in*: *Proceedings of SPIE - The International Society for Optical Engineering* 7744 (July 2010), DOI: 10.1117/12.862431 (cit. on p. 102).

[30] B. Gardlo et al., « Microworkers vs. facebook: The impact of crowdsourcing platform choice on experimental results », *in*: *2012 Fourth International Workshop on Quality of Multimedia Experience*, 2012, pp. 35–36, DOI: 10.1109/QoMEX.2012.6263885 (cit. on p. 26).

[31] Bruno Gardlo et al., « Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based QoE testing », *in*: *2014 IEEE International Conference on*

Communications (ICC), 2014, pp. 1070–1075, DOI: 10.1109/ICC.2014.6883463 (cit. on pp. 17, 26, 47, 61).

[32] *Google Draco*, https://google.github.io/draco/, Accessed: Aug 2021. [Online] (cit. on p. 127).

[33] Abhishek Goswami et al., « Reliability of Crowdsourcing for Subjective Quality Evaluation of Tone Mapping Operators », *in*: *IEEE International Workshop on Multimedia Signal Processing (MMSP'2021)*, Tampere, Finland, Oct. 2021, URL: https://hal.archives-ouvertes.fr/hal-03298957 (cit. on p. 27).

[34] Abhishek Goswami et al., « Reliability of Crowdsourcing for Subjective Quality Evaluation of Tone Mapping Operators », *in*: *IEEE International Workshop on Multimedia Signal Processing (MMSP'2021)*, Tampere, Finland, Oct. 2021, URL: https://hal.archives-ouvertes.fr/hal-03298957 (cit. on pp. 33, 43).

[35] Abhishek Goswami et al., « Tone Mapping Operators: Progressing Towards Semantic-awareness », *in*: *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, 2020, pp. 1–6 (cit. on pp. 29, 34, 46, 60, 73, 86).

[36] Ke Gu et al., « Blind Quality Assessment of Tone-Mapped Images Via Analysis of Information, Naturalness, and Structure », *in*: *IEEE Transactions on Multimedia* 18 (Mar. 2016), pp. 1–1, DOI: 10.1109/TMM.2016.2518868 (cit. on p. 68).

[37] Ke Gu et al., « Model-Based Referenceless Quality Metric of 3D Synthesized Images Using Local Image Description », *in*: *IEEE Transactions on Image Processing* (2017) (cit. on p. 131).

[38] Ke Gu et al., « No-Reference Quality Metric of Contrast-Distorted Images Based on Information Maximization », *in*: *IEEE Transactions on Cybernetics* 47.*12* (2017), pp. 4559–4565, DOI: 10.1109/TCYB.2016.2575544 (cit. on p. 68).

[39] Ke Gu et al., « Using Free Energy Principle For Blind Image Quality Assessment », *in*: *Multimedia, IEEE Transactions on* 17 (Jan. 2015), pp. 50–63, DOI: 10.1109/TMM.2014.2373812 (cit. on p. 122).

[40] R. Hadsell, S. Chopra, and Y. LeCun, « Dimensionality Reduction by Learning an Invariant Mapping », *in*: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, 2006, pp. 1735–1742, DOI: 10.1109/CVPR.2006.100 (cit. on pp. 103, 106).

[41] J.A. Hanley and Barbara Mcneil, « A Method of Comparing the Areas Under Receiver Operating Characteristic Curves Derived from the Same Cases », *in*: *Radiology* 148 (Oct. 1983), pp. 839–43, DOI: `10.1148/radiology.148.3.6878708` (cit. on p. 97).

[42] J.A. Hanley and Barbara Mcneil, « The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve », *in*: *Radiology* 143 (May 1982), pp. 29–36, DOI: `10.1148/radiology.143.1.7063747` (cit. on p. 97).

[43] David Hasler and Sabine Suesstrunk, « Measuring Colourfulness in Natural Images », *in*: *Proceedings of SPIE - The International Society for Optical Engineering* 5007 (June 2003), pp. 87–95, DOI: `10.1117/12.477378` (cit. on p. 22).

[44] T. Hossfeld et al., « Best Practices and Recommendations for Crowdsourced QoE - Lessons learned from the Qualinet Task Force Crowdsourcing », *in*: 2014 (cit. on pp. 26, 27, 43, 61).

[45] Tobias Hoßfeld et al., « Survey of web-based crowdsourcing frameworks for subjective quality assessment », *in*: *2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP)*, 2014, pp. 1–6, DOI: `10.1109/MMSP.2014.6958831` (cit. on p. 27).

[46] Qin Huang et al., « Measure and Prediction of HEVC Perceptually Lossy/Lossless Boundary QP Values », *in*: *2017 Data Compression Conference (DCC)* (2017), pp. 42–51 (cit. on p. 100).

[47] Vedad Hulusic et al., « Perceived dynamic range of HDR images », *in*: *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6, DOI: `10.1109/QoMEX.2016.7498953` (cit. on p. 22).

[48] ITU-T, *Methodologies for the subjective assessment of the quality of television images*, ITU-R Recommendation BT.500-14, ITU-R Std. 2019 (cit. on pp. 17, 18, 23, 24, 47, 75, 79).

[49] ITU-T, *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*, ITU-T Recommendation P.1401, ITU-T Std. 2019 (cit. on p. 95).

[50] ITU-T, *Subjective evaluation of media quality using a crowdsourcing approach*, ITU-T Technical Report PSTR-CROWDS, 2018 (cit. on pp. 17, 26, 43).

[51] ITU-T, *Subjective video quality assessment methods for multimedia applications*, ITU-T Recommendation P.910, 2008 (cit. on p. 127).

[52] ITU-T SG09, *Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, phase II (FR-TV2)*, ITU-T SG09, 2004 (cit. on p. 95).

[53] Paul Jaccard, « Étude comparative de la distribution florale dans une portion des Alpes et des Jura », *in*: *Bulletin del la Société Vaudoise des Sciences Naturelles* 37 (1901), pp. 547–579 (cit. on pp. 52, 80).

[54] Lina Jin et al., « Statistical Study on Perceived JPEG Image Quality via MCL-JCI Dataset Construction and Analysis », *in*: *electronic imaging* 2016 (2016), pp. 1–9 (cit. on pp. 100–106).

[55] M. G. KENDALL, « A New Measure of Rank Correlation », *in*: *Biometrika* 30.*1-2* (June 1938), pp. 81–93, ISSN: 0006-3444, DOI: `10.1093/biomet/30.1-2.81`, eprint: `https://academic.oup.com/biomet/article-pdf/30/1-2/81/423380/30-1-2-81.pdf`, URL: `https://doi.org/10.1093/biomet/30.1-2.81` (cit. on p. 25).

[56] Min H Kim, Jan Kautz, et al., « Consistent tone reproduction », *in*: *Proceedings of the Tenth IASTED International Conference on Computer Graphics and Imaging*, ACTA Press Anaheim, 2008, pp. 152–159 (cit. on pp. 29, 34, 38, 40, 46, 60, 73, 86).

[57] Lukáš Krasula, Karel Fliegel, and Patrick Le Callet, « FFTMI: Features Fusion for Natural Tone-Mapped Images Quality Evaluation », *in*: *IEEE Transactions on Multimedia* 22.*8* (2019), pp. 2038–2047 (cit. on p. 68).

[58] Lukas Krasula et al., « Preference of Experience in Image Tone-Mapping: Dataset and Framework for Objective Measures Comparison », *in*: *IEEE Journal of Selected Topics in Signal Processing* 11.*1* (Feb. 2017), pp. 64 –74, DOI: `10.1109/JSTSP.2016.2637168`, URL: `https://hal.archives-ouvertes.fr/hal-01633843` (cit. on pp. 30, 31, 62).

[59] Lukáš Krasula et al., « On the accuracy of objective image and video quality models: New methodology for performance evaluation », *in*: *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6, DOI: `10.1109/QoMEX.2016.7498936` (cit. on pp. 68, 69, 73, 95, 108).

[60] Grzegorz Krawczyk, Karol Myszkowski, and Hans-Peter Seidel, « Lightness perception in tone reproduction for high dynamic range images », *in*: *Computer Graphics Forum*, vol. 24, 3, Citeseer, 2005, pp. 635–646 (cit. on pp. 29, 34, 38, 40, 46, 60, 73, 86).

[61] Klaus Krippendorff, « Estimating the Reliability, Systematic Error and Random Error of Interval Data », *in*: *Educational and Psychological Measurement* 30.*1* (1970), pp. 61–70, DOI: `10.1177/001316447003000105`, URL: `https://doi.org/10.1177/001316447003000105` (cit. on pp. 25, 52, 54).

[62] Debarati Kundu et al., « Large-Scale Crowdsourced Study for Tone-Mapped HDR Pictures », *in*: *IEEE Transactions on Image Processing* 26.*10* (2017), pp. 4725–4740, DOI: `10.1109/TIP.2017.2713945` (cit. on p. 31).

[63] G.W. Larson, H. Rushmeier, and C. Piatko, « A visibility matching tone reproduction operator for high dynamic range scenes », *in*: *IEEE Transactions on Visualization and Computer Graphics* 3.*4* (1997), pp. 291–306, DOI: `10.1109/2945.646233` (cit. on p. 29).

[64] Patrick Ledda et al., « Evaluation of Tone Mapping Operators Using a High Dynamic Range Display », *in*: *ACM Trans. Graph.* 24.*3* (July 2005), 640–648, ISSN: 0730-0301, DOI: `10.1145/1073204.1073242`, URL: `https://doi.org/10.1145/1073204.1073242` (cit. on p. 31).

[65] Jing Li, Marcus Barkowsky, and Patrick Le Callet, « Boosting paired comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs », *in*: *Stereoscopic Displays and Applications XXIV*, ed. by Andrew J. Woods, Nicolas S. Holliman, and Gregg E. Favalora, vol. 8648, International Society for Optics and Photonics, SPIE, 2013, pp. 547 –558, DOI: `10.1117/12.2002075`, URL: `https://doi.org/10.1117/12.2002075` (cit. on p. 21).

[66] Jing Li, Marcus Barkowsky, and Patrick Le Callet, « Boosting Paired Comparison methodology in measuring visual discomfort of 3DTV: Performances of three different designs », *in*: *Proceedings of SPIE - The International Society for Optical Engineering* 8648 (Mar. 2013), DOI: `10.1117/12.2002075` (cit. on p. 63).

[67] Hanhe Lin et al., « SUR-FeatNet: Predicting the Satisfied User Ratio Curvefor Image Compression with Deep Feature Learning », *in*: (Jan. 2020), DOI: `10.1007/s41233-020-00034-1`, URL: `http://arxiv.org/abs/2001.02002http://dx.doi.org/10.1007/s41233-020-00034-1` (cit. on p. 105).

[68] Suiyi Ling, Gene Cheung, and Patrick Le Callet, « No-Reference Quality Assessment for Stitched Panoramic Images Using Convolutional Sparse Coding and Compound Feature Selection », *in*: *2018 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2018, pp. 1–6 (cit. on p. 121).

[69] Suiyi Ling and Patrick Le Callet, « How to Learn the Effect of Non-Uniform Distortion on Perceived Visual Quality? Case Study Using Convolutional Sparse Coding for Quality Assessment of Synthesized Views », *in*: *2018 25th IEEE International Conference on Image Processing (ICIP)*, IEEE, 2018, pp. 286–290 (cit. on p. 121).

[70] Suiyi Ling and Patrick Le Callet, « Image quality assessment for DIBR synthesized views using elastic metric », *in*: *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1157–1163 (cit. on p. 131).

[71] Suiyi Ling and Patrick Le Callet, « Image quality assessment for free viewpoint video based on mid-level contours feature », *in*: *2017 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2017, pp. 79–84 (cit. on p. 131).

[72] Xiaohua Liu et al., « JND-Pano: Database for Just Noticeable Difference of JPEG Compressed Panoramic Images », *in*: *Advances in Multimedia Information Processing – PCM 2018*, ed. by Richang Hong et al., Cham: Springer International Publishing, 2018, pp. 458–468, ISBN: 978-3-030-00776-8 (cit. on p. 100).

[73] Rafal Mantiuk, Anna Tomaszewska, and Radoslaw Mantiuk, « Comparison of Four Subjective Methods for Image Quality Assessment », *in*: *Computer Graphics Forum* 31 (Nov. 2012), DOI: `10.1111/j.1467-8659.2012.03188.x` (cit. on p. 19).

[74] Rafal Mantiuk et al., « HDR-VDP-2: A Calibrated Visual Metric for Visibility and Quality Predictions in All Luminance Conditions », *in*: *ACM Trans. Graph.* 30.*4* (July 2011), ISSN: 0730-0301, DOI: `10.1145/2010324.1964935`, URL: `https://doi.org/10.1145/2010324.1964935` (cit. on pp. 93, 102, 103, 105).

[75] Cyrus R Mehta and Pralay Senchaudhuri, « Conditional versus unconditional exact tests for comparing two binomials », *in*: *Cytel Software Corporation* 675 (2003), pp. 1–5 (cit. on pp. 24, 67).

[76] Rufael Mekuria et al., *Evaluation criteria for PCC (Point Cloud Compression)*, ISO/IEC JTC 1/SC29/WG11 Doc. N16332, 2016 (cit. on pp. 94, 131).

[77] Anish Mittal, Anush K. Moorthy, and Alan C. Bovik, « Blind/Referenceless Image Spatial Quality Evaluator », *in*: *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 2011, pp. 723–727, DOI: `10.1109/ACSSC.2011.6190099` (cit. on p. 122).

[78] Frederick Mosteller, « Remarks on the Method of Paired Comparisons: I. The Least Squares Solution Assuming Equal Standard Deviations and Equal Correlations », *in*: *Selected Papers of Frederick Mosteller*, ed. by Stephen E. Fienberg and David C. Hoaglin, New York, NY: Springer New York, 2006, pp. 157–162, ISBN: 978-0-387-44956-2, DOI: `10.1007/978-0-387-44956-2_8` (cit. on p. 23).

[79] Bennet B. Murdock, « The Serial Position Effect of Free Recall », *in*: *Journal of Experimental Psychology* 64.*5* (1962), p. 482, DOI: `10.1037/h0045106` (cit. on p. 130).

[80] Bruno A Olshausen and David J Field, « Sparse coding of sensory inputs », *in*: *Current opinion in neurobiology* 14.*4* (2004), pp. 481–487 (cit. on p. 120).

[81] Jincheol Park et al., « Video Quality Pooling Adaptive to Perceptual Distortion Severity », *in*: *IEEE Transactions on Image Processing* 22 (2013), pp. 610–620 (cit. on pp. 129, 130).

[82] Eyal Peer et al., « Beyond the Turk: Alternative platforms for crowdsourcing behavioral research », *in*: *Journal of Experimental Social Psychology* 70 (2017), pp. 153–163, ISSN: 0022-1031, DOI: `https://doi.org/10.1016/j.jesp.2017.01.006` (cit. on p. 65).

[83] Josselin Petit and Rafał K. Mantiuk, « Assessment of Video Tone-Mapping: Are Cameras' S-Shaped Tone-Curves Good Enough? », *in*: 24.*7* (Oct. 2013), 1020–1030, ISSN: 1047-3203, DOI: `10.1016/j.jvcir.2013.06.014`, URL: `https://doi.org/10.1016/j.jvcir.2013.06.014` (cit. on pp. 30, 31).

[84] Sergio Pezzulli, Maria G. Martini, and Nabajeet Barman, « Estimation of Quality Scores From Subjective Tests-Beyond Subjects' MOS », *in*: *IEEE Transactions on Multimedia* 23 (2021), pp. 2505–2519, DOI: `10.1109/TMM.2020.3013349` (cit. on p. 23).

[85] Nikolay Ponomarenko et al., « Image database TID2013: Peculiarities, results and perspectives », *in*: *Signal Processing: Image Communication* 30 (2015), pp. 57 –77,

ISSN: 0923-5965, DOI: https://doi.org/10.1016/j.image.2014.10.009 (cit. on pp. 92, 106–108).

[86] Ekta Prashnani et al., *PieAPP: Perceptual Image-Error Assessment through Pairwise Preference*, 2018, arXiv: 1806.02067 [cs.CV] (cit. on pp. 64, 73).

[87] J. M. S. PREWITT, « Object enhancement and extraction », *in*: *Picture Processing and. Psychopictorics* (1970), URL: https://ci.nii.ac.jp/naid/10017095478/en/ (cit. on p. 117).

[88] *Prolific*, https://www.prolific.co/, Accessed: Aug 2021. [Online] (cit. on pp. 34, 44, 47, 52, 64, 75, 85, 86).

[89] María Pérez-Ortiz and Rafal Mantiuk, « A practical guide and software for analysing pairwise comparison experiments », *in*: (Dec. 2017) (cit. on p. 24).

[90] Erik Reinhard et al., « Photographic tone reproduction for digital images », *in*: *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 2002, pp. 267–276 (cit. on pp. 29, 34, 38, 40, 46, 60, 73, 86).

[91] David J. Rogers and Taffee T. Tanimoto, « A Computer Program for Classifying Plants », *in*: *Science* 132.*3434* (1960), pp. 1115–1118, ISSN: 0036-8075, DOI: 10.1126/science.132.3434.1115 (cit. on pp. 52, 76, 79, 80).

[92] D. M. Rouse and S. S. Hemami, « Natural image utility assessment using image contours », *in*: *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 2217–2220, DOI: 10.1109/ICIP.2009.5413882 (cit. on p. 117).

[93] S. Roy et al., « Siamese Networks: The Tale of Two Manifolds », *in*: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3046–3055, DOI: 10.1109/ICCV.2019.00314 (cit. on p. 103).

[94] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, « DIBR synthesized image quality assessment based on morphological wavelets », *in*: *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, 2015, pp. 1–6, DOI: 10.1109/QoMEX.2015.7148143 (cit. on pp. 117, 118, 122, 131).

[95] Dragana Sandić-Stanković, Dragan Kukolj, and Patrick Le Callet, « DIBR-synthesized image quality assessment based on morphological multi-scale approach », *in*: *EURASIP Journal on Image and Video Processing* 2017 (July 2016), DOI: 10.1186/s13640-016-0124-7 (cit. on pp. 122, 131).

[96] Dragana Sandić-Stanković, Dragan Kukolj, and Patrick Le Callet, « Multi-Scale Synthesized View Assessment Based on Morphological Pyramids », *in*: *Journal of Electrical Engineering* 67 (Jan. 2016), pp. 3–11, DOI: `10.1515/jee-2016-0001` (cit. on p. 131).

[97] Dragana Sandić-Stanković et al., « Free viewpoint video quality assessment based on morphological multiscale metrics », *in*: *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6, DOI: `10.1109/QoMEX.2016.7498949` (cit. on pp. 115–117).

[98] Kyoshiro Sasaki and Yuki Yamada, « Crowdsourcing visual perception experiments: a case of contrast threshold », *in*: *PeerJ* 7 (Dec. 2019), e8339, ISSN: 2167-8359, DOI: `10.7717/peerj.8339` (cit. on p. 27).

[99] Sebastian Schwarz et al., « Emerging MPEG Standards for Point Cloud Compression », *in*: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9.*1* (2019), pp. 133–148, DOI: `10.1109/JETCAS.2018.2885981` (cit. on pp. 126, 127, 134).

[100] M. Seufert et al., « "To pool or not to pool": A comparison of temporal pooling methods for HTTP adaptive video streaming », *in*: *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2013, pp. 52–57, DOI: `10.1109/QoMEX.2013.6603210` (cit. on pp. 128, 129).

[101] H.R. Sheikh and A.C. Bovik, « Image information and visual quality », *in*: *IEEE Transactions on Image Processing* 15.*2* (2006), pp. 430–444, DOI: `10.1109/TIP.2005.859378` (cit. on pp. 93, 122).

[102] X. Shen et al., « A JND Dataset Based on VVC Compressed Images », *in*: *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, Los Alamitos, CA, USA: IEEE Computer Society, 2020, pp. 1–6, DOI: `10.1109/ICMEW46912.2020.9105955` (cit. on p. 100).

[103] Likun Shi, Shengyang Zhao, and Zhibo Chen, « Belif: Blind Quality Evaluator Of Light Field Image With Tensor Structure Variation Index », *in*: *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 3781–3785, DOI: `10.1109/ICIP.2019.8803559` (cit. on pp. 94, 122, 123).

[104] Irwin Sobel, « An Isotropic 3x3 Image Gradient Operator », *in*: *Presentation at Stanford A.I. Project 1968* (Feb. 2014) (cit. on pp. 22, 117).

[105] C. Spearman, « The Proof and Measurement of Association between Two Things », *in*: *The American Journal of Psychology* 15 (1904), pp. 72–101 (cit. on p. 25).

[106] John A Swets, « Book Reviews : Signal Detection Theory and ROC Analysis in Psychology and Diagnostics : Collected Papers. », *in*: *Medical Decision Making* 19.*2* (1999), pp. 217–217, DOI: `10.1177/0272989X9901900216` (cit. on p. 96).

[107] Louis Leon Thurstone, « A law of comparative judgment. », *in*: *Psychological Review* 34 (1994), pp. 273–286 (cit. on p. 23).

[108] Dong Tian et al., « Geometric distortion metrics for point cloud compression », *in*: *IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 3460–3464, DOI: `10.1109/ICIP.2017.8296925` (cit. on pp. 94, 131).

[109] Shishun Tian et al., « NIQSV: A no reference image quality assessment metric for 3D synthesized views », *in*: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 1248–1252 (cit. on p. 131).

[110] Shishun Tian et al., « NIQSV+: A no-reference synthesized view quality assessment metric », *in*: *IEEE Transactions on Image Processing* 27.*4* (2017), pp. 1652–1664 (cit. on p. 131).

[111] Yu Tian et al., « A Light Field Image Quality Assessment Model Based on Symmetry and Depth Features », *in*: *IEEE Transactions on Circuits and Systems for Video Technology* 31.*5* (2021), pp. 2046–2050, DOI: `10.1109/TCSVT.2020.2971256` (cit. on p. 94).

[112] Kristi Tsukida and Maya R. Gupta, « How to Analyze Paired Comparison Data », *in*: 2011 (cit. on p. 23).

[113] Z. Tu et al., « A Comparative Evaluation Of Temporal Pooling Methods For Blind Video Quality Assessment », *in*: *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 141–145, DOI: `10.1109/ICIP40778.2020.9191169` (cit. on pp. 128–130).

[114] Michal Šorel and Filip Šroubek, « Fast Convolutional Sparse Coding Using Matrix Inversion Lemma », *in*: *Digit. Signal Process.* 55.*C* (Aug. 2016), pp. 44–51, ISSN: 1051-2004, DOI: `10.1016/j.dsp.2016.04.012` (cit. on p. 121).

[115] Haiqiang Wang et al., « MCL-JCV: A JND-based H.264/AVC video quality assessment dataset », *in*: *2016 IEEE International Conference on Image Processing (ICIP)* (2016), pp. 1509–1513 (cit. on p. 100).

[116] Haiqiang Wang et al., « VideoSet: A Large-Scale Compressed Video Quality Dataset Based on JND Measurement », *in*: *CoRR* abs/1701.01500 (2017), arXiv: `1701.01500`, URL: `http://arxiv.org/abs/1701.01500` (cit. on p. 100).

[117] Z. Wang, Eero Simoncelli, and Alan Bovik, « Multiscale structural similarity for image quality assessment », *in*: vol. 2, Dec. 2003, 1398 –1402 Vol.2, ISBN: 0-7803-8104-1, DOI: `10.1109/ACSSC.2003.1292216` (cit. on pp. 93, 122).

[118] Zhou Wang and Qiang Li, « Information Content Weighting for Perceptual Image Quality Assessment », *in*: *IEEE Transactions on Image Processing* 20.*5* (2011), pp. 1185–1198, DOI: `10.1109/TIP.2010.2092435` (cit. on p. 122).

[119] Zhou Wang et al., « Image quality assessment: from error visibility to structural similarity », *in*: *IEEE transactions on image processing* 13.*4* (2004), pp. 600–612 (cit. on pp. 92, 122, 131).

[120] Stefan Winkler, « Analysis of Public Image and Video Databases for Quality Assessment », *in*: *IEEE Journal of Selected Topics in Signal Processing* 6.*6* (2012), pp. 616–625, DOI: `10.1109/JSTSP.2012.2215007` (cit. on p. 22).

[121] Wufeng Xue et al., « Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index », *in*: *IEEE Transactions on Image Processing* 23.*2* (2014), pp. 684–695, DOI: `10.1109/TIP.2013.2293423` (cit. on p. 117).

[122] H. Yeganeh and Z. Wang, « Objective Quality Assessment of Tone-Mapped Images », *in*: *IEEE Transactions on Image Processing* () (cit. on pp. 38, 40, 61, 68).

[123] Akiko Yoshida et al., « Perceptual evaluation of tone mapping operators with real-world scenes », *in*: *Human Vision and Electronic Imaging X*, ed. by Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Scott J. Daly, vol. 5666, International Society for Optics and Photonics, SPIE, 2005, pp. 192 –203, DOI: `10.1117/12.587782` (cit. on p. 31).

[124] Honghai Yu and Stefan Winkler, « Image complexity and spatial information », *in*: *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2013, pp. 12–17, DOI: `10.1109/QoMEX.2013.6603194` (cit. on p. 22).

[125] Emin Zerman et al., « Textured mesh vs coloured point cloud: A subjective study for volumetric video compression », *in*: *12th International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2020 (cit. on pp. 125, 126, 128, 132).

[126] Emin Zerman et al., « The Relation Between MOS and Pairwise Comparisons and the Importance of Cross-Content Comparisons », *in*: *Human Vision and Electronic Imaging 2018, Burlingame, CA, USA*, 2018, DOI: `10.2352/ISSN.2470-1173.2018.14.HVEI-517` (cit. on pp. 68, 72).

[127] Guangtao Zhai and Xiongkuo Min, « Perceptual image quality assessment: a survey », *in*: *Science China Information Sciences* 63 (2020), pp. 1–52 (cit. on p. 92).

[128] Jianming Zhang et al., « Minimum Barrier Salient Object Detection at 80 FPS », *in*: *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1404–1412, DOI: `10.1109/ICCV.2015.165` (cit. on p. 38).

[129] Lin Zhang et al., « FSIM: A Feature Similarity Index for Image Quality Assessment », *in*: *IEEE Transactions on Image Processing* 20.*8* (2011), pp. 2378–2386, DOI: `10.1109/TIP.2011.2109730` (cit. on pp. 93, 122).

[130] Wei Zhang, Ralph R. Martin, and Hantao Liu, « A Saliency Dispersion Measure for Improving Saliency-Based Image Quality Metrics », *in*: *IEEE Transactions on Circuits and Systems for Video Technology* 28.*6* (2018), pp. 1462–1466, DOI: `10.1109/TCSVT.2017.2650910` (cit. on p. 38).

[131] Martin Čadík et al., « Evaluation of HDR tone mapping methods using essential perceptual attributes », *in*: *Computers Graphics* 32.*3* (2008), pp. 330–349, ISSN: 0097-8493, DOI: `https://doi.org/10.1016/j.cag.2008.04.003` (cit. on p. 31).

DOCTORAT
BRETAGNE
LOIRE / MATHSTIC

Nantes
Université

Titre : Évaluation de la qualité perceptuelle de contenus multimédias immersifs : HDR, champs lumineux et vidéos volumétriques.

Mot clés : Évaluation de la qualité, médias immersifs, mappage ton local, champs lumineux, vidéo volumétrique

Résumé : Des formats multimédias immersifs ont émergé comme un puissant canevas dans de nombreuses disciplines pour offrir une expérience utilisateur hyperréaliste. Ils peuvent prendre de nombreuses formes, telles que des images HDR, des champs lumineux, des nuages de points et des vidéos volumétriques. L'objectif de cette thèse est de proposer de nouvelles méthodologies pour l'évaluation de la qualité de tels contenus. La première partie de la thèse porte sur l'évaluation subjective de la qualité d'image. Plus précisément, nous proposons une stratégie de sélection de contenu et d'observateurs, ainsi qu'une analyse approfondie de la fiabilité des plateformes de crowdsourcing pour collecter des données subjectives à grande échelle. Nos résultats montrent une amélioration de la fiabilité des annotations subjectives collectées et répondent aux exigences liées en crowdsourcing à la reproduction d'expériences menés en laboratoire. La deuxième partie contribue à l'évaluation objective de la qualité avec une métrique de qualité d'image basée sur l'apprentissage automatique utilisant les informations de seuil de discrimination, et une métrique de qualité d'image pour les champs lumineux sans référence basée sur des représentations d'images planes épipolaires. Enfin, nous étudions l'impact des méthodologies d'agrégation temporel sur les performances des métriques de qualité objective pour les vidéos volumétriques. Dans l'ensemble, nous démontrons comment nos résultats peuvent être utilisés pour améliorer l'optimisation des outils de traitement pour les contenus multimédias immersifs.

Title: Perceptual quality evaluation of immersive multimedia content: HDR, Light Field and Volumetric Video

Keywords: Quality evaluation, immersive media, tone mapped images, light fields, volumetric video

Abstract: Immersive multimedia formats emerged as a powerful canvas in numerous disciplines for delivering hyper-realistic user experience. They can take many forms, such as HDR images, Light Fields, Point Clouds, and Volumetric Videos. The goal of this thesis is to propose novel methodologies for the quality assessment of such multimedia content. The first part of the thesis focuses on subjective image quality assessment. More specifically, we propose a content selection strategy, observer screening tools, and an extensive analysis on the reliability of crowdsourcing platforms to produce a large-scale dataset. Our findings improve the reliability of the collected subjective annotations and address issues to transfer laboratory experiments into crowdsourcing. The second part contributes to the objective quality evaluation with a learning-based image quality metric utilizing the just noticeable difference information and a no-reference light field image quality metric based on epipolar plane image representations. Finally, we investigate the impact of temporal pooling methodologies in objective quality metric performances for volumetric videos. Overall, we demonstrate how our findings can be used to improve the optimization of processing tools for immersive multimedia content.