Visual Quality of Experience: A Metric-Driven Perspective

Siahaan, Ernestasia

DOI
10.4233/uuid:d0a8f1b0-d829-4a34-be5a-1ff7aa8679ca

Publication date
2018

Document Version
Final published version

Citation (APA)
Siahaan, E. (2018). *Visual Quality of Experience: A Metric-Driven Perspective*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:d0a8f1b0-d829-4a34-be5a-1ff7aa8679ca

VISUAL QUALITY OF EXPERIENCE

A METRIC-DRIVEN PERSPECTIVE

VISUAL QUALITY OF EXPERIENCE

A METRIC-DRIVEN PERSPECTIVE

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus, prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates,
to be defended publicly on Tuesday 16 October 2018 at 10:00 o'clock

by

Ernestasia SIAHAAN

Master of Science in Computer Science and Information Engineering,
National Central University, Zhongli, Taiwan

born in Medan, Indonesia.

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. A. Hanjalic,	Delft University of Technology, promotor
Dr. J. A. Redi,	Delft University of Technology, copromotor

Independent members:

Prof. dr. P. le Callet,	Polytech Nantes/University of Nantes, France
Prof. dr.-Ing. A. Raake,	Ilmenau University of Technology, Germany
Prof. dr. I. E. J. Heynderickx,	Eindhoven University of Technology, The Netherlands
Prof. dr. H. de Ridder,	Delft University of Technology
Prof. dr. C. M. Jonker,	Delft University of Technology

Keywords: Quality of Experience (QoE), image quality metrics, subjective methodologies

Printed by: Ipskamp Printing

ISBN/EAN: 978-94-028-1188-9

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

To O. Tiarma Doli & Boru.

CONTENTS

SUMMARY

Multimedia systems are typically optimized in a way that maximizes users' satisfaction of using the systems/services. This user satisfaction is what is commonly referred to as *Quality of Experience (QoE)*. For visual media, such as images and videos, the optimization of QoE has meant reducing the visibility of artifacts (e.g. noise or other disturbing factors) in the visual media. This is based on the assumption that the sole appearance of artifacts would disrupt the whole visual experience, in a world where media were mostly consumed passively, and in well defined contexts (e.g., TV broadcasts). Nowadays, the way users experience visual media has changed, thanks to the diffusion of mobile, interactive, immersive, and on-demand technology. Media are now consumed in many different contexts, for example, in the interactive and customizable contexts of social media, or in the immersive contexts of virtual and augmented reality. As consequence of these developments, a user's visual QoE is no longer determined solely on the appearance of artifacts, but also by factors relevant to the viewing context.

This thesis brings in new insights in modeling and automatically assessing users' visual QoE in view of the developments above. The thesis starts with looking into subjective methodologies for QoE assessments, and continues with developing objective quality metrics that incorporate QoE influencing factors to improve state-of-the-art metrics.

Developing reliable and accurate objective metrics to automatically assess users' visual QoE requires subjective data that are reliable as well. This thesis argues that existing methodologies for collecting subjective data might not be reliable when used to evaluate QoE factors that are highly subjective, or that are new to the research community. Highly subjective quantities may yield different conclusions across experiments. As for new types of media, they often bring the uncertainty on how to evaluate them. Two studies are then presented in this regard. The first study considers the assessment of image aesthetic appeal, as one example of a *highly subjective* quantity. A large scale study was conducted to compare the use of different subjective methodologies to collect aesthetic appeal data, and some ways to measure the data reliably were proposed. The second study considers the assessment of point cloud quality, as one example of a new type of media (i.e. immersive media). The study explores quantitative and qualitative approaches to understand the way users judge point cloud images.

Following the studies on subjective QoE assessments, two studies on objective QoE

metrics are presented in this thesis. Despite existing efforts to model the influence of different factors on visual QoE, limited work have proposed to incorporate these factors into existing objective quality metrics to improve state-of-the-art. The first study on objective QoE metrics in this thesis investigates the influence of image content/semantic categories (i.e. scene and object categories) on visual QoE, and proposes to include semantic category features in objective image quality metrics. The proposed approach shows improvement from state-of-the-art in predicting image quality. The next study on objective quality metrics investigates new QoE influencing factors for point cloud images, and proposes to incorporate these into an objective quality metric for point cloud images.

The results of the studies presented in this thesis show how existing subjective methodologies could yield reliable aesthetic appeal data, and explore point cloud QoE influencing factors. Moreover, the results show that incorporating new QoE influencing factors into objective image quality metrics could improve state-of-the-art performance in predicting users' QoE. At the end of this thesis, some recommendations are given for future research following up the findings in this thesis.

SAMENVATTING

Multimediasystemen zijn normaal gezien in een manier geoptimaliseerd wat voor een gemaximaliseerde gebruikerstevredenheid zorgt wanneer de systemen/diensten gebruikt worden. Deze gebruikerstevredenheid wordt meestal aangeduid als een *Quality of Experience* (QoE). Voor visuele media, zoals beelden en video's, betekent de optimalisatie van QoE het verminderen van de zichtbaarheid van artefacten (e.g. lawaai of andere storende factoren) in visuele media. Dit is gebaseerd op de aanname dat alleen het verschijnen van artefacten de hele visuele ervaring zou verstoren, in een wereld waar media vooral passief wordt geconsumeerd, en in goed gedefinieerde contexten (e.g. tv-uitzendingen). Tegenwoordig is de manier waarop gebruikers visuele media ervaren veranderd, dankzij de verspreiding van de mobiele telefoon, interactieve, immersieve en on-demand technologie. Media wordt nu verbruikt in vele verschillende contexten, bijvoorbeeld in interactieve en aanpasbare contexten van social media, of in de immersieve contexten van *Virtual* en *Augmented Reality*. Als een gevolg van deze ontwikkelingen, is de visuele QoE van een gebruiker niet meer enkel gebaseerd op de verschijning van artefacten, maar ook door factoren die relevant zijn tot de kijkcontext.

Dit proefschrift leveert nieuwe inzichten in het modeleren en het automatisch beoordelen van de visuele QoE van gebruikers gezien de ontwikkelingen hierboven. Het proefschrift begint met het onderzoeken van subjectieve methodologieën voor het beoordelen van QoE, en gaat verder met het ontwikkelen van objectieve kwaliteitsmetrieken die QoE-beïnvloedende factoren bevatten om de state-of-the-art (meest geavanceerde meetgegevens) te verbeteren.

Het ontwikkelen van betrouwbare en nauwkeurige objectieve metrics om automatisch visuele QoE van gebruikers te beoordelen vereist subjectieve data dat ook betrouwbaar moet zijn. Dit proefschrift beweert dat de bestaande methodologieën voor het verzamelen van subjectieve data wellicht niet betrouwbaar is wanneer het gebruikt wordt om zeer subjectieve QoE factoren te evalueren, of die nieuw zijn voor de onderzoeksgemeenschap. Zeer subjectieve kwantiteiten kunnen verschillende conclusies opbrengen over verschillende experimenten. Wat betreft de nieuwe types van media, vaak brengen die een onzekerheid over hoe ze geëvalueerd moeten worden. In dit verband worden er twee studies gepresenteerd. De eerste studie beschouwt de beoordeling van image

aesthetic appeal als een voorbeeld van een *zeer subjectieve* kwantiteit. Een grootschalige studie was uitgevoerd om het gebruik van verschillende subjectieve methodologieën die image aesthetic appeal (esthetische aantrekkingskracht) verzamelen te vergelijken. Naar aanleiding daarvan werden er een aantal manieren om de data betrouwbaar te meten voorgesteld. De tweede studie beschouwt de beoordeling van point cloud QoE als een voorbeeld van een nieuw type media (i.e. immersieve media). De studie onderzoekt kwantitatieve en kwalitatieve benaderingen om te begrijpen hoe gebruikers point cloud images beoordelen.

Na het presenteren van de studies over subjectieve QoE beoordelingen, worden er twee studies over objectieve QoE metrics (metriek) voorgesteld in dit proefschrift. Ondanks bestaande inspanningen om een model te maken van de invloed van verschillende factoren op visuele QoE, is er gelimiteerd werk die voorgesteld heeft om deze factoren te incorporeren in een bestaan objectieve kwaliteitsmetrics om de state-of-the-art metrics te verbeteren. De eerste studie over objectieve QoE metrics in deze studie kijkt naar de invloed van beeldcontent/semantische categorieën (i.e. scène en object categorieën) op visuele QoE, en stelt voor om de semantische categorie kenmerken te incorporeren in objectieve beeldkwaliteitsmetrics. The voorgestelde benadering toont een verbetering van de state-of-the-art in het voorspellen van beeldkwaliteit. De volgende studie over objectieve kwaliteitsmetrics onderzoekt nieuwe QoE factoren die invloed hebben op point cloud beelden en stelt voor deze te incorporeren in een objectieve kwaliteitsmetric voor point cloud beelden.

De resultaten van de studies gepresenteerd in dit proefschrift tonen hoe bestaande subjectieve methodologieën betrouwbare image aesthetic appeal data kunnen opbrengen, en onderzoeken point cloud QoE beïnvloedende factoren. Bovendien tonen de resultaten dat het incorporeren van nieuwe QoE beïnvloedende factoren in objectieve beeldkwaliteitsmetrics, de uitvoering van state-of-the-art kunnen verbeteren in het voorspelen van de QoE van gebruikers. Tot slot geeft dit proefschrift een aantal aanbevelingen voor toekomstig onderzoek naar aanleiding van de resultaten uit dit onderzoek.

I

PRELUDE

1

INTRODUCTION

Tak kenal maka tak sayang.
(English transl.: Ignorance leads to indifference.)

Indonesian Proverb

1

Multimedia systems have become an integral part of human life. Nowadays, most of our waking hours are spent using different multimedia systems, ranging from web applications to streaming services. Even when asleep, we sometimes rely on multimedia systems to perform tasks for us, such as surveillance cameras to control our home, or smart watches to monitor our sleeping conditions.

As more companies compete to deliver multimedia services to users, market research has shown that users have become more demanding of what they get in return for their money [1, 2, 3]. A consumer review from Deloitte in 2014 suggests that businesses in general are struggling to keep up with users'/ consumers' ever growing expectations [1]. For digital services, specifically, users' high expectations have led to less loyalty to service providers. Conviva, in its 2015 Consumer Report, shows that one in three over-the-top (OTT) users abandons its current provider when experiencing a lower quality in the service [2]. According to Ofcom, the percentage of UK consumers switching TV or mobile services in the year 2015 almost doubled that of 2014 [3].

To keep up with users' expectations, it is therefore essential that multimedia systems are developed such that their optimization is targeted towards maximizing users' satisfaction of using the service. This user satisfaction is what is commonly referred to as *Quality of Experience* (QoE). Formally, QoE is defined as "the degree of delight or annoyance of the user of an application or service" [4]. Further, this delight or annoyance comes from the fulfillment of users' expectations of the utility and enjoyment provided by the application or service, as influenced by users' personality and current state.

In this thesis, we focus on multimedia systems and services built around visual media (i.e., images and video) consumption. This is a critical range of systems, as visual media consumption is responsible for more than 70% of internet traffic in the world, according to the Cisco Visual Networking Index 2017 [5], and it will continue to take more traffic in the future. These systems include video services, such as virtual and augmented reality, video-on-demand, visual surveillance, or mobile video services, but also still images (photos): around 3.9 trillion images were taken in 2016 [6], and at least 1.8 billion are uploaded to various online platforms every day [7]. This thesis focuses on the optimization of these systems by targeting the assessment of Visual Quality of Experience (Visual QoE) as a measure of user satisfaction.

For a long time, optimizing Visual QoE meant reducing the visibility of artifacts (e.g. noise or other disturbing factors) that occur due to technology limitations [8] in the visual media. This view was based on the assumption that the sole appearance of artifacts would disrupt the whole visual experience, in a world where media were mostly consumed passively, and in well defined contexts (e.g., TV broadcasts).

Nowadays, the diffusion of mobile, interactive, immersive, on demand media technology has changed the way users experience media. Media are now consumed in many different contexts, for example social, interactive and customizable contexts (e.g. social media, mobile), or immersive contexts (e.g. virtual and augmented reality). Different contexts bring in additional factors that influence the QoE elicited by a visual medium. Some studies have shown, for example, that visual QoE is also influenced by the type of device through which they access multimedia [9, 10], or the content type of the media [9, 11]. In other studies, the aesthetic appeal of a visual medium was shown to influence user perception of its overall quality [12]. In other words, the visibility of an artifact alone is not the sole factor determining the user satisfaction with media experiences anymore [13].

This thesis brings in new insights in modeling and assessing Visual QoE in view of the developments sketched above. Due to the complexity of the QoE problem in the broad domain of visual media, we will focus on multimedia systems and services handling still images only. It is worthwhile noting that, although addressing image quality, the results contributed in this thesis can be extended to video quality as well. Image quality estimation is a stepping stone in the development of reliable video quality metrics, where the temporal dimension is also taken into account [14, 15]. As the contribution in this thesis touches upon fundamental issues of visual QoE, the results may also be applied to video QoE assessment.

In the remainder of this chapter, we first describe the general problem of visual quality assessment in Section 1.1. In Section 1.2, we then continue with reviewing the traditional approach to tackle the problem (i.e. visual quality estimation based on artifact visibility). Afterwards, we move on to an overview of Visual QoE estimation beyond artifact visibility, and provide a glimpse of existing research on the topic (Section 1.3), which will guide us to formulating the research questions addressed in this thesis (Section 1.4). The chapter concludes with an overview of the thesis contributions in Section 1.5. In the remaining parts of this thesis, we will use the term QoE and visual QoE interchangeably.

1.1. THE PROBLEM OF VISUAL QOE ASSESSMENT

Visual media (images and videos) go through various steps in their life cycle (see Figure 1.1), that can tamper with their quality [8]. It is important to review the most common steps, along with the type of artifacts they generate, in order to better contextualize the role of visual quality assessment systems in the process of visual QoE optimization.

Acquisition. During acquisition or capture of visual media, problems may occur due

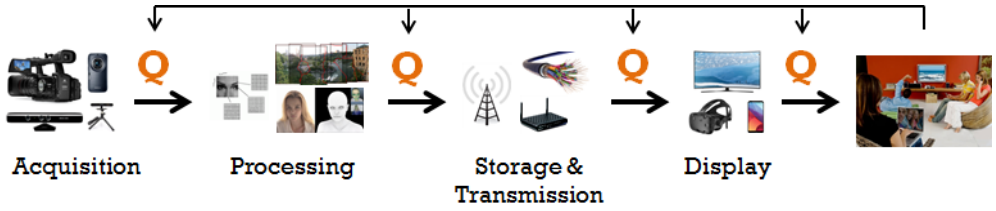


Figure 1.1: Visual media life cycle, from acquisition to display. In each step, quality degradation may occur, and a quality estimator (Q in the figure) may be used to decide on an appropriate enhancement or optimization.

to movement of camera, or movement of the object that is being captured. One example of artifacts that may occur in this case is blur [8].Another acquisition problem is the limitation of sensing technology in the camera. For example, a camera's A/D converter may have an improper sampling rate during capture, in which case aliasing artifacts would occur. As another example, inaccurate camera parameters when using multiple cameras to capture a 3D object may create noise in the captured image.

Processing. Captured images and videos are often processed before being stored and delivered to the final user. Among the processing algorithms, the most typical ones are compression (encoding and decoding), enhancement, or combining multiple images (to create a pano- ramic image, or construct a 3D image). Artifacts introduced at this stage are typically caused by the approach used in the processing algorithms in transforming the images. For example, compression using block-based codecs typically results in block artifacts, while wavelet-based codecs result in ringing or aliasing artifacts [16, 17, 18]. Meanwhile, automatic image stitching to create panoramic images may introduce parallax errors [19], while depth triangulation in constructing 3D images may create geometry noise [20].

Storage and transmission. When storing or transmitting (e.g., streaming) visual media, loss of information often occurs due to constraints in the network, or fault in device. Example of artifacts that occur at this stage include packet loss or delay in videos [21, 22].

Display. Artifacts during display usually occur due to the need to adapt an image or video to a particular device setup. Examples of this process include re-sampling, transcoding or tone-mapping [23, 24, 25]. Artifacts may also occur due to the display technology itself. For example, liquid crystal display (LCD)'s slow temporal response and hold-type LCD rendering method creates motion blur artifacts [26].

As indicated above, in each step of the image lifecycle, one or multiple systems or algorithms need to be optimized in order to minimize the losses in visual QoE. This is where a *quality estimator* or *quality metric* (denoted as Q in Figure 1.1) comes into play:

measuring the quality of the media and feeding it back to the system for quality control, adaptation or enhancement. The goal of visual QoE studies is to build better quality estimators or metrics for either stage of the image life cycle. Traditionally, the approach taken to devise these quality estimators has been to model the visibility of artifacts in the media. However, as explained in the beginning of this chapter, this approach no longer fully represents users' perception of visual QoE. In the next two sections, we will look into these different approaches to building visual QoE metrics, focusing especially on image QoE, being the focus of this thesis.

1.2. VISUAL QUALITY BASED ON ARTIFACT VISIBILITY

Traditionally, visual QoE has been estimated in terms of artifact visibility, i.e. by designing quality metrics that could estimate the extent to which artifacts in the image would be visible to the human eye [8]. A large body of extremely valuable work has been produced in the past two decades [14, 15, 27, 28], which we will be briefly review as it represents the basis for modern QoE assessment.

Visual Quality metrics are usually categorized based on the availability of a reference, pristine signal (image or video) to compare the processed signal with. When there is no reference signal needed in the metric, the metric is referred to as a no-reference (NR) quality metric. A reduced-reference (RR) metric uses extracted features from the reference signal, and compares it with the corresponding features of the processed signal. Lastly, a full-reference (FR) metric needs the complete reference signal to assess the processed signal's quality.

A quality metric typically follows three steps in estimating image quality [8]. The first step is measuring properties or features of the image that could be used to estimate its quality. These features could be the direct value of the image pixels, some image statistics (e.g. distribution of the image luminance values), or a model of distortions or artifacts in the image (e.g. a measure of texture masking or luminance masking to estimate blockiness). Afterwards, a pooling step is performed to combine the features into an appropriate scale. For example, if the image features have been calculated locally (on partitions of the image), the pooling step would aggregate the local features into a global feature value. Finally, features are mapped into quality scores. One way to do this is to train a machine learning algorithm to map pooled features into quality scores based on ground truth data.

Over the years, various image quality metrics have been proposed. Traditional metrics are typically full-reference, and measure signal fidelity, i.e. pixel-by-pixel differences

between the processed image and its reference. These metrics include mean square error (MSE), peak signal-to-noise ratio (PSNR), and other similar metrics [29]. Another type of metrics attempts to model quality based on the properties of the human visual system (HVS), for example by modeling the contrast sensitivity function (CSF), or the luminance adaptation [30, 31]. Using these properties, the metrics aim at estimating the HVS response to different artifacts in an image, and subsequently the image quality. Finally, some metrics estimates quality by analyzing properties of the processed image's signal. We will refer to this category of metrics as signal-driven metrics [27]. The Structural Similarity (SSIM) Index [32] is a well-known full-reference metric that belongs in this category. The metric performs a comparison of local luminance, contrast and structure measurements of the degraded and reference images. Another well-known example of signal-driven metrics are natural scene statistics (NSS)-based metrics [33, 34, 35]. The idea behind NSS-based metrics is that the signal (e.g. luminance values) of a pristine, high-quality image, follows a certain distribution. The presence of artifacts typically alters the shape of this distribution, and so the image quality can be estimated based on this principle. Artifact-specific metrics usually also belong to the signal-driven metrics category [16, 17, 36], and so are the more recently proposed deep learning-based metrics [37, 38].

1.3. VISUAL QUALITY OF EXPERIENCE (QOE)

In this section, we will provide a brief overview of the past research on Visual QoE beyond artifact visibility (referred simply as Visual QoE or QoE from here on). We will start in Section 1.3.1. with work related to objectively modeling visual QoE, and afterwards, section 1.3.2 will focus on the challenge of subjectivity in assessing visual QoE. At the end of both subsections, we describe the challenges and open problems related to image QoE research that we address in this thesis.

1.3.1. VISUAL QOE: OBJECTIVE QUALITY BEYOND ARTIFACT VISIBILITY

The way visual media are produced and consumed have changed and diversified considerably in the past years, thanks to various technology developments. Acquisition devices have become more ubiquitous and complex: high resolution cameras are widely available even to naive users, and more immersive images could be captured using plenoptic cameras and 3D scanners. Images and videos are not only consumed passively on TV or through prints, but in more interactive and social contexts such as through social media or on demand applications. More advanced display technologies, such as 3D screens

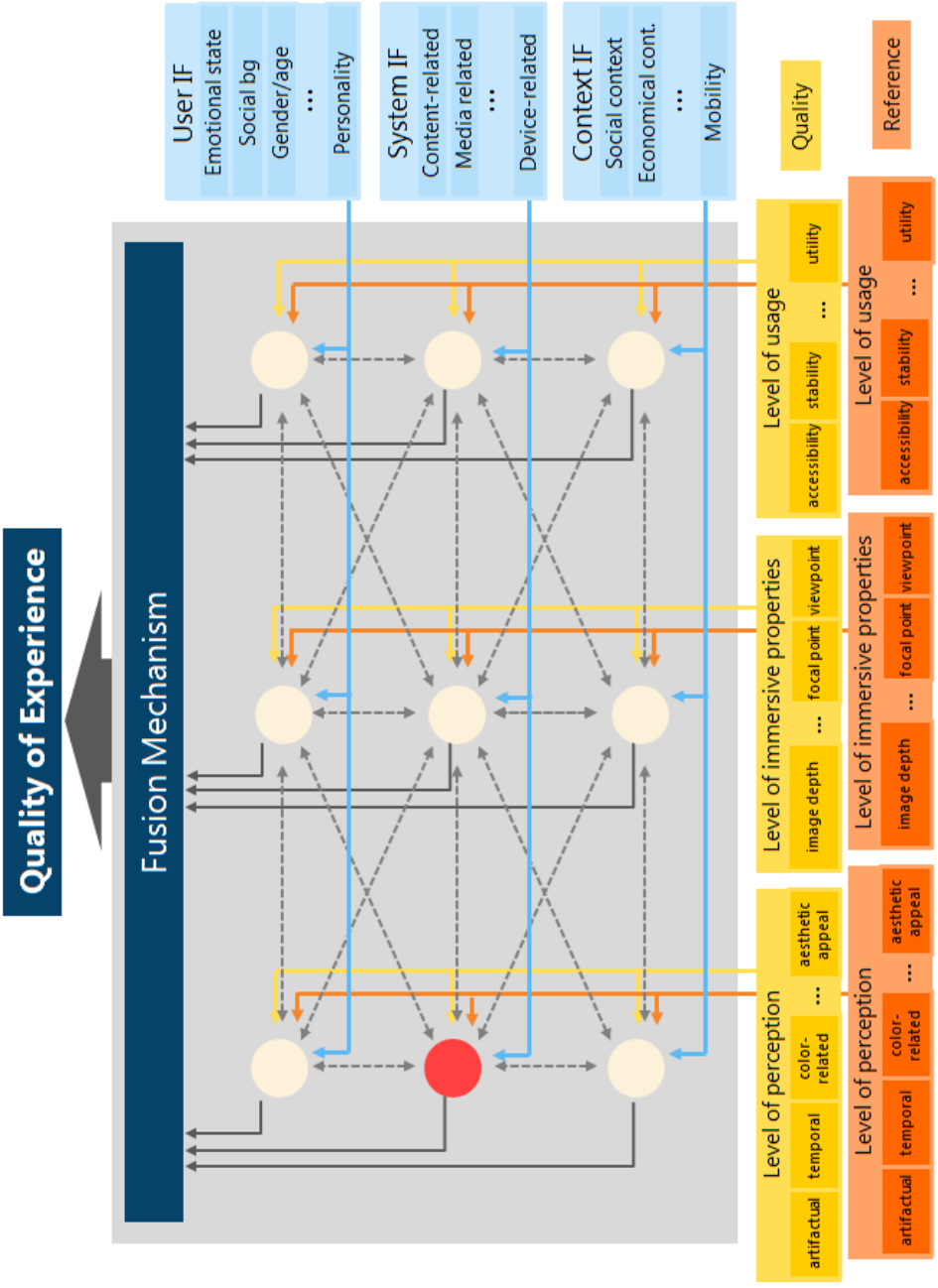


Figure 1.2: Schematic representation of the Quality of Experience (QoE) as presented in [12].

and head mounted displays (HMDs), have also allowed for more immersive viewing contexts. Consequently, visual QoE assessment can no longer rely on visual signal alone, but needs to take into account additional factors that may arise due to the different viewing contexts. Studies on visual QoE have looked into uncovering these factors and understanding how these factors interact in forming user QoE. We describe the factors that have been studied in previous research below.

Physical environment or viewing environment. Some studies have looked into the effect of viewing context on users' perception of quality. For example, Xue et al. showed that mobile viewing affects user QoE due to contextual noise masking the signal noise [39]. Another example comes from the work of Zhu et al. [40, 41], showing that having other people viewing content together has an effect on user's enjoyment, which in turn positively influences their overall perception of QoE.

Media content. Different studies have also proposed image and video content as an influencing factor of QoE, i.e. users' ability to apprehend an image or video's content affects their overall perception of the image/video quality. Pereira [42] included knowledge or information acquired in his proposed model for evaluating QoE of videos. Similarly, the term usefulness of an image is used in the image quality model by De Ridder and Endrikhovski [43], suggesting the importance of users being able to extract information from images. Results of experiments by Alers et al. [44] showed the importance of having clear image region of interest (ROI), further presenting the influence of content on user QoE.

Affective attributes. Image aesthetics, interestingness, or likability are often considered important attributes in viewing contexts such as social media or image recommendation applications [45, 46, 47]. Some studies in image QoE have looked into how these attributes influence overall image quality perception. In her user experiment, Redi [12] found that users are highly critical of the technical quality (i.e., annoyance brought about by artifacts) of images with high aesthetic appeal, but less so for images with low aesthetic appeal. Halonen et al. [48] suggested the importance of balancing the naturalness and interestingness attributes of images used in subjective quality evaluations, and performed a study to understand how users rate these attributes in images.

Immersive properties. With new acquisition and display technology in recent years, various types of new media have emerged that offer more immersive viewing experiences, such as stereoscopic images/videos, light fields, omnidirectional content, or point clouds. Immersive viewing experiences refer to viewing experiences that create a sense of presence (immersion) for users, thanks to features such as depth, or multiple degrees of freedom. Various studies have looked into factors influencing user QoE when viewing

this type of media. For example, image depth perception was found to be an influencing factor to users' overall viewing experience when watching content on 3D displays [49]. Another study showed the effect of different head-mounted displays (HDM) and motion sickness on the QoE of watching omnidirectional videos [50]. They compared two popular HMDs in the market, and found that one offers higher integral quality than the other. Garau et al. studied the effect of avatar realism and eye-gaze control on users' QoE in shared virtual environments, and showed that inferred eye animations have significant positive effect on users' overall QoE. [51].

Most work related to influencing factors of visual QoE, as described above, have pointed out the effect of previously neglected factors on QoE, but have not proposed an actual way of incorporating this effect in the existing quality metrics to build better quality predictors. Zhu et al. [41], after showing the influence of user demographics on video QoE, have proposed a demographics-based metric to automatically assess the perceived quality of individual users. This approach is useful especially for personalized systems which have user background information to use as features. However, there are a lot of application contexts that cannot rely on such information. Moreover, in some cases, it is still desirable to have an automatic assessment or estimation of an average quality value across users, instead of individual scores for each user.

To illustrate this research gap, we build upon the schematic representation of QoE as proposed by Redi [12] and shown in Figure 1.2. The figure shows how the eventual quality of experience depends on technical factors/features that refer to the media or system that the user is viewing or interacting with (orange and yellow-colored boxes in the figure), and contextual influencing factors which are the conditions of the user, environment or system that pre-exist the viewing experience (light blue-colored boxes in the figure). Referring to image, the technical factors include, but are not limited to, the level of perceived artifacts, the aesthetic appeal of the image, or the viewpoint of the (immersive) media. The contextual influencing factors include, for example, users' emotional state, demographic background and device limitations.

Current work on QoE mostly look into the modeling of the relationship of these factors with each other (for example, image aesthetics and image artifacts [12], or user mobility and video bit rate [39]). However, only few [39] have proposed a computational integration of these factors to predict image quality (i.e. the circles inside the fusion mechanism box in Figure 1.2). Our contribution in this thesis is represented by the red circle in Figure 1.2. Not only do we uncover how different perceptual factors and system influencing factors influence image QoE, but we also incorporate these factors into an objective quality metric (Chapters 4 and 5 of this thesis).

To address the design of these more accurate metrics, not only do we need to design better algorithm that incorporate novel information types (e.g. about immersion or content), but we also, first and foremost, need to fully clarify which role these new factors play in QoE perception. To do so, the most reliable approach is based on empirical research in the so-called Subjective Quality Assessment sphere.

1.3.2. SUBJECTIVE QUALITY ASSESSMENTS IN VISUAL QOE MODELING

Subjective assessments are essential to understand users' perception of QoE changes along with change of different factors, as they directly involve humans in the assessment process. A branch of empirical research and psychometrics in particular, subjective quality assessment is typically based on experiments where visual experiences are manipulated by controlling for one or more factors of interest, and users are asked to evaluate the QoE of visual stimuli without being aware of the different factors that are being tested. Subjetive quality assessments are also the sole source of ground truth for the validation of quality metrics (see Section 1.2) [52, 53].

Typically, subjective assessments in QoE studies adopt standard/recommended experiment setups in controlled lab settings [52, 53]. However, more studies have begun to look into conducting subjective assessments in uncontrolled environments [54, 55], as this allows for reaching out to a wider, more diverse and therefore more representative user group. These studies usually compare the results with controlled lab-based studies, and some would recommend best-practices for conducting experiments in uncontrolled settings, like, for example, a crowdsourcing environment [56].

QoE studies often involve factors that, unlike image artifacts, are not always straightforward to measure. These factors include image aesthetics, immersive properties, interestingness, among others, and are often viewed as highly subjective factors. It is generally unknown whether or not standard methods for assessing perceptual (visual) quality could be directly adopted in experiments that measure these *highly subjective* factors. Some papers raise the concern of the reliability of the assessments collected, or whether or not current practices of subjective assessments are the correct approach to evaluate these factors [57, 58, 59]. It is important to address these concerns, as the quality of data collected through subjective assessments will affect the quality of metrics built with it. In this thesis, we contribute to resolving the above concerns for the subjective assessments (Chapters 2 and 3 of this thesis).

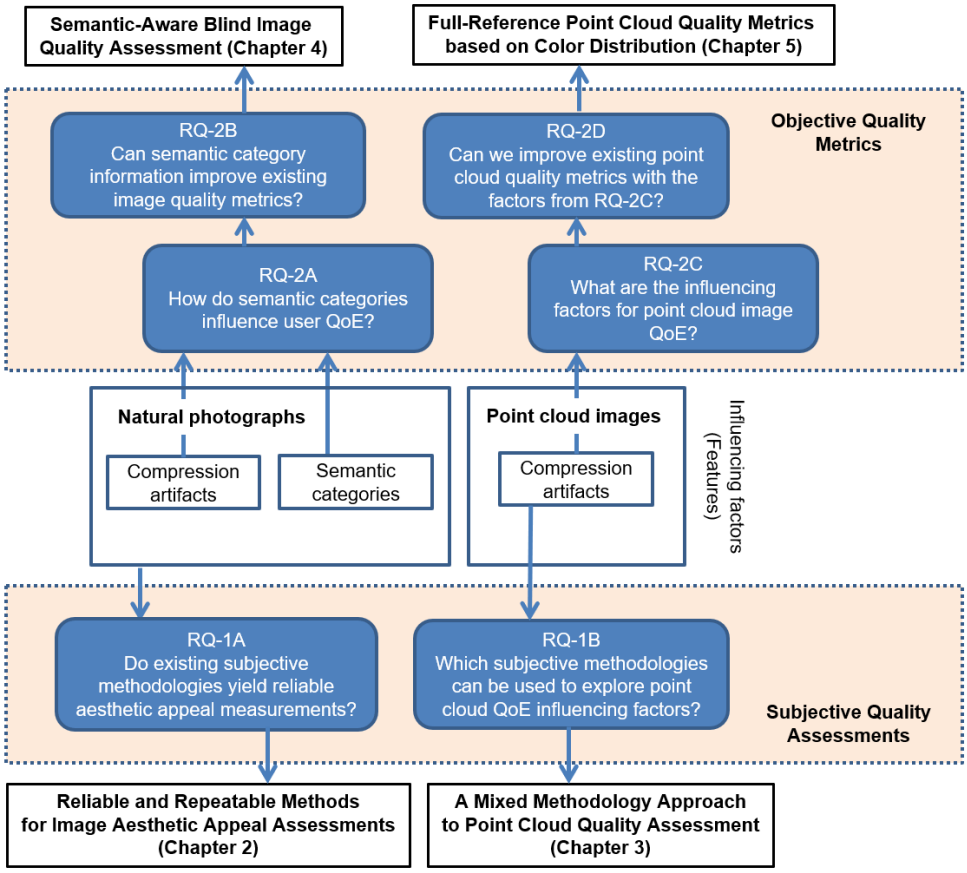


Figure 1.3: Scope and contribution of this thesis. The two orange boxes indicate the two main research scopes of this thesis. Each blue box represents one research questions addressed in this thesis. The main contribution of this thesis is indicated with black-bordered boxes.

1

1.4. RESEARCH PROBLEMS

Based on our survey of the state of the art for visual QoE in Section 1.3., we identify open research problems related to both the objective quality metrics and subjective quality assessments of visual QoE.

- **Research Scope 1: Subjective Quality Assessments.**

  Understanding the reliability of existing subjective methodologies in assessing highly subjective QoE factors and new types of media.

  Developing objective quality metrics requires subjective data. To build reliable and accurate quality metrics, we need to be certain that the subjective data we are using is also reliable. For this reason, standards and recommendations on how to perform subjective quality assessment exist. However, in certain cases, it is not always clear whether or not these standard methods for assessing perceptual (visual) quality could be directly adopted to other cases.

  One of such cases is the assessment of quantities that are considered *highly subjective* (e.g. affective factors: aesthetic appeal, interestingness). For example, some studies have shown concern on the reliability of scores collected for aesthetic appeal evaluations [57, 58], even showing that in some cases, different experiments on the same image set yield different conclusions. For this reason, we look into comparing the use of different subjective methodologies to collect aesthetic appeal data, and propose some ways to measure the data reliably.

  Another case is when the visual media being assessed is of a new type, for example, immersive media. Dealing with new types of media often brings the uncertainty on how to evaluate their quality. As we do not yet understand what users perceive in these types of media and how they form their judgment of these media, existing subjective methodologies may not be appropriate to evaluate these media.

  In this thesis, we address this problem for point cloud quality assessment. Point clouds have sparked interest as a promising technology for 3D representations in immersive systems. Studies on its objective quality assessment show that state-of-the-art metrics do not correlate well with human assessments [60, 20]. In response to these findings, none have looked into different subjective methodologies to fully explore how users perceive point cloud quality. We explore quantitative and qualitative approaches to understand the way users judge point cloud images, and present insights on performing subjective quality assessments of point cloud images.

Figure 1.3 shows the specific research questions under the Subjective Quality Assessments research scope. We also list the questions below:

RQ-1A. Do existing subjective methodologies yield reliable aesthetic appeal measurements?

RQ-1B. Which subjective methodologies can be used to explore point cloud QoE influencing factors?

- **Research Scope 2: Objective Quality Metrics.**

  Incorporating new influencing factors into an objective quality metric.

  Existing studies have looked into quantifying, using statistical models, how different factors influence user QoE of images and videos. However, very few have shown how to incorporate these factors to improve existing objective quality metrics.

  In this thesis, we aim to incorporate media content into image quality metrics to improve quality prediction. As explained in Section 1.3.1, various studies have suggested the influence of media content (e.g. acquired information, clarity of Region of Interest/ROI) on visual QoE. One aspect of media content that has not received attention in relation to visual QoE is semantics. Semantics refers to the meaning of words, phrases, or systems [1]. For images, semantics can refer to scene or object categories of an image, which are meaningful entities that people recognize to appear in them. Image semantics is also the first visual element fully disambiguated by a viewer [61, 62], hence playing a major role in visual experiences. This aspect is also hinted by the existing work on visual attention and image quality [44], showing how artifacts in visually important (and semantically rich) regions of the image provoke higher annoyance. In addition, from a computational point of view, work on automatic recognition of image semantics is well established, allowing us to easily incorporate image semantic information into quality metrics. For these reasons, we look into modeling of the influence of content (semantic) categories on user QoE, and using content category information to improve image quality prediction.

  Another aspect that is gaining a lot of attention in relation to media experiences is immersiveness, which relates to features, such as depth or multiple degrees of freedom. Many companies and researchers are looking into immersive media to develop virtual and augmented reality experiences. Still, many factors remain to

[1] Oxford dictionary, http://www.oxforddictionaries.com/

be understood on how immersive properties of visual media may influence user QoE. In this thesis, we focus on the QoE of point cloud images. Point cloud is an up-and-coming media type for 3D representations in immersive media, as it does not require expensive computational resources to manipulate it. Until now, not much work has looked into understanding the QoE of point clouds, and even less has looked into improving state-of-the-art quality metrics for point clouds. For this reason, we look into uncovering new influencing factors to consider in point cloud QoE, and incorporate them into an objective metric for point cloud images.

To summarize, the following are the two main problems related with our second research scope (objective quality metrics): investigating new influence factors and their relationship to QoE for images, and incorporating the above factors into new metrics that improve state-of-the-art objective quality metrics. We break these two main problems into four more specific research questions below (as also shown in Figure 1.3).

RQ-2A. How do semantic categories influence user QoE?

RQ-2B. Can semantic category information improve existing image quality metrics?

RQ-2C. What are the influencing factors for point cloud image QoE?

RQ-2D. Can we improve existing point cloud quality metrics with the factors from RQ-2C?

1.5. CONTRIBUTIONS OF THIS THESIS

We now describe the thesis contribution per chapter, including our experimental methodology, and provide the list of publications that served as the basis of this PhD thesis.

1.5.1. THESIS OUTLINE

This thesis consists of four main parts, Prelude, Methods, Metrics and Outlook.

PART I: PRELUDE

In this part, we include the necessary background information for readers to understand the scope of this thesis and its motivation. We introduce the topic of Visual Quality of Experience (QoE), and position the thesis with respect to the research area. We also give a review of the related work to give readers an idea of research that has been done, and what still needs to be done in the field.

PART II: METHODS

This part presents our contribution related to subjective assessment methodologies. We first look into the reliability of existing subjective methodologies in collecting user judgment of image aesthetic appeal. We also explore whether or not user judgment of aesthetic appeal could be repeated in different environments in which subjective assessments are setup. Our findings show that existing subjective methodologies could be used to reliably measure image aesthetic appeal, and that image aesthetic appeal assessments could be repeatable across different experimental environments. Another contribution presented in this part is on exploring the use of mixed methods to understand users' judgment of point cloud quality. Using a mixed method approach, we show that user perception of color fidelity plays an important role in their perception of point cloud QoE. This part includes chapters 2 and 3, which are based on the following publications:

1. **E. Siahaan**, A. Hanjalic, J.A. Redi, *A Reliable Methodology to Collect Ground Truth Data of Image Aesthetic Appeal*, IEEE Transactions on Multimedia **18(7)**, 1338 - 1350 (2016). **[Chapter 2]**

2. **E. Siahaan**, J.A. Redi, A. Hanjalic, P. Cesar, *Full-reference quality metrics for point cloud compression based on color distribution*, Under Review, (2018) **[Chapter 3]**

PART III: METRICS

This part presents our contribution related to building objective metrics for image quality prediction. Our first contribution in this part looks into using image semantic information (scene and object category information) as features to predict image quality. After showing the influence of semantic categories on user QoE, we incorporate semantic category features into no-reference image quality metrics and show that this approach outperforms the state-of-the-art. In the next chapter of this part, we propose a quality metric based on point cloud color distribution, and improve the state-of-the-art in point cloud quality assessment. This part includes Chapters 4 and 5, and is based on the following publications:

1. **E. Siahaan**, A. Hanjalic, J.A. Redi, *Semantic-aware blind image quality assessment*, Signal Processing: Image Communication **60**, 237-252 (2018). **[Chapter 4]**

2. **E. Siahaan**, J.A. Redi, A. Hanjalic, P. Cesar, *Full-reference quality metrics for point cloud compression based on color distribution*, Under Review, (2018) **[Chapter 5]**

PART IV: OUTLOOK

After presenting all our studies and findings in the previous parts, we present a discussion of their implications and provide recommendations for future research. We also conclude our thesis in this part. This part includes chapter 6.

Table 1.1: Overview of user studies/subjective assessments performed in each chapter of this thesis

Chapter Number	Research Question	Methodology	Environment	Number of Subjects
2	RQ-1A	Quantitative	Laboratory and Crowdsourcing (CS)	24 (lab) >360 (CS)
3	RQ-1B	Quantitative and Qualitative	Laboratory	22
4	RQ-2A and RQ2-B	Quantitative	Laboratory and Crowdsourcing	20 (lab) 337 (CS)
5	RQ-2C and RQ2-D	Quantitative	Laboratory	22

1.5.2. METHODOLOGY

To answer our research questions, we first perform subjective quality assessments to understand how the influencing factors influence user QoE. The scope of our subjective assessments is given in Table1.1. As shown in the table, we perform quantitative and qualitative studies, in controlled laboratory setups as well as crowdsourcing setups. Except for the assessments performed in Chapter 2 (RQ-1A), our subjective assessments are conducted to understand how certain factors influence user QoE. The collected data are then used to build new QoE metrics. In Chapter 2 (RQ-1A), our subjective assessments are conducted to compare the reliability of data collected using different subjective methodologies.

After conducting subjective assessments, we perform statistical analysis to obtain a statistical model of our subjective assessment results. Different types of analysis are performed in this thesis. In Chapter 2 (RQ-1A), we conduct a reliability analysis of different subjective methodologies. In the other chapters, we use Generalized Linear Mixed Models and Hierarchical Multiple Factor Analysis to model the relationship between the influence factors and user QoE. Our statistical analysis is performed using statistical tools available in Matlab, SPSS, and R.

Based on the subjective assessment and statistical analysis, we then propose new quality metrics. We do this in chapters 4 and 5. In Chapter 4, we propose no-reference image quality metrics, based on a machine learning approach. In Chapter 5, we propose a full-reference metric for point cloud images, where we estimate a point cloud image's quality based on the distance between selected features of the point cloud and its reference. To evaluate our quality metrics, we use the Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC) to evaluate the correlations of our predicted quality scores and the corresponding subjective assessments.

1.5.3. FULL LIST OF PUBLICATIONS

The following articles have been published in the process of completing this PhD thesis. Articles directly serving as chapters of this thesis are indicated accordingly.

8. **E. Siahaan**, J.A. Redi, A. Hanjalic, P. Cesar, *Full-reference quality metrics for point cloud compression based on color distribution*, Under Review, (2018) **[Chapters 3 and 5]**

7. **E. Siahaan**, A. Hanjalic, J.A. Redi, *Semantic-aware blind image quality assessment*, Signal Processing: Image Communication **60**, 237-252 (2018). **[Chapter 4]**

6. D. Martin, S. Carpendale, N. Gupta, T. Hoßfeld, B. Naderi, J. Redi, **E. Siahaan**, Ina Wechsung, *Understanding the Crowd: Ethical and Practical Matters in the Academic Use of Crowdsourcing*, Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments, 27-691 (2017).

5. **E. Siahaan**, A. Hanjalic, J.A. Redi, *Augmenting Blind Image Quality Assessment Using Image Semantics*, 2016 IEEE International Symposium on Multimedia (ISM), San Jose, USA, December 2016.

4. **E. Siahaan**, A. Hanjalic, J.A. Redi, *A Reliable Methodology to Collect Ground Truth Data of Image Aesthetic Appeal*, IEEE Transactions on Multimedia **18(7)**, 1338 - 1350 (2016). **[Chapter 2]**

3. **E. Siahaan**, A. Hanjalic, J.A. Redi, *Does visual quality depend on semantics? A study on the relationship between impairment annoyance and image semantics at early attentive stages*, Human Vision and Electronic Imaging 2016, San Fransisco, USA, February 2016.

2. J. Redi, **E. Siahaan**, P. Korshunov, J. Habigt, T. Hoßfeld, *When the Crowd Challenges the Lab: Lessons Learnt from Subjective Studies on Image Aesthetic Appeal*, Proceedings of the Fourth International Workshop on Crowdsourcing for Multimedia, Brisbane, Australia, October 2015.

1. **E. Siahaan**, J.A. Redi, A. Hanjalic, *Beauty is in the scale of the beholder: Comparison of methodologies for the subjective assessment of image aesthetic appeal*, 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX), Singapore, Singapore, September 2014.

II

SUBJECTIVE METHODOLOGIES

2

Reliable & Repeatable Methods for Image Aesthetic Appeal Assessments

This chapter addresses the problem of reliably conducting subjective assessments for quantities that are considered "highly subjective". We also look into the repeatability of such studies across different experiment environments. Our subjective assessments in this chapter focus on image aesthetic appeal, and we show that existing subjective methodologies could yield reliable aesthetic appeal assessments. Moreover, we show that we could obtain comparable results between lab and crowdsourcing experiment environment.

This chapter is published as: E. Siahaan, A. Hanjalic, J.A. Redi, *A Reliable Methodology to Collect Ground Truth Data of Image Aesthetic Appeal*, IEEE Transactions on Multimedia **18(7)**, 1338 - 1350 (2016) [63]

2.1. INTRODUCTION

The importance of multimedia systems that comply to specific user states and needs has been widely acknowledged, as more studies appear that attempt to optimize multimedia systems according to, e.g., the user affective state, cultural background, and social context [64, 65, 66]. Facilitating user preferences and affective state to guarantee high Quality of Experience (QoE) [4] is therefore of major interest when developing multimedia systems. A hot topic within this trend is computational aesthetics [67, 68, 69, 70].

Computational aesthetics aims to build models that can automatically quantify the aesthetic appeal of media. According to the Oxford Dictionary, an aesthetic object is "Giving or designed to give pleasure through beauty." [1] Therefore, understanding the mechanisms that regulate the aesthetic preferences of users with respect to media (e.g. images) is essential to improve users' overall satisfaction with a system. In the past few years, aesthetic appeal has been shown to be critical in determining the perceptual quality of images [12], preferences in visual summarization of large image collections [71], and in steering image retrieval and photo-editing recommendation systems [67, 47].

To design models that predict the aesthetic appeal of an image as perceived by a user, based on image and/or user properties, the availability of ground truth for model training and/or validation is vital. Such ground truth consists of a set of images, each provided with a numerical aesthetic appeal rating (often referred to as mean opinion score, or MOS) that represents the target of the prediction model. These MOSs are the mean of individual ratings given by a pool of users evaluating the images. As these individual ratings act as random variables (i.e. there is no determinate outcome that can be associated with every single individual rating), it is important that the ground truth, as expressed in MOSs, is both reliable and repeatable.

Reliability is verified when high agreement exists in the ratings that the different users gave to the same image. High variability in users' individual ratings results in low confidence in MOS estimation: the value obtained from the participants sample may vary within a large confidence interval (CI). Feeding a model with poorly reliable MOSs as target labels will lead to learning values that have high uncertainty, and therefore may be inexact. In addition, wide CI lead to statistical indistinguishability of a large number of MOSs (and therefore images), making discrimination between beautiful and less beautiful images more difficult.

Repeatability implies that the aesthetic appeal ratings are as independent as possible of the specific experiment in which the images were evaluated. Various factors in an ex-

[1] [Online]. Available: http://www.oxforddictionaries.com/definition/english/ aesthetic

periment session (environmental conditions, user sampling, scaling methodology, etc. [72]) can bias the individual ratings as well as the MOSs. It is possible that very different MOSs are obtained when the same set of images is re-evaluated in a different experiment session (even simply with a different set of users). This can happen (and is yet more worrisome) even when MOS are highly reliable (have small CI). Thus, both repeatability and reliability need to be verified when choosing ground truth for training computational aesthetics models.

Literature on psychometrics (study of users' quantification of their perception of a certain attribute of a stimulus) has often pointed out that many factors may influence the reliability and repeatability of collected data, such as the test instructions, range of stimuli, rating scale, etc. [73, 72, 56, 74, 54, 75]. In fact, recommendations or best practices exist for conducting experiments to collect subjective image ratings; yet they mostly target perceptual quality assessment [52, 56], and it is unclear whether they can be extended directly to the domain of aesthetic appeal. As shown in [58] and [76], aesthetic appeal assessment is impacted by individual differences much more than perceptual quality assessment. Moreover, aesthetic appeal assessments have been found to be less repeatable than image quality ones [57]. Therefore, there is a great need to establish best practices on how to perform aesthetic appeal evaluations that ensure repeatable and reliable ratings.

Among experimental design factors, two seem to have a prominent role in repeatability and reliability: the scale used to rate images, and the experimental environment. Studies in computational aesthetics have used a wide variety of rating scales (i.e. the tool used to quantify human perception of a certain attribute of a stimulus, such as the aesthetic appeal of an image) for collecting data (see Table 2.1), from binary scales to star-based rating systems. It is known, though, from literature in e.g., image quality assessment [72], that the rating scale needs to be chosen carefully to minimize individual differences and maximize user agreement in the ratings. This may be the case for aesthetic appeal evaluations too.

In addition, we look into the influence of the experimental environment on reliability and repeatability of the ratings. Specifically, we compare aesthetic appeal assessments taken in controlled (lab) and uncontrolled (e.g., crowdsourcing) environments. Crowdsourcing (CS) has become a very appealing alternative to lab-based evaluations, given its ability to reach a large number of users with very diverse demographics, in a relatively short amount of time and for a fraction of the cost that lab-based evaluations would usually entail [56]. However, the lack of control on task understanding and execution makes the reliability and repeatability of CS-based ratings questionable, as shown in [57].

Table 2.1: Datasets and subjective methodologies across studies on image aesthetic appeal

Ref no.	Image set			Experiment setting			
	Image source	Number of images	Image type	Rating scale	Attribute(s) scored	Number of users	Experiment environment
[77]	[78]	221	Photographs	10-point rating scale binary rating	Image composition like/dislike	8 experts [78] 168 Amazon MTurk workers	Lab (controlled) Crowdsourcing (uncontrolled)
[57]	[12], internet, and private collection of amateur photographer	200	Photographs	5-point discrete numerical scales with semantic labels at the ends 5-point ACR scales	likeability, familiarity, recognizability, aesthetic appeal recognizability, aesthetic appeal	14 users 1170 users	Lab (controlled) Crowdsourcing (uncontrolled)
[79]	Flickr, Google+, personal collections	339	Photographs	7-point discrete numerical scale	aesthetic appeal	32	Lab (controlled)
[68]	Photo.net	3581	Photographs	7-point scale	aesthetic appeal, originality	n/a	Uncontrolled
[70]	n/a	100	Copies of paintings	5-point scale 5-point scale with "no concern" option	general impression color	42	Lab (controlled)
[80]	Flickr, Kodak picture of the day, study observers, and archive of consumer image sets	450	Photographs	0-100 point scale with semantic labels at the ends	pleasantness artistically	30	Lab (controlled)
[67]	Image sharing portals (e.g. Flickr)	632	Photographs	5-point discrete scale	aesthetic appeal	15	Lab (controlled)
[81]	Dpchallenge.com	255000	Photographs	10-point discrete scale	aesthetic appeal	78 to 549 users per image (mean 210)	Photo-sharing website (uncontrolled)
[82]	Professional photography websites, and collection of amateur photographer	17673	Photographs	3 categories: high quality, low quality, uncertain about quality	aesthetic quality	10	n/a

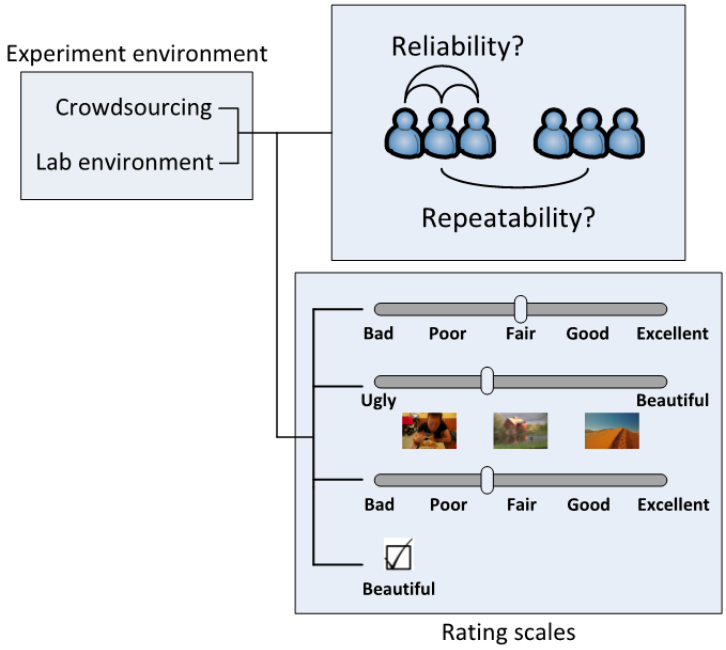


Figure 2.1: Overview of the study: we investigate the effect of experiment environments (laboratory and CS) and four rating scales (further detailed in Section III) on reliability and repeatability of aesthetic appeal ratings.

As a result, the research questions below call for an answer.

1. Which rating scale maximizes the reliability and repeatability of aesthetic appeal subjective evaluations?

2. To what extent are subjective aesthetic appeal measurements reliable and repeatable, when performed in different experiment environments?

To answer these questions, we perform an empirical study in which we carefully setup experiments for aesthetic appeal data collection. In performing this study, we focus on images consumed in web application contexts, such as those found in social media or blog posts, for which management and retrieval systems are more needed. Therefore, we cover content and resolutions typical of consumer images retrievable on the web (detailed in Section 2.3.1). We compare four different rating scales (see Figure 2.1, details in Section 2.3.2), and check which of these yields the most reliable aesthetic appeal ratings.We repeat aesthetic appeal measurements in a lab (highly controlled) environment first, and later on in a CS (less controlled) environment. We then check if there are discrepancies between the resulting measurements, and when applicable, we investigate

their causes, eventually defining clear recommendations on how to perform subjective evaluations of image aesthetic appeal in the different environments to promote more reliability and repeatability.

The rest of this chapter is organized as follows. Factors potentially influencing reliability and repeatability of aesthetic appeal ratings are reviewed in Section 2.2. Section 2.3 gives an explanation of our experimental setup. Section 2.4 presents the measurements we use to analyze reliability and repeatability of aesthetic appeal ratings. In Section 2.5, we discuss our observation on reliability of aesthetic appeal evaluations, and in Section 2.6 our observation on their repeatability. We conclude this chapter in Section 2.7.

2.2. RATING SCALES AND EXPERIMENT ENVIRONMENT FOR IMAGE AESTHETIC APPEAL ASSESSMENTS

Different rating scales and experimental environments affect the reliability and repeatability of image and video evaluations in general, and image aesthetic appeal in particular. As rating scales are related with scaling methodologies, we firstly present a brief discussion of scaling methodologies.

2.2.1. SCALING METHODOLOGIES

Sorting visual media, such as images or videos according to their perceived attributes, has long been investigated, and resulted in the development of various scaling methodologies for quantifying users' perception and/or preference of visual media [72, 83]. On top of the large body of work done in psychophysics to scale and quantify users' sensations and perceptions (e.g. visibility thresholds) [83, 84], the multimedia community has dedicated efforts to scale and quantify multimedia preferences, related to audiovisual characteristics, as well as user expectations and affective state [72, 85, 86]. Besides the rating scales, an important element in the scaling methodology is the way in which stimuli are presented to users, which we will discuss first.

Aesthetic appeal evaluations are typically based on Single Stimulus (SS) methodology [77, 67, 80, 70, 57]. Users are presented with one image at a time and evaluate it on a given rating scale before being presented with another image. The way users rate images in photo-sharing websites also follows this methodology [68, 81]. The SS methodology allows fast evaluations of large sets of images, but can promote unreliability in ratings, since it is not always straightforward for users to quantify their experience of an image without having references [72, 85]. In the field of perceptual quality assessment, where the evaluation task consists of quantifying annoyance induced by visual distortions in

images and videos, the limitation of SS has been addressed by the use of Double Stimulus (DS) methodologies. Here, a reference (undistorted, pristine) image or video is shown along with the (distorted) images or videos to be rated, to function as visual anchor. This reduces the rating task to the task of image comparison [72]. Examples of DS methodologies that have been used in image quality assessment studies are the DS continuous quality scale (DSCQS) [52], DS impairment scale [52], paired comparison [53, 83], and the quality ruler (QR) [85]. The QR method [85], in particular, has been proven to provide more reliable measurements than SS [75]. It is based on the comparison of each image with a set of images (visual anchors) equally spaced in perceptual quality, spanning a predefined quality scale. The users' task is then reduced to comparing the current image with the visual anchors and scoring the image with respect to the anchor's position on the scale. This has been shown to lead to higher inter-rater reliability for perceptual image quality; so far, no adaptation has been proposed for aesthetic appeal. In fact, DS methodologies are not easy to apply for image aesthetic appeal evaluations: it is very difficult, if not impossible, to determine a 'reference' image with maximum aesthetic appeal.

In our study, we resort to using SS methodology, as is commonly adopted in computational aesthetics studies. However, we also adapt the QR method through one of our rating scales (see Section 2.3), to check the reliability of the scores when users have the ease to assess images through comparison with visual anchors.

2.2.2. RATING SCALES

Studies in perceptual quality assessments looked into the effect of rating scales on the reliability of image and video evaluations [74, 87]. Although the evaluations using different rating scales could correlate well with each other in [74], analysis on data collected from different experiments in [87] points to rating scales having an effect on the MOS reliability. The study compared the standard deviations (SDs) of mid-range MOS obtained from different experiments using 5-point discrete and 11-point discrete absolute category rating (ACR) scales, and 100-point discrete and continuous scales (DSCQS methodology). The continuous rating scales gave a lower SD around MOS compared to the other rating scales.

In the field of computational aesthetics, different studies have collected aesthetic appeal ratings based on different types of rating scales: either discrete or continuous, with numerical or categorical labels, even binary (see Table 2.1). However, very few have paid attention to the reliability and repeatability of the ratings produced by these scales. In [77], a comparison of image aesthetic appeal MOS using a 10-point scale and binary

scale (like/dislike) was presented, showing an acceptable, albeit relatively low, correlation between the two (Pearson correlation coefficient of 0.59), posing a question on repeatability of aesthetic appeal evaluation with different rating scales. However, the evaluations with each rating scale were performed in different experimental settings, so it is difficult to pinpoint the factor that generated this lack of repeatability. Moreover, the comparison does not include an in-depth analysis on how reliable the image aesthetic appeal evaluations were per rating scale. In fact, a proper reliability analysis is almost never reported along with the outcomes of image aesthetic appeal evaluations.

2.2.3. EXPERIMENT ENVIRONMENTS

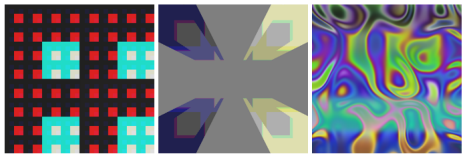
For a long time, subjective tests for the assessment of perceptual quality have been performed in controlled lab settings. In this type of setting, users are asked to perform their evaluation of some image or video stimuli in a controlled room/laboratory, with fixed size monitor, lighting level, viewing distance, and display calibration, for which the International Telecommunication Union (ITU) has published recommendations [52]. In real life, though, users are less likely to see an image or watch a video in a setting such as that recommended for lab experiments. Consequently, researchers have started looking into the assessment of (perceptual) quality in less controlled environments [54, 55]. A large body of work has started exploring the feasibility of CS [56]. Nevertheless, some initial concerns have been raised about reliability and repeatability of e.g. video QoE assessment in CS [88, 56].

In work related to computational aesthetics, as shown in Table 2.1, studies in controlled lab environment [67, 80, 70, 79] have also been complemented by evaluations in uncontrolled environments such as CS platforms (e.g., MechanicalTurk[2] or Microworkers[3]) [77, 57], or photo-sharing websites [81]. CS-based evaluations differ from photo-sharing website-based evaluations for two fundamental factors: 1) the presence of a monetary compensation for the evaluators, and 2) the existence of a specific evaluation protocol. In this study, we choose to use CS platforms instead of photo-sharing websites, to represent uncontrolled experiment environments. This is to keep the subjective test setup (i.e. instructions, experimental protocol, presentation methodology) as close as possible to the lab experiment, for easier comparison.

Although allowing a higher level of control with respect to ratings in photo-sharing websites (CS tasks can be designed in a very rigorous way), the remote nature of the experimental task distribution in CS still poses some serious concerns about reliability and

[2] [Online]. Available: https://www.mturk.com/
[3] [Online]. Available: https://microworkers.com/

(a)



(b)

Figure 2.2: Examples of (a) consumer photographs, and (b) abstract images used as stimuli in our experiment

repeatability of the collected ratings. In fact, previous work related to aesthetic appeal evaluation has shown conflicting results. In [77], the aesthetic appeal ratings obtained in CS and lab for the same set of images presented a rather low correlation. Nevertheless, the CS and lab experiments used different scale types, and the number of users in the lab experiment was very small (8 people). Also in [57], aesthetic appeal ratings collected in lab and CS were found to have a low correlation. The lab and CS experiments used a similar type of scale, a 5-point ACR scale, however, the graphical user interface where the scale was presented was different, which the authors argue may have affected the results. Further reasons for the discrepancy could be a lack of comprehension of the test instructions in CS, and/or a tendency of the crowd-workers to perform the task in a less-than rigorous way. These problems are known in the community [16], and it is so far unclear whether it is possible to overcome them towards reliable and repeatable evaluation of aesthetic appeal in CS.

2.3. EXPERIMENTAL SETUP

To investigate the effect of rating scales on the reliability of aesthetic appeal ratings, we compared in a controlled lab setting the reliability of the aesthetic appeal ratings

collected using four different types of scales (described in Section 2.3.2). Then, to address the effect of experiment environment on rating repeatability and reliability, we compared ratings collected in the controlled lab setting with those collected through CS (which setup is described in Section 2.3.3).

2.3.1. STIMULI

Fifty four images were selected to be used in this experiment, following the design choices of previous literature devoted to the comparison of reliability and repeatability of subjective ratings [74, 89, 75, 90, 91]. Although this number of images may be lower than that of images used in studies aimed at collecting ground truth for model training, it has been shown to be sufficient to estimate with good confidence reliability and repeatability of evaluations [74, 89, 75, 90, 91].

Of the 54 images, forty were consumer photographs, in line with our application scenario. As literature suggests that "the bigger-the better" does not apply for image aesthetic appeal [92], we focused on the type and size of images typically used in social media and blog posts. We selected a subset of the 200 images used in [57]. The photographs were JPEG and BMP images with sizes ranging between 600 × 600 and 720 × 478 pixels, in line with image sizes adopted in previous studies targeting similar applications [77, 67, 81]. The subset uniformly covered a wide range of aesthetic appeal, as well as a wide range of content categories. All JPEG images were compressed at a high quality to avoid the appearance of blocky or blurry artifacts. The remaining fourteen images were abstract images generated using Jene evolutionary art program [93], based on [94]. These were included to challenge the scale usage, as it is known that users disagree the most on assessing the beauty of abstract images [95]. The abstract images were 256 × 256 PNG images. Examples of the images used in our experiments can be seen in Figure 2.2.

2.3.2. RATING SCALES

Four types of scales were compared in our experiment, as shown in Figure 2.3. First, we included a scale commonly used in image/video quality testing, being advised by the ITU [52], and adopted in [72]: the discrete 5-point ACR scale [hereafter referred to as ACR, see Figure 2.3(a)]. The scale has five category levels "Bad—Poor—Fair—Good—Excellent" attached to it that are equally spaced along a linear scale. Users evaluate the aesthetic appeal of images by choosing one of the category levels.

To check whether continuous scales would provide more reliable MOS than the discrete ACR (as found in [87]), we also included in our experiment a continuous scale with 5-point category labels corresponding to those in the discrete 5-point ACR scale.

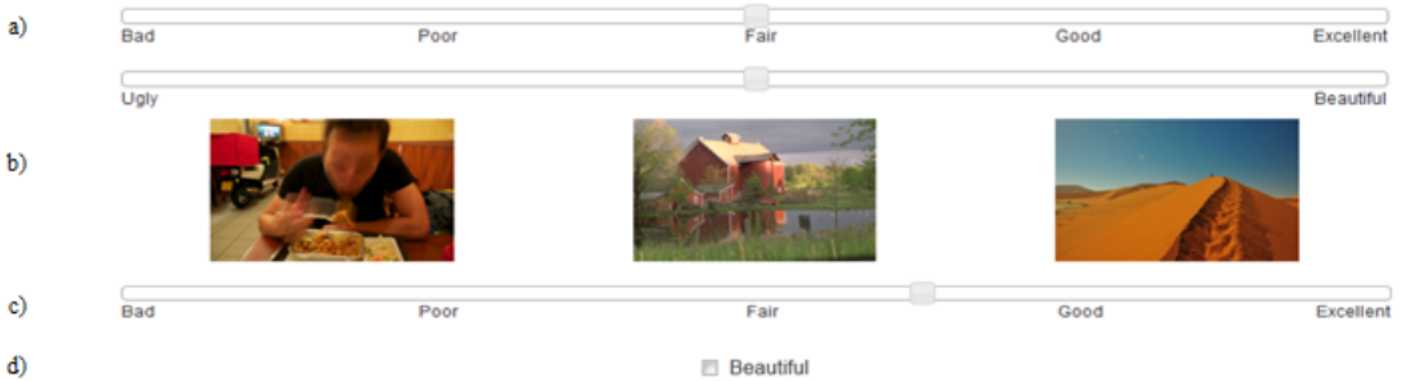


Figure 2.3: Interface of scales used in experiment: (a) discrete 5-point ACR scale (ACR), (b) continuous scale with visual anchors (VA), (c) continuous scale with 5-point category labels (CONT), and (d) binary scale (BIN).

A continuous scale might also provide more flexibility for users, as they would not have to round up or down the rating of an image when they think it is positioned more in-between two aesthetic appeal levels. This scale (referred to as CONT) is shown in Figure 2.3(c).

The next scale used in our study is one that we implemented based on the QR [85].We were interested in observing whether or not the presence of visual anchors would be beneficial for the reliability of the aesthetic appeal scores. Thus, we presented a scale with semantic labels ["Ugly" and "Beautiful" were placed at the extreme ends of the scale, see Figure 2.3(b)], as well as three visual anchors attached to it, equally spaced across the length of the scale. We will refer to this scale as VA. The function of the visual anchors was to indicate visually the level of image beauty a certain part of the scale corresponded to. Three anchors were selected: one to characterize the top (right) part of the scale (beautiful images), one for the middle part of the scale, and one for the bottom (left) end of the scale (ugly images).We chose the visual anchors based on an existing dataset of images rated in terms of aesthetic appeal [57]. The beautiful image [right-hand side of Figure 2.3(b)] was selected among those from [57] rated in the range between the third quartile and the maximum of the aesthetic appeal ratings distribution. The image characterizing the middle part of the scale was selected as the one with the aesthetic appeal score closest to the median of the aesthetic appeal distribution. Finally, the image representing ugly images was chosen from those that had aesthetic appeal values in the range between the minimum and the first quartile of the aesthetic appeal ratings distribution of the image set [57]. The images characterizing the top and bottom parts of the scale were not positioned at the scale ends but slightly towards the center, so that room was left to rate images deemed more beautiful or uglier than the above-mentioned anchors.

Finally, we included a binary scale (BIN), with which users solely needed to tick a box when they thought the image was beautiful. This scale was chosen as it implements a straightforward evaluation method that demands low cognitive load, and has shown to yield comparable judgment with a 10-point discrete scale in a previous study [77]. Moreover, it might be an easy scale for users to interpret and use, especially in an uncontrolled setting such as the CS platform. The interface that we used for this scale is shown in Figure 2.3(d).

2.3.3. CONTROLLED LAB EXPERIMENT SETUP

24 users took part in the lab experiment. Each user had to perform four subtasks in one session. In each subtask, users were asked to rate all the images using a particular rating scale. To avoid learning effects in the ratings, users were presented with each scale in

a different order. At the beginning of the experiment, users were given a brief explanation of the experiment setup and their tasks. They were not aware of the experiment's purpose (i.e., to compare the effect of using different rating scales on aesthetic appeal evaluation). Before each subtask, an explanation was given on how to rate images using the scale for that subtask. Two training images were then shown so that the user could get used to the scale interface. The user could then proceed with the task. After the second session, there was a short break to prevent fatigue and minimize memory effects on the image ratings. After the four subtasks, a questionnaire was filled in, asking the user to rank the four scales based on the preference of usage. One session of the experiment took 30 minutes on average. Figure 2.4 shows the interface presented to users for rating an image.

As recommended for subjective evaluations related to image and video perceptual quality [52], our lab experiment was performed in a fixed environmental setup for every user. The viewing conditions in the lab experiment were set with constant illumination at approximately 70 lux, using a 23" LED backlight monitor at 1920 × 1080 resolution. Users were sitting approximately 70 cm away from the display. There was no time constraint on image observation and scoring.

2.3.4. CS EXPERIMENT SETUP

The CS experiment aimed at establishing whether aesthetic appeal evaluations performed in a lab setting were repeatable (and as much reliable) in a less controlled environment, and whether this repeatability would depend on the specific scale used. Thus, we set up a series of campaigns on Microworkers to evaluate the image set with all four scales. After accessing the campaign page, workers[4] followed a link to a webpage from where the evaluations would be performed. Images and scales were presented through the same interface as in our lab experiment (Figure 2.4). The webpage used for our experiment was built using Ruby on Rails 5[5] version 2.2.2.

Since CS tasks need to be kept short to maximize workers engagement (and seriousness in carrying out the task) [96], we divided the 54 images into 3 sets for 3 campaigns. To make sure every set of images in a campaign would span a similar range of aesthetic appeal, five fixed images with significantly different levels of aesthetic appeal were included in every set. Based on aesthetic appeal scores from the lab experiment, we chose images with the minimum MOS, MOS closest to the 25th, 50th, and 75th percentile of the aesthetic appeal MOS distribution, and the maximum MOS. This practice has been

[4] For the crowdsourcing experiment, we will refer to participants/users as "workers."
[5] [Online]. Available: http://rubyonrails.org/

shown to be effective in limiting context effects in the aesthetic appeal ratings [96]. A Kruskal–Wallis test was performed to ascertain that the five images' scores were statistically different ($\chi^2 = 56.54$, $df = 4$, $p = 0.000$). Further check with pairwise comparison showed that except for the pair of images corresponding to the 25th and 50th percentile of the MOS distribution, all other pairs were indeed statistically different. In the end, two campaigns included 21 images and one campaign included 22 images.

The task of each worker was to rate all images in one set, in a single campaign. For each campaign, workers were given instructions on how to use the scale in that campaign to evaluate images, and two training images were first shown to familiarize them with the scale interface. To control for workers' trustworthiness and commitment to the task, two content questions were added to each campaign, in the form of multiple choice questions after certain images, asking the workers what the main object was on the last image they saw.

For each scale type, 3 campaigns were set up, giving a total of 12 campaigns. 30 to 44 workers participated in each campaign, and every worker could only take part in one campaign. This was to maintain the workers naive about the presence of content questions (avoiding therefore learning effects on task completion). Workers' origin was restricted to countries with recognized proficiency in English (English-speaking countries and Western Europe), to avoid language barrier in understanding the instructions of the task. Each worker received $0.50 for the completion of one campaign.

2.4. DATA AND ANALYSIS PREPARATION

2.4.1. MOS CALCULATION AND OUTLIER ANALYSIS

The individual aesthetic appeal ratings for each image i and observer n (Opinion Scores $OS_s(i,n)$)were obtained from each rating scale s as follows. For ACR, each category label was converted to a rating score of 1, 2, 3, 4, and 5, respectively, with 1 corresponding to 'bad' and 5 corresponding to 'excellent'. For CONT, ratings were translated into scores between 0–100, with the 5 category labels having the scores 0, 25, 50, 75, and 100, respectively. Ratings on the VA scale were similarly translated into scores between 0–100, with 0 being the extreme closest to the 'ugliest' visual anchor. For BIN, when a user rated an image as 'beautiful', the image was assigned a 1, and 0 otherwise.

Before conducting any analysis of the data collected from our experiments, a preprocessing step was necessary. For data collected in the lab, we calculated the mean opinion scores (MOSs) of every image i as rated on scale s, by averaging the individual opinion scores of the N users rating i on scale s

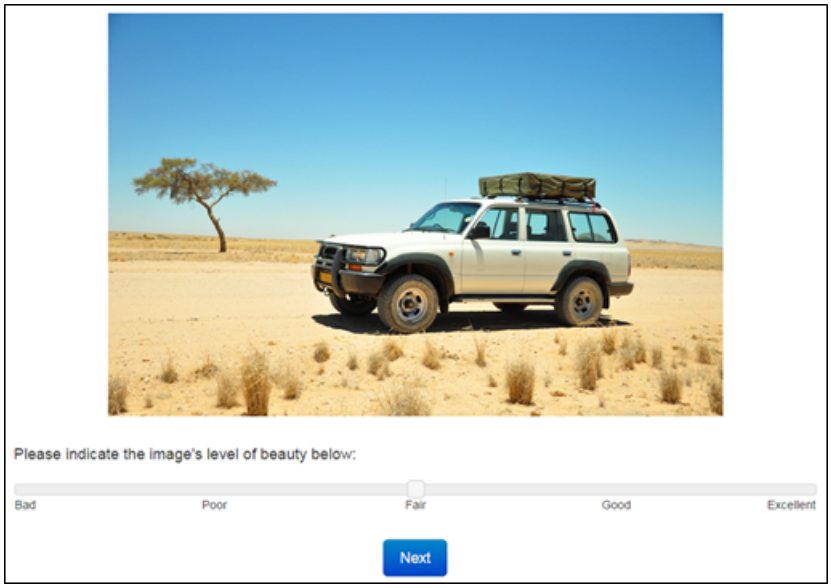


Figure 2.4: Interface presented to user for rating an image's aesthetic appeal

$$MOS_s(i) = \frac{1}{N}\sum_{n=1}^{N} OS_s(i,n) \tag{2.1}$$

with $OS_s(i,n)$ being the individual rating for image i on scale s by user n. For all scales, we performed outlier detection.

Outlier users were removed if their scoring behavior was considered random (i.e., significantly uncorrelated to the rest of the user population), according to the procedure recommended in [52]. A random scoring behavior was detected if, within the set of images, multiple ratings of a user were higher than the MOS plus twice its SD and multiple ratings were lower than the MOS minus twice its SD (so, the scoring behavior was not systematically higher and lower of that of the population, but randomly either higher or lower) [52]. One user in our lab experiment resulted to be an outlier and was discarded from the data analysis.

For the CS experiment results, we considered best practices for analyzing data gathered through CS [56], and conducted several steps to filter out possibly unreliable workers. We firstly filtered workers that did not show sufficient commitment to the task, i.e. did not give the right answers to the content questions, or did not rate all 21 or 22 images in the campaign they participated in. We then performed outlier analysis based on time statistics and scoring behavior [57]. We check the time statistics by adapting the outlier

Table 2.2: Overview of the number of workers considered unreliable according to the criteria defined in Section 2.4.1

Campaign #	Total Workers	No. workers giving correct answers to content questions	No. workers completing the test	Time statistics	Outlier detection	Final number of subjects
1	30	28	28	No outlier	No outlier	28
2	30	23	22	No outlier	1 outlier	21
3	31	29	29	No outlier	2 outliers	27
4	39	33	32	No outlier	No outlier	32
5	44	24	22	No outlier	No outlier	22
6	32	31	31	No outlier	No outlier	31
7	31	30	29	No outlier	No outlier	29
8	44	31	30	No outlier	1 outlier	29
9	33	32	32	No outlier	No outlier	32
10	32	31	31	No outlier	n/a	31
11	34	29	28	No outlier	n/a	28
12	33	31	30	No outlier	n/a	30

detection in [52], where for each image in a campaign, the mean and SD of time taken to rate the image was calculated. For every worker, then, we computed the number of images whose rating took a time outside the interquartile range of the worker's rating time for that image. Workers who rate more than half of the images in a time outside this interquartile range were considered as outliers. Outliers were also removed if a random scoring behavior was detected [52], as already done for the lab experiment participants. However, the outlier detection was not applicable on binary data. Details of the filtering performed on our CS data are given in Table 2.2. The table shows that from the 12 campaigns, more than 75% of the data could be retained in 9 campaigns. Most of the data was discarded due to incorrect answers to the content questions.

After the data filtering, we also aligned the ratings from the different scales to be expressed within the same numerical range. We chose as a common range that of the ACR scale, i.e. [53, 4]. We used linear normalization to rescale our data collected with VA, CONT, and BIN scale types.

2.4.2. RELIABILITY MEASUREMENTS

To evaluate reliability of subjective ratings, some well established statistical indicators exist that are widely used across different domains of study. These indicators serve as diagnostic tools for reliability, and their accuracy is independent of the quantity evaluated. Two of these indicators, which we adopt in this chapter, are Cronbach's alpha and the inter-rater correlation coefficient (ICC). Literature on interrater reliability stresses

the importance of interpreting reliability values in comparison with those from studies using similar subjective assessment methodologies [97]. Therefore, in addition to Cronbach's alpha and ICC, we look into reliability indicators used in studies of perceptual quality assessment, namely the SD of opinion scores (SOS) hypothesis [98], the subject bias [99], and the scale discriminative power [74]. Below, we describe in more detail each of the reliability measurements that we use. Each measure is computed per scale and experimental environment (lab and CS) separately.

Cronbach's Alpha: Cronbach's alpha is a widely established statistical measure for internal consistency, often used to check if different test items in a questionnaire consistently measure the same construct. Alpha is obtained by computing the average of correlations between scores for each pair of items in a questionnaire. In our study, we consider users as questionnaire items and the images the different subjects. Our goal is to verify that users are measuring the same underlying construct, i.e. aesthetic appeal of images. Thus to measure the reliability of the different scale types, we calculate Cronbach's alpha on the individual ratings of every user for each image in the set. A high Cronbach's alpha would indicate that on average, the ratings of every user for the image sets are positively correlated, hence denoting high agreement (at least in the ordering of the images).

2) Intra-Class Correlation Coefficient: ICC is the most widely adopted indicator to measure interrater reliability, and measures the extent to which the variance in the data is due to the fact that different images are being rated with respect to the variance of ratings between users. A high ICC indicates that most of the variance can be explained by differences in the images, and not by individual differences in their evaluations by the users, thus indicating a high degree of agreement among users. Low ICC indicates poor interrater reliability. For the lab data, we used a two-way random model with absolute agreement to calculate the ICC of each scale type. To check reliability in CS settings, because not all images were rated by the same users due to the fact that they were subdivided in campaigns, we used a one-way random model [96].

3) SOS Hypothesis: The SOS hypothesis was proposed in [98] to measure consistency in subjective QoE measurements. The hypothesis models a square relationship between the SOS and the MOS. Per each scale *s*, the SOS of image i is given by the following equation:

$$SOS_s(i) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(OS_s(i,n) - MOS_s(i))^2} \tag{2.2}$$

and the SOS hypothesis for a 5-point scale (i.e., the range within which we normalized

all MOS from all scales) is given by [98]:

$$SOS_s(i)^2 = \alpha(-MOS_s(i)^2 + 6MOS_s(i) - 5). \tag{2.3}$$

The parameter α represents the extent to which the SOS changes depending on the MOS values. A bigger value of α gives a steeper quadratic curve indicating that the SOS towards the middle scale value (e.g. 3 in a 5-point scale) is considerably larger than the SOS of the values toward the ends of the scale range. This translates to a wider (95%) CI especially in the middle of the scale, as shown by Equation 2.4, indicating lower agreement among users.

$$CI_s(i) = MOS_s(i) \pm \big(1.96 * \frac{SOS_s(i)}{\sqrt{N}}\big). \tag{2.4}$$

To obtain reliable evaluations of image aesthetic appeal, we would favour rating scales that show lower SOS hypothesis α value compared with the other scales.

4) Scale Subject Bias: The notion of subject bias was proposed in [99], as part of a model for user rating behaviour. In the model, the rating expressed by user n for image i on scale s is expressed as

$$OS_s(i,n) = MOS_s(i) + \Delta_{n,s} + \epsilon_{i,n,s} \tag{2.5}$$

with $\Delta_{n,s}$ being the subject bias characterizing the rating behavior of user n, and $\epsilon_{i,n,s}$ an error term for scale s and image i. Subject bias is computed per user: the smaller it is, the closer the user rating is to the image true value, estimated as the MOS. It may be the case that different types of scales intrinsically lead users to express different opinions: in such case, we would expect subject bias values to be high for all users. In the following, we use the mean subject bias across all users using a given scale in a lab experiment or CS campaign to characterize the reliability of such scale.

5) Scale Discriminative Power: The scale discriminative power [74] indicates to what extent a scale delivers MOSs that are clearly discriminable (i.e., significantly different) for a given pair of images. In [74], the scale discriminative power is calculated as the percentage of pairs of stimuli that had significant differences through ANOVA tests followed by post-hoc multiple comparisons using the Bonferroni method. Ideally, if a scale had full discriminative power, all MOSs had negligible SD, indicating high agreement across users and hence high reliability. In this chapter, we measure each scale's discriminative power as the percentage of pairs of images that are statistically different using a Kruskal–Wallis tests followed by multiple pairwise comparisons.

2.4.3. REPEATABILITY MEASUREMENTS

For repeatability measurements, we evaluate whether the MOS produced by different scales in different environment are compatible, in terms of (linear) ranking and distribution. Also in this case, we resort to indicators that are independent on the quantity for which repeatability is being assessed. Specifically, we use the following measures:

1) Pearson Correlation: We use Pearson correlation to verify MOS(i) (per each scale s) collected in our lab and CS experiments are linearly related, i.e. if their values grow in a linear monotonic way across different rating scales and environments.

2) Rating Distribution Shape: Since correlation is independent on the absolute value of the MOS, we further check if ratings collected on different scales and in different environments cover similar aesthetic appeal range and have similar distribution. We compute the skewness of the individual scores distributions, and we calculate pairwise the Kullback–Leibler divergence between the rating distribution per environmental condition. The lower the distance, the more similar the distribution. In addition, similar skewness values point to distributions with similar shape.

2.5. RELIABILITY OF AESTHETIC APPEAL EVALUATIONS

We performed three analysis related with reliability of aesthetic appeal scores across rating scales. The first targets the ratings obtained in the lab environment. The second targets the lab ratings of the subset of abstract images. As it has been observed that people disagree most when assessing the beauty of abstract images [95], it is interesting to check the extent to which the four scales would still deliver reliable ratings. The third analysis concerns the ratings obtained in the CS environment.

2.5.1. RELIABILITY OF RATING SCALES IN LAB ENVIRONMENT

In our lab experiment, every user went through four subtasks in which they rated all images using the four different scales. We firstly observed whether or not this procedure had an effect on the users' ratings, in the form of learning or memory effects due to rating the same images four times, albeit on different rating scales. As reported in our previous analysis [100], each scale yielded highly correlated MOSs across subtasks ($p > 0.93$ in all cases). This correlation remained stable throughout the experiment, as shown in Figure 2.5. Each column represents the correlation between the MOSs obtained for each possible pair of scales, when MOSs are computed based only on the users who used those scales in a specific subtask. The high correlation then was not due to learning or memory effects. If a learning or memory effect was in place (e.g., users attempted to

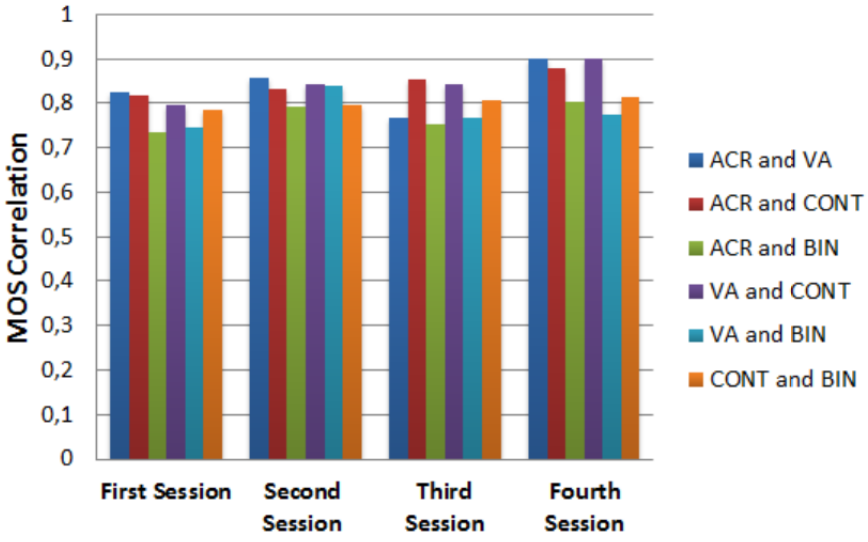


Figure 2.5: Correlation between MOS of the four scale types through the four subtasks in lab experiment.

give always the same score to the same image across the scales), we would expect the correlation to grow over subtasks, but that is not the case.

We proceed with our reliability analysis by computing the measures outlined in Section 2.4.2. We first look at the SOS hypothesis α values of the four scales, as shown in Table 2.3. The α values for ACR, CONT, and VA are comparable to each other, and are similar to those obtained for evaluations of QoE in other domains such as web surfing and cloud gaming [98], for which α values usually range between 0.2342 and 0.3497 (albeit higher than those typical of perceptual image quality assessment tasks). The α for BIN, instead, is three times bigger. This implies that the SD of the MOSs obtained with BIN is higher, on average, than obtained with ACR, CONT, or VA. On the other hand, the Cronbach's alpha and ICC value of the four scales are similar, and close to 1, which indicates high correlation of image ratings among different users, and low variability of ratings for the same image, respectively. It is important to note that an ICC > 0.9 is typically considered to indicate excellent inter-rater reliability. Nevertheless, it is well known that the sensitivity of Chronbach's alpha decreases as the number of raters increases, thus this result may not be very representative.

The average subject bias values for ratings on the four scales are comparable with those of subjective data mentioned in literature, having absolute value mostly between 0 and 1 [99]. With the average subject bias measure, we see again that the BIN scale performs differently than the other three scale types. This could be an indication of users

Table 2.3: Comparison of reliability measures for ACR, VA, CONT, and BIN based on lab experiment

	SOS hypothesis alpha	Cronbach's alpha	ICC	Average subject bias	Scale discriminative power
ACR	0.24	0.94	0.942	0.281	26.48%
VA	0.24	0.94	0.944	0.231	24.95%
CONT	0.25	0.94	0.939	0.290	25.58%
BIN	1.04	0.93	0.890	0.441	22.43%

having difficulty interpreting the binary scale; in fact, several users in the lab experiment indicated after their session that they had doubts deciding their "internal boundary" above which an image would be judged as beautiful. This might also explain the scale VA having the lowest subject bias among the scales. The design of the VA scale reduced the rating task to a comparison one, making it easier for users to interpret where they should locate the position of the scale for a particular image.

For the calculation of scale discriminative power, in all cases, the Kruskal–Wallis test found a significant effect of the image content on individual ratings (for ACR: df = 53, chi^2 = 549.38, p = 0.0; for VA: df = 53, chi^2 = 543.76, p = 0.0; for CONT: df = 53, chi^2 = 534.89, p = 0.0; for BIN: df = 53, chi^2 = 453.94, p = 0.0). Here, we see that the ACR, VA, and CONT scales perform slightly better than the BIN scale. However, compared with scale discriminative power measured for n-point discrete scales in the assessment of video quality [18] (around 50%), the discriminative power of the scales in assessing image aesthetic appeal seems to be generally lower. This result shows less agreement in general across subjects on image aesthetic appeal compared with image or video perceptual quality, as already pointed out by the SOS values.

In summary, from the measures shown in Table 2.3, we see that ACR, CONT, and VA are more reliable than the BIN scale, as they yield ratings with higher inter-user agreement. This suggests that for building image aesthetic appeal models, it is more reliable to use scores collected with ACR, CONT, or VA scales instead of BIN as the model's target scores.

2.5.2. RELIABILITY OF RATING SCALES IN ASSESSING ABSTRACT IMAGES

Abstract images are most challenging to be reliably evaluated by users: the lack of a clear semantic meaning leads users to perform a more affective evaluation of the image, thereby resulting in lower inter-user reliability [95]. It is therefore interesting to check whether the different scales can help compensate for the intrinsic unreliability of these measurements. The results for reliability of abstract image evaluations are presented

Table 2.4: Comparison of reliability measures for ACR, VA, CONT, and BIN based on abstract images assessment

	SOS hypothesis alpha	Cronbach's alpha	ICC	Average subject bias	Scale discriminative power
ACR	0.304	0.477	0.403	0.422	0
VA	0.280	0.487	0.381	0.426	0
CONT	0.260	0.615	0.557	0.488	0
BIN	1.045	0.233	0.218	0.699	0

in Table 2.4. The SOS hypothesis α values are comparable (although slightly higher) to those obtained for the whole image set (Table 2.3). Again, ACR, VA, and CONT show more reliability than BIN. Looking at the Cronbach's alpha and ICC values, we see that the three scale types, ACR, VA, and CONT, also indicate higher reliability than BIN. However, the Cronbach's alpha and ICC for the abstract images are considerably lower compared to the values obtained for the whole image set. This could be due to both low interrater reliability and low variability in the aesthetic appeal of the stimuli. To check this case further, we observe the average subject bias and scale discriminative power. The former is almost twice as the average subject bias for the whole image set. In addition, the scale discriminative power indicates that, on all four scales, no images have significantly different MOS from others. Details of the Kruskal–Wallis tests performed to calculate the scale discriminative power are as follows: ACR: df = 13, chi^2 = 17.05, p = 0.197; VA: df = 13, chi^2 = 16.89, p = 0.204; CONT: df = 13, chi^2 = 22.78, p = 0.04; BIN: df = 13, chi^2 = 14.32, p = 0.352. In fact, all of the abstract images had MOS values between 1.5 and 2.7 (in a [1, 5] range). This low variability in the aesthetic appeal of the stimuli partially explains the low scale discriminative power and ICC. We can therefore conclude that users generally disagreed more in rating abstract rather than real-life images [46], leading to lower reliability of all scales, with the CONT being seemingly able to provide more reliable scores than the other three scales.

2.5.3. RELIABILITY OF RATING SCALES IN CS ENVIRONMENT

We verified whether the reliability findings reported for lab experiments would hold in an uncontrolled environment by repeating the previous analysis (in Section 2.5.1) on the aesthetic appeal scores collected in CS. For some measures, such as SOS hypothesis alpha, ICC and scale discriminative power, we calculated the measures for all image ratings given with the same scale type. For the other measures, we calculated the measures per campaign, as they are specific to the pool of users and stimuli scored. Since each campaign had different sets of images, we could not average out Cronbach's alpha and

subject bias across campaigns.

Our results in Table 2.5 show that in general, the SOS hypothesis α values for the scales are higher in CS than in the lab experiment. However, the alpha value of ACR shows to be the most comparable with the alpha values of ACR, CONT, and VA in lab experiment. The values also confirm that ACR, CONT, and VA yield more reliability than BIN. When we look at the Cronbach's alpha values, we see that the values remain similar to those measured from the lab experiment. This indicates a strong correlation of image scores among different groups of users for every scale type; nevertheless, these high values may be due to the high number of raters involved (see [101]). The ICC indicator also shows excellent reliability for the four scales, indicating that users were able to agree on the ratings of the images, with BIN performing comparable to the other scales in this case.

The average subject bias values for the four scales in the CS environment are, in general, higher than those in the lab. This could be caused by users interpreting the scales differently in CS, as they had to rely solely on written instructions to carry out the evaluation task. However, in some of the campaigns, the average subject bias for some scales is smaller or comparable to that of the corresponding scale in the lab experiment. This happened for one campaign of VA and BIN. The scale that has the lowest average subject bias in all three campaigns (and most comparable to its average subject bias in lab experiment) is ACR. Finally, for the calculation of scale discriminative power, details of the Kruskal–Wallis test performed for each scale type are as follows: ACR: $df = 53$, $chi^2 = 830.66$, $p = 0.0$; VA: $df = 53$, $chi^2 = 656.29$, $p = 0.0$; CONT: $df = 53$, $chi^2 = 737.25$, $p = 0.0$; BIN: $df = 53$, $chi^2 = 629.23$, $p = 0.0$. From our CS experiment, the four scales show higher discriminative power compared to that of the lab experiment. An explanation for this is the smaller number of images being rated in each CS campaign, giving a bigger ratio of significantly different pairs of images per campaign, and thus a higher discriminative power on average across campaigns. Similarly to the lab experiment, though, ACR shows higher scale discriminative power compared to the other three scales.

From the data shown in Table 2.5, then, we could conclude that, among the four scales, ACR conveys the most reliable judgments in CS, yielding similar (and highest) user agreement in both CS and lab environments.

2.6. REPEATABILITY OF AESTHETIC APPEAL EVALUATIONS

Having shown that most of the scales above yield reliable MOSs in the lab, and (albeit to a lesser extent) in CS, we are now interested in looking into the repeatability of image

Table 2.5: Comparison of Reliability Measures for ACR, VA, CONT, and BIN based on crowdsourcing experiment

Scale	Campaign Set #	SOS Hypothesis alpha	Cronbach's alpha	ICC	Average subject bias	Scale discriminative power
ACR	1	0.268	0.961	0.957	0.300	34.41%
	2		0.967		0.250	
	3		0.98		0.255	
VA	1	0.353	0.973	0.953	0.503	30.12%
	2		0.964		0.223	
	3		0.955		0.591	
CONT	1	0.390	0.963	0.951	0.411	31.66%
	2		0.96		0.441	
	3		0.968		0.402	
BIN	1	1.034	0.943	0.931	0.520	28.30%
	2		0.946		0.353	
	3		0.954		0.749	

Table 2.6: Pearson Correlation Coefficient of All Image Scores Per Scale Between Lab Experiment and Crowdsourcing Experiment

ACR	VA	CONT	BIN
0.872	0.868	0.886	0.899

aesthetic appeal ratings, i.e. whether or not the ratings obtained are consistent across different experiment environments.

Table 2.6 shows that BIN yields the highest MOS correlation between lab and CS scores. Nevertheless, we find that the correlation between lab and CS MOS is relatively high for all types of scale. Given the BIN ratings are considerably less reliable than those of the other scales (see Section 2.5), the correlation value of BIN may be misleading, and the MOS values highly uncertain (i.e. varying within wide CIs; depending on the goodness of the estimation of the true value of the MOS, correlation values may change).

The individual rating distribution for the four scales in lab environment is shown in Figure 2.6, while Figure 2.7 shows the rating distribution of the four scales in CS. The skewness value of each distribution can be found on top of each distribution plot in the same figures. We see that ACR and BIN maintain their distribution shape more than VA and CONT between lab and CS experiments. The tendency of ratings to be skewed towards the right or left part of the scale is maintained between lab and CS experiments. For ACR and CONT, for example, the distributions are skewed to the right in both lab and CS, while for VA and BIN, the distributions are skewed to the left. These tendencies possibly hint at scale types creating a positive or negative bias in users when giving their evaluations (in the case of VA, it may be due to the selection of the anchors). To further check which scale gives the most repeatability, we computed the Kullback–Leibler divergence between the rating distributions of lab and CS for each scale type. The outcome is presented in Table 2.7. Among the four scales, BIN shows the smallest divergence be-

tween lab and CS rating distributions, with ACR following close by after.

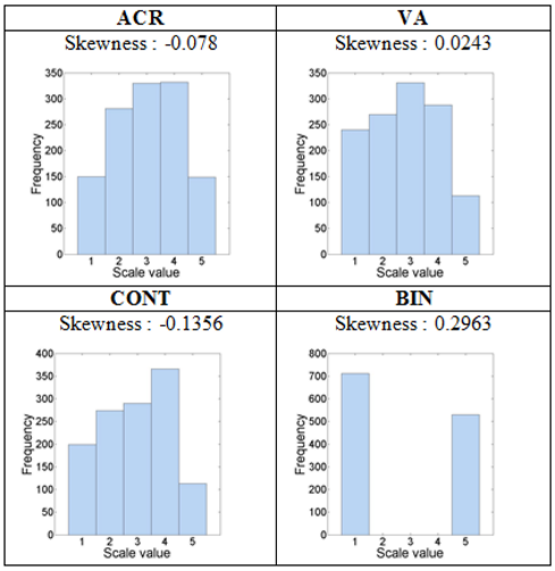


Figure 2.6: Rating distribution for ACR, VA, CONT, and BIN in lab experiment environment

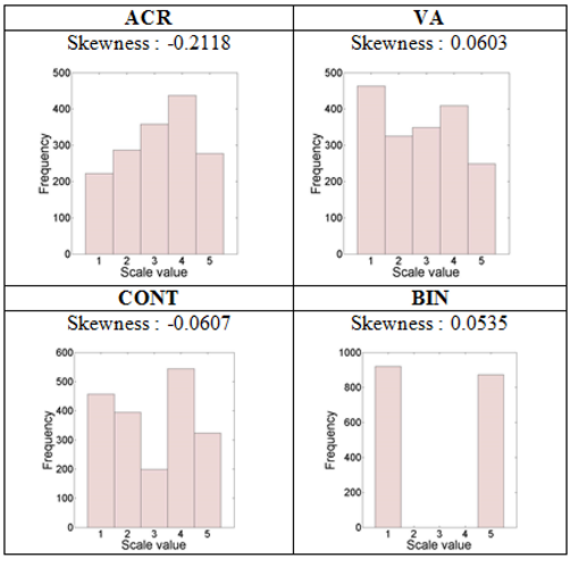


Figure 2.7: Rating distribution for ACR, VA, CONT, and BIN in CS experiment environment

As our findings show better repeatability of image aesthetic appeal evaluations from that presented in previous work on comparing lab and CS experimental results [57], it

Table 2.7: KL-Divergence Between Distributions of All Image Scores Per Scale Type Between Lab Experiment and Crowdsourcing Experiment

ACR	VA	CONT	BIN
0.0197	0.0337	0.096	0.0072

is interesting to analyze what might have influenced the repeatability of evaluations in our case of study. It should be noted that in [57], users were asked to rate images on other aspects besides their aesthetic appeal, such as the image's familiarity, colorfulness, and recognizability. In our experiment, we asked users to only rate image aesthetic appeal. Moreover, we used identical scoring interface between lab and CS, which was not the case in [57]. This result suggests that aesthetic appeal measurements may be repeatable across different environments, given that extra care is put in maintaining the instructions, and that the scoring interface is identical in the different experiments. Another factor that might have influenced repeatability of evaluations between lab and CS experiments is the simplicity of task given to users. In the previous study, users were instructed to rate the "aesthetic appeal" of the image presented to them, whereas in our study, we asked users to rate the "beauty" of the image presented. Although both terms are synonymous, it could be argued that the term "beauty" is more familiar and easier to interpret by users, making their task easier to understand. Further work is necessary to verify whether this is the case.

2.7. CONCLUSION

In this chapter, we looked into the problem of reliability and repeatability of the subjective evaluations of image aesthetic appeal. We performed subjective tests to compare four different commonly used scoring scales, and also two different experiment environments (laboratory and CS) to check which would yield the most repeatable and reliable results. Our study showed that among the four scale types that we used, the 5- point ACR and the scale with visual anchors yielded the most reliable ratings in lab environment. Although in CS the rating reliability decreased for all scales, ACR was the one providing the highest user agreement in this case. We also noted that the continuous scale with 5-point ACR labels (CONT scale) promoted good reliability compared to the other three scales when used to assess abstract images, which intrinsically challenge agreement among users. This may indicate CONT to be more suitable for judging artificial images (e.g. computer graphics), or images spanning a small aesthetic appeal range (in our case, the abstract images spanned a very limited portion of the aesthetic appeal scale).

Our analysis also showed that image aesthetic appeal evaluations are repeatable across different experiment environments, particularly between lab and CS environments, with ACR and BIN scales giving more repeatability than VA and CONT scales. We note that the repeatability of image aesthetic appeal evaluations between lab and CS experiments may be influenced by the presentation of scales or rating methodologies in both environments, and the simplicity of the task that users are asked to perform. Although the BIN scale shows good repeatability across experiment environments, considering the reliability measures of ratings obtained with that scale, we do not recommend its usage, as the aesthetic appeal MOS produced appears to be very imprecise. Summarizing, a continuous scale with visual anchors or an ACR scale seem to be most appropriate to rate aesthetic appeal in controlled lab conditions. When using CS, it is necessary to keep in mind that the reliability of the ratings will decrease, so a larger number of workers (participants) may need to be involved. Maximum reliability in CS will be achieved when adopting an ACR scale, which will also guarantee ratings to be as close as possible to those that would be obtained by performing the experiment in a controlled lab environment (repeatability).

In all cases, it is important to keep in mind that reliability of aesthetic appeal ratings is bounded to be measured with relatively low reliability, when compared e.g., to perceptual quality assessment. This is due to individual preferences and background impact on image appreciation, sometimes even more than that of physical, objectively measurable properties of the image. Researchers aiming at predicting aesthetic appeal of images should take this into account: besides selecting a methodology which maximizes reliability (for which we give indication), researchers may want to involve more subjects in the ratings, so as to increase the confidence with which the MOS is estimated, or even abandon the prediction of MOS in favor of e.g., individual scores, or distribution of ratings over the rating scale for the same image. In general, finally, when publishing results of subjective evaluations of aesthetic appeal, researchers should report a reliability analysis to enable the community to make appropriate choices for the usage of the data as ground truth.

Of course, the recommendations above are made based on experiments having certain limitations.We chose to compare four rating scales, and only considered CS to represent uncontrolled experiment environments. As a result, whether these results can be fully extended to e.g., evaluations collected through photo-sharing websites needs further investigation. Moreover, we did not consider other factors that might influence the outcome of aesthetic appeal ratings, such as task instructions or range and context effects (we used the same instructions and range of image across our experiments). An

interesting direction for future work would be to observe how far image aesthetic appeal evaluations would be repeatable when the same images are evaluated within different ranges and contexts. This could provide insights into how to realign image aesthetic appeal scores from different datasets to provide larger image aesthetic appeal ground truth.

3

A Mixed Methodology Approach to Point Cloud Quality Assessment

In the previous chapter, we addressed the problem of reliably conducting subjective assessments for "highly subjective" quantities. In this chapter, we investigate subjective quality assessments for point clouds. Dealing with new types of media, for example point clouds, often brings the uncertainty on how to evaluate their quality, i.e. existing subjective methodologies may not be appropriate to evaluate these media. This chapter explores quantitative and qualitative approaches to understand the way users judge point cloud images, and presents insights on performing subjective quality assessments of point cloud images.

This chapter is based on: E. Siahaan, J.A. Redi, A. Hanjalic, P. Cesar, *Full-reference quality metrics for point cloud compression based on color distribution*, Under Review, (2018)

3.1. INTRODUCTION

Recent years have seen growing interest in the development of immersive systems, i.e. systems that create a sense of presence (immersion) by providing highly realistic representation of objects or places [102]. Immersive systems have become more accessible thanks to advanced and cheaper acquisition, display and network technologies, together with the availability of more computing power. Figure 5.1 illustrates examples of immersive systems, such as a virtual museum exhibition, a multi-party tele-presence system, and an augmented reality system.

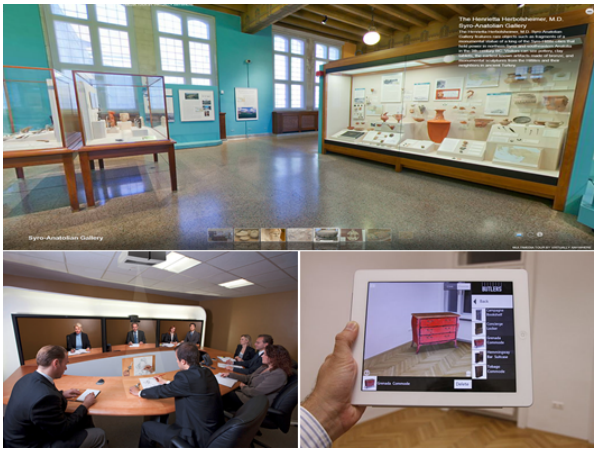


Figure 3.1: Examples of immersive systems (top to bottom, counter-clockwise): virtual museum[1], tele-presence system[2], and augmented reality systems[2].

Point cloud is one of the technologies used for representing 3D scenes and objects in virtual immersive systems. A point cloud is a set of points in a three-dimensional space, that together make up the surface of an object. Point clouds are typically acquired through depth sensors or multiple cameras, and a point cloud may consist of up to millions of points. Besides its coordinates in the 3D space, each point may also include additional attributes, such as its color value, transparency, normal, etc. A point cloud does not require triangulation computations (as is required for meshes), or heavy pre-processing, and it is more resilient to noise. All these characteristics make it compelling to use point clouds as media type for real time immersive systems.

Nevertheless, point clouds may still demand high resources. For example, a high

[1] Screenshot taken from virtual tour of the Oriental Institute Museum (https://oi.uchicago.edu/virtualtour)
[2] Images taken from Wikimedia Commons.

quality point cloud of a human object can contain up to $9 * 10^5$ points, taking 18 MB of space. Consequently, much interest is given to point cloud compression technology. Compression technology is typically evaluated using quality metrics, to observe whether or not the compressed image or video satisfies consumers' or users' expectations for viewing quality. Various point cloud compression technology have been proposed until now ([103, 104, 105, 106, 107]), however, not many have looked into its quality assessment ([60, 108, 107]).

Existing studies on the quality assessment of point cloud compression have mainly shown that state-of-the-art objective quality metrics do not yield scores that correlate well with human judgment for point cloud compression [20, 108, 60]. Not much insight has been presented on fully exploring users' perception of point cloud quality through different subjective methodologies. One study by Alexiou and Ebrahimi [109] looked into comparing the use of Absolute Category Rating (ACR) and Double Stimulus Impairment Scale (DSIS) methodologies to assess point cloud geometry. Their study shows that although the two methodologies are statistically equivalent, the DSIS methodology yields higher confidence interval for assessing point clouds with compression-like artifacts.

While it is important to understand how different standard methodologies affect point cloud quality assessment, it is also necessary to investigate what users perceive when evaluating point cloud quality. As point cloud is a relatively new type of visual media, we may not be aware of all quality dimensions that users consider in evaluating it, and consequently be ignoring these dimensions in designing our metrics. This concern occurred in the past, when other types of new visual technology emerged, for example 3D displays. Research showed that users considered perceived depth as an additional factor to image fidelity (which traditionally indicates quality), thus bringing forth the importance of evaluating viewing experience and naturalness [49]. To explore the factors that affect user perception of point cloud quality, we propose to use a mixed methodology in our subjective experiments.

This chapter aims to fill in the above research gaps related with the subjective quality assessment of point cloud compression (PCC). Our contribution in this chapter is as follows.

1. We perform a user study on a realistic point cloud dataset, using a mixed methodology (quantitative and qualitative) to collect ground truth subjective scores, and investigate more deeply how users perceive degraded point clouds due to compression.

2. We conduct a reliability analysis on the subjective methodology for point cloud quality assessment, to give insights on how to design subjective assessment tests

for point cloud quality compression

This chapter is organized as follows. In the next section, we discuss some background work related to our study. In Section 3.3, we describe the experiments we performed, and in Section 3.4, we present some analysis on the subjective quality assessment of point cloud, and observe how users perceive and evaluate point cloud quality. In Section 3.5, we discuss the main findings of our chapter and their implications. We conclude this chapter in Section 3.6.

3.2. BACKGROUND

In this section, we describe the type of point clouds and their applications that we consider in our study, and give a short overview of related work to point cloud subjective quality assessment.

3.2.1. APPLICATION AND DATASETS

There are many application context in which point clouds are utilized, such as 3D immersive tele-presence systems [107], VR content viewing [110], free viewpoint sports replay [111], indoor/outdoor spatial reconstruction ([112, 113]), and cultural heritage archiving [114]. In most applications, the point clouds used are complex, i.e. containing not only 3D-coordinates of the points, but also additional attributes such as color, normals, reflectance, etc.

Up until now, point cloud quality assessment has mainly focused on point cloud geometry in their studies [108, 20, 60, 109]. Recently, several point cloud datasets were made publicly available that provide more complex sequences such as human close-ups [115], full human body [116], architecture facades [117], and more. These datasets would allow for quality assessment studies, especially those on QoE, to focus on the types of applications users would use. We therefore consider these datasets in this chapter.

3.2.2. IMMERSIVE QUALITY ASSESSMENT

In recent years, various types of new media have emerged that offer upgraded viewing experiences from the conventional 2D images or videos. These include stereoscopic images/videos, light fields, point clouds, among others. Research has shown that these new types of media often involve additional factors that influence their quality assessment other than signal fidelity alone (which is the typical indicator of quality). For example, when evaluating 3D display technologies, it was shown that image depth perception is an additional factor to image fidelity, and contributes to users' overall viewing experi-

ence and perception of image naturalness [49]. Understanding these factors would help design better metrics for immersive media.

Not much research has been done on the the subjective assessment of point cloud quality, and the factors that users consider in their assessments. A study by Zhang et al. presented a subjective experiment to compare user judgment on point clouds degraded with coordinate noise and those degraded with color noise, and show that people seem to be less sensitive to color noise than to coordinate noise[118]. Another study by Alexiou and Ebrahimi [109] took on comparing the use of two different subjective methodologies (ACR and DSIS) to assess point cloud geometry. Their study shows that the DSIS methodology yields higher confidence interval for assessing point clouds with compression-like artifacts.

In general, there is still a lack of in-depth analysis on users' perception of point cloud quality, for example, what factors users consider when asked to judge point cloud quality. To answer this question, we propose to perform a subjective experiment using mixed methodology, i.e. a combination of quantitative and qualitative studies. A qualitative study would allow users to freely explain and describe their evaluation process, i.e. the characteristics in a point cloud that influenced their judgment of quality. This could help discover point cloud features that need to be considered in designing better objective metrics.

3.3. EXPERIMENT SETUP

We perform a subjective experiment using mixed methodology to evaluate point cloud compression quality. Our quantitative study aims to collect ground truth subjective scores, and observe user reliability when assessing point cloud quality. Our qualitative study aims to explore in more detail users perception of point cloud quality, and uncover some quality dimensions that we may consider when building objective metrics for point clouds. We first explain the dataset that we use to conduct our experiments, and continue with the setup of our quantitative and qualitative study.

3.3.1. DATASET

Figure 5.2 shows the point cloud sequences that we use in our study. We use 6 frames from 6 different sequences of full-body humans from [116]. We limit our experiments to these sequences, and do not include other types of sequences (for example, non-human objects, or close-up human objects) available through other datasets for the following reasons. Firstly, we wish to avoid content categories becoming a confounding factor in

Figure 3.2: Point cloud sequences used in the experiment, from the publicly available 8i Voxelized Full Bodies (8iVFB v2) dataset [116].

our analysis. Secondly, sequences taken from different datasets often vary in perceptual quality due to different acquisition techniques.

Each frame of point clouds from the dataset is compressed into 4 different levels of compression using the compression algorithm in [107]. The compression parameters used are Level of Detail (LoD) 10, LoD 9, LoD 8, and LoD 7. LoD 10 means that the compression uses a 10-b octree setting, i.e., 10 quantization bits per direction. In the rest of this chapter, we refer to compression with LoD 10 as compression level 1, and so on, compression with the lowest LoD (LoD 7) as compression level 4. We obtain a total of 24 point cloud sequences to be rated by our participants. During the rendering of the point clouds, we assign different point sizes for each compression level, such that each object can be seen in full (i.e., no hollow parts due to missing points can be seen).

We then created video sequences for each point cloud (reference and compressed), which show the point cloud rotated along the vertical axis. This allows users to have a 360 degree view of the point clouds. The resolution of the video sequences is 1280x720, and were 40 seconds long each, with 30 frames per second. Figure 3.3 illustrates some of the viewpoints taken from one of the resulting test videos.

3.3.2. QUANTITATIVE SUBJECTIVE STUDY

We perform a quantitative subjective experiment to obtain ground truth subjective scores of our point cloud sequences. 23 people participated in the experiment, 6 of which are

Figure 3.3: In the test video sequences, each point cloud was rotated along the vertical axis such that users had a 360 degree viewpoint of it.

female and 17 are male. The participants' age ranges from 22 to 33 years old, and most of them are naive to point clouds.

A training session was given before the test, so that participants could familiarize themselves with the task and rating interface. We used different point clouds from our test set for the training session. Users had full control of the time taken to evaluate a test point cloud, as they had to press a button on the test screen to display the next pair of test sequences. A 32 inch DELL UHD 4K monitor was used in the experiment, and users were seated at a distance of twice the height of the monitor, in compliance with the recommendation in [53]. The experiment room had controlled lighting as recommended in [52].

Participants were asked to evaluate the quality of point clouds using a Simultaneous Double Stimulus presentation, in which the reference point cloud is shown alongside the test point cloud. Participants were then asked to rate the level of degradation for the test point cloud using a 5-point Degradation Category Rating (DCR) scale [53]. Following the current standard in subjective assessments [53], the point cloud sequences were shown in a non-interactive manner, meaning that participants could not interact with the stimulus and change its viewpoint or its scale. We created two different viewpoint sequences for our point clouds, to prevent the collected scores biasing one particular viewpoint. The viewpoints rotate the point clouds along the vertical axis, and slowly zooms in or out on the point clouds at certain moments. A reference point cloud would always be shown with the same viewpoint as the test point cloud.

3.3.3. QUALITATIVE SUBJECTIVE STUDY

A qualitative subjective study is performed to allow users to express more freely what they perceive when asked to judge point cloud quality, and give us a better understanding of how they perceive point cloud quality. In this chapter, we use the Descriptive Sorted Napping method ([119, 120]). This method presents users or participants with a

number of test items, and asks them to sort and group the items on a *Nappe*, or a blank space, based on how similar or dissimilar users find them. The closer two items are placed on the Nappe, the more similar they are perceived by a user, and vice versa. After all items are placed on the Nappe, users are asked to draw a circle around the different groups of items they have decided on, and write down or explain the similar characteristics attributed to items in the same group.

This method was originally developed in the food sciences field [119], and recently has been used in the quality assessment field to explore the way users construct their judgment of Quality of Experience (QoE) ([121, 122]). We use this method to learn what attributes users would associate with degraded point clouds when asked to group point clouds based on their similarity in quality. Participants were given an interface through a tablet device in which they are presented with a set of numbered blocks and a blank space that represent the Nappe. On a big computer screen, participants could watch the point cloud sequences, each corresponding to a numbered block on the tablet. Participants could then place the numbered blocks on the Nappe based on how similar they perceive the quality of point clouds corresponding to each block. Users were allowed to watch the point cloud sequences multiple times, and were not restricted in time to complete the task.

Once users were done placing the numbered blocks on the Nappe, they could draw circles around numbered blocks that they consider to belong in the same group, and type in the attributes or characteristics related to the perceived quality of that group. This is done until all numbered blocks have been included in at least one group. This whole process was explained to participants before they started the task, and they were instructed specifically to sort and group the point clouds based on the perceived quality. Figure 3.4 shows the interface used for the study.

3.4. SUBJECTIVE QUALITY ASSESSMENT OF POINT CLOUD COMPRESSION

3.4.1. QUANTITATIVE STUDY ANALYSIS

We conducted an outlier analysis on the scores obtained from our quantitative study, according to the recommendation in [52]. No outlier was found in the analysis. We then calculated the mean opinion score (MOS) and standard deviation of opinion score (SOS) of each image.

To compare the level of user agreement in this task with user agreement in other quality assessment tasks, we calculated the SOS hypothesis alpha [98] of our collected

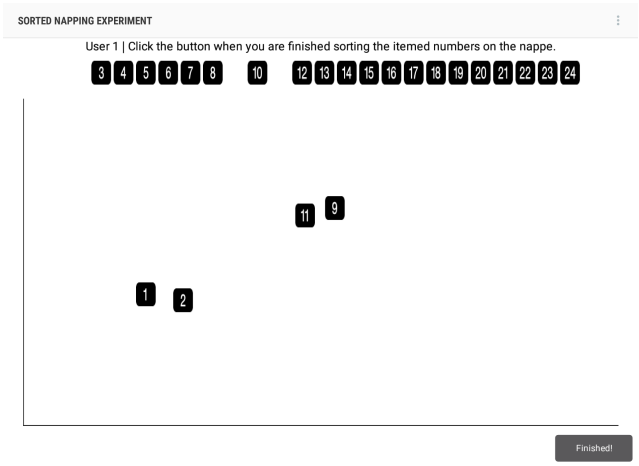


Figure 3.4: The interface used to perform the sorted napping experiment. Users could watch the videos corresponding with each number on a computer screen in front of them.

data. The SOS hypothesis alpha represents how the standard deviation of opinion scores (SOS) changes with the mean opinion scores (MOS) values using a parameter α. A higher value of α would indicate higher disagreement among user scores. The hypothesis for a point cloud p is expressed as in Eq. 4.1 below.

$$SOS^2(p) = \alpha(-MOS^2(p) + (V_1 + V_K)MOS(p) - V_1 V_K), \tag{3.1}$$

where V_1 and V_K indicate the lowest and highest end of a rating scale, respectively.

Table 3.1 shows the SOS Hypothesis Alpha values taken for different image or video quality assessment tasks, as presented in [98], and the SOS Hypothesis Alpha value for our study. The table shows that the level of user agreement obtained in our study lies within the range of alpha values for visual quality assessment tasks. This indicates that the use of Simultaneous Double Stimulus presentation with Degradation Category Rating (DCR) scale yields reliable scores for assessing point cloud compression quality in our study.

Figure 3.5 plots the MOS distribution across impairment levels and content. At the higher levels of quality (low compression rate), we observe that the content "Soldier" is significantly rated higher than other contents for the same compression level. We confirm this through an ANOVA test on the sequences with compression levels 1 and 2 separately; opinion scores being the dependent variable and content being the independent variable. For both compression levels, the MOS for content "Soldier" has a statistically significant difference with the MOS for other contents (p<0.05). This may indicate at

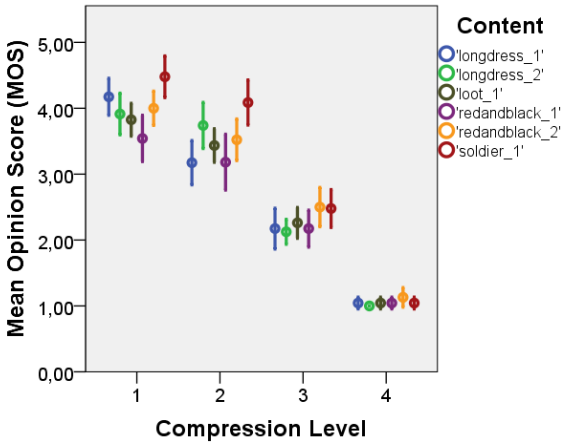


Figure 3.5: Mean Opinion Scores (y-axis) over the different compression levels (x-axis), and content (colored). Compression level 1 indicates the lowest compression level, and 4 indicates the highest compression level. Error bar indicates 95% confidence interval.

Table 3.1: Comparison of SOS Hypothesis α Values for Different Quality Assessment Tasks (as reported in [98])

Application and Subjective Study	**SOS α**
Point cloud compression quality assessment (our study)	0.146
Image quality assessment, JPEG images of LIVE dataset [123]	0.0400
Image quality assessment, IRCCyN/IVC Scores on Toyama [124]	0.1715
Video streaming quality assessment, H.264 codec [125]	0.1078
Video streaming quality assessment, MPEG2 codec [125]	0.1137

the higher range of quality, certain characteristics of the content "Solider" could mask artifacts that users otherwise perceive in other content.

3.4.2. QUALITATIVE STUDY ANALYSIS

Next, we perform a hierarchical multiple factor analysis (HMFA) on our collected data from the qualitative study, to find shared structures among the individual sortings. Figure 3.6 shows Confidence Ellipses of the sorted napping configurations obtained through the study on 24 and 12 point cloud sequences, respectively. The study on 24 point cloud sequences includes the whole quality range in our dataset, while the study on 12 sequences only includes the upper-half of the quality range in our dataset. Each colored ellipse in the figures represents the spread of each item along the two dimensions that explain most of the variance of the data. The two dimensions of the plot for the 24 items

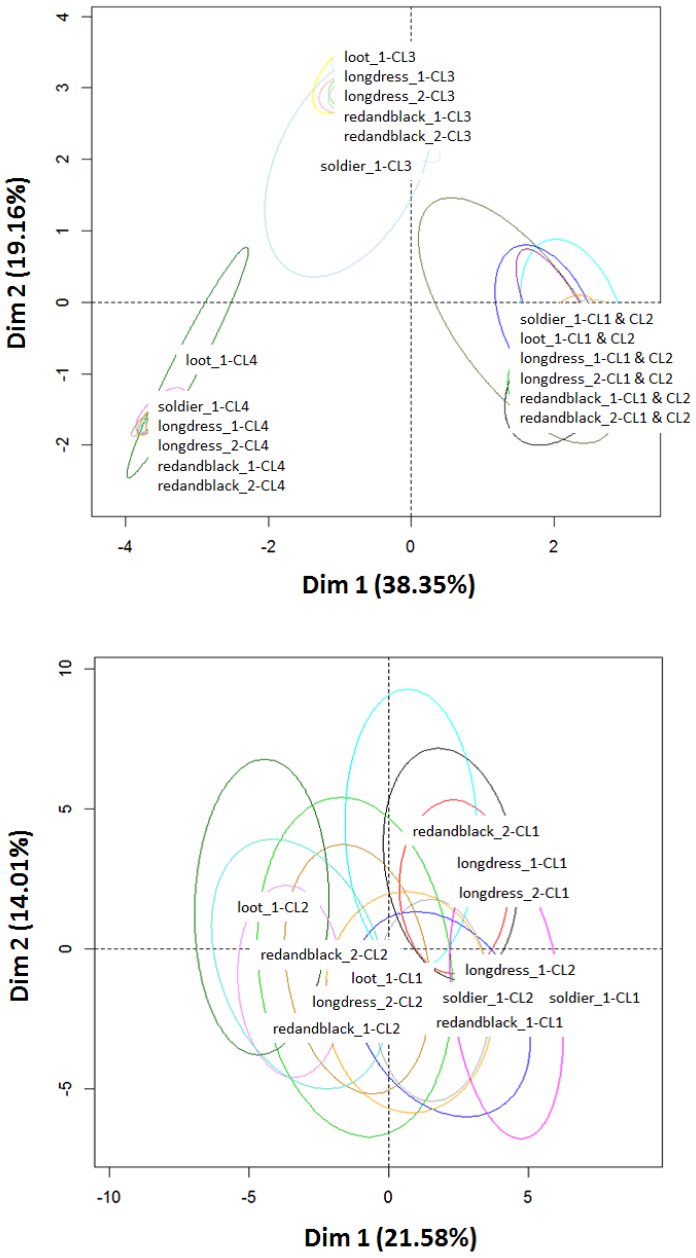


Figure 3.6: Confidence ellipses of the sorted napping configuration for the study on the whole quality range/24 items (top figure) and the upper-half of the quality range/12 items (bottom figure) in our dataset. The suffix "CL*n*" indicates the compression level with n=1 indicating the lowest compression level.

explains 57.51% of variance in the data together, while the two dimensions of the plot for the 12 items explains 35.59% of variance of the data together. When possible, we place the name of a point cloud sequence at the center of its corresponding ellipse. The suffix "CLn" at the end of a sequence name indicates the compression level of the point cloud, with n=1 being the lowest compression level (and thus, highest quality).

From the plot for the 24 items, participants seem to agree on three clear separate clusters of items. The two clusters on the negative side of the Dim 1 axis show clusters of point clouds with highest level of compression (most left-side cluster), and point clouds of the second highest level of compression. We look into the attributes that participants associated with these three different clusters, and find that for the high levels of compression (the two lowest quality ranges), participants noted that the point clouds appear blurry or blocky (due to the large point sizes assigned to the point clouds during rendering). Meanwhile, the higher quality point clouds were described as being clear. Some participants specifically mentioned facial features and patterns in clothing as cues to judge the clarity of the point cloud videos.

As shown in the Confidence Ellipses plot for the 24 items (Figure 3.6, top figure), the sequences belonging to the higher levels of quality fall into the same cluster at the positive side of the Dim 1 axis. To look more closely at how users perceive the quality of these sequences, we conducted the Sorted Napping study only on these 12 sequences with different participants, as described in Section 3.3.3. The Confidence Ellipse plot for the 12 items shows that there are still no clear separate clusters formed from the 12 items.

Looking at the attributes that participants associated with the sequences that fall into the negative side of the Dim 1 axis for the 12 point clouds, we find that half of the participants mentioned color distortion to describe these sequences. When asked to explain what they meant by color distortion, participants commented that some of the sequences had colors from their clothes projected onto their skin. For example, they could see some red dots in the skin of the woman in the "Red and Black" sequence, or some blueish tones on the skin of the man in the "Loot" sequence.

3.5. DISCUSSION

Our subjective experiment aimed to understand better what artifacts users observe when asked to judge the quality of compressed point clouds. Using mixed methods (combining quantitative and qualitative studies), we observe that while users looked into the clarity of faces and overall structure to judge low quality point clouds, they noticed color

alterations in higher quality point clouds. Having a qualitative study in our experiment allowed us to have a deeper probe into users' perception of point cloud quality, and thus shows what a valuable tool mixed methods can be in quality assessment studies.

We also looked into the level of agreement that users had in evaluating point cloud quality, compared with user level of agreement in other quality assessment studies. This user agreement measure is useful to compare different subjective methodologies for the same task, or different tasks using the same subjective methodology. Future studies on point cloud quality assessment using the same setup as ours (Simultaneous Double Stimulus with DCR scale) or a different setup could refer to this study to compare their obtained level of user agreement.

In this study, we use point cloud sequences that would be closer to users' application context compared to point cloud geometry. However, we realize that our experiment setup may be argued as not being fully ecologically valid. One participant asked in our studies whether or not we wanted her to evaluate the point clouds as if she was consuming them in a gaming context, or normal viewing context. This brings up an essential issue: evaluating the quality of experience (QoE) of point clouds depending on its application context may need a more elaborate setting that mimics the context as closely as possible.

Furthermore, we note here that in our experiment, we created our dataset using 24 point cloud stimuli. This means that this study is more of a first step towards more studies on point cloud quality assessments.

3.6. CONCLUSION

In this chapter, we present a study on the subjective assessment of point cloud image compression. In our subjective assessment, we perform a mixed method experiment on 24 compressed point cloud stimuli of full human body representations. Our quantitative and qualitative studies were done using the Double Stimulus Impairment Scale, and Sorted Napping, respectively. Our experiment shows us that users take notice of the clarity of the face and overall structure of the body while evaluating point cloud quality, as well as color alterations or distortions.

This study is limited to the 24 stimuli that we use in our experiments. No other dataset of point cloud quality scores is available up until now, which limits our ability to evaluate our metrics on different point cloud stimuli. We hope that future work on point cloud quality assessment would result in more datasets available for the research community. Another future direction would be to look into the effect of user interac-

tion with the stimuli on quality evaluation. Research on light field images have looked into this problem, comparing whether or not interactivity (letting users freely change a stimulus' viewpoint and focal point poses) would influence the subjective quality assessment of light field images [126]. As point cloud also allows for interactivity, and could be used in applications that encourage interactivity (such as in VR displays), this research problem would be worthwhile to look into.

III

OBJECTIVE QUALITY METRICS

4

SEMANTIC-AWARE BLIND IMAGE QUALITY ASSESSMENT

After looking at different aspects of QoE subjective assessments in the previous part, this chapter marks the start of our research related with objective QoE metrics. We begin by looking into how semantic categories, as a representation of image content, influence user perception of image QoE. Afterwards, we propose to incorporate semantic category features into no-reference image quality metrics (NR-IQMs), and show that our approach improves state-of-the-art NR-IQMs.

This chapter is published as: E. Siahaan, A. Hanjalic, J.A. Redi, *Semantic-aware blind image quality assessment,* Signal Processing: Image Communication **60**, 237-252 (2018).

4.1. INTRODUCTION

A recent report on viewer experience by Conviva shows that users are becoming more and more demanding of the quality of media (images, videos) delivered to them: 75% of users will give a sub-par media experience less than 5 minutes before abandoning it [127]. In this scenario, mechanisms are needed that can control and adaptively optimize the quality of the delivered media, depending on the current user perception. Such optimization is only possible if guided by an unobtrusive, automatic measure of the perceived Quality of Experience (QoE) [4] of users.

Algorithms that predict perceived quality from an analysis of the encoded or decoded bitstream of the media content are often referred to as Image Quality Metrics (IQM), and are typically categorized into full-reference (FR) or no-reference (NR) methods [8]. FR methods predict quality by comparing (features of) an impaired image with its original, pristine version. NR methods, on the other hand, do not rely on the availability of such reference images, and are therefore preferred for real time and adaptive control of visual quality.

NR methods often approach the problem of predicting quality by modeling how the human visual system (HVS) responds to impairments in images or videos [8, 27]. This approach implies that users' QoE depends mostly on the visibility of impairments, and that a measure of visual sensitivity alone is enough to predict visual quality. In this chapter, we challenge this view, and we prove empirically that semantic content, besides impairment visibility, plays an important role in determining the perceived quality of images. Based on this result, we propose a new paradigm for IQM which considers semantic content information, on top of impairment visibility, to more accurately estimate perceived image quality.

The potential to exploit image semantics in QoE assessment has already been recognized in previous research that investigated the influence of various factors, besides impairment visibility, on the formation of QoE judgments. Context and user influencing factors [4, 39, 41, 128, 129], such as physical environment, task, affective state of the user and demographics, have been shown to be strong predictors for QoE, to the point that they could be used to automatically assess the perceived quality of individual (rather than average) users [41]. A main drawback of this approach is that information about context of media consumption or preferences and personality of the user may prove difficult to collect unobtrusively, or may require specific physical infrastructure (e.g., cameras) or data structure (e.g., preference records). As a result, albeit promising, this approach has limited applicability to date.

A separate but related trend has instead looked into incorporating in image quality metrics higher level features of the HVS that enable cognition, such as visual attention [130]. This has been shown to bring significant accuracy improvements without an excessive computational and infrastructural overhead, as all information can be worked out from the (decoded) bitstream. The first steps in this direction have investigated the role of visual attention in quality assessment [130]. In [44], it was shown that impairments located in salient or visually important areas of images are perceived as more annoying by users. Because those areas are more likely to attract visual attention, the impairments they present will be more visible and therefore more annoying. Based on this rationale, a number of studies have confirmed that, by adding saliency and/or visual importance information to quality metrics, their accuracy can be significantly improved [131, 132, 133].

The study in [134] brought this concept further by identifying visually important regions with those having richer semantics, and incorporating a measure of semantic obviousness into image quality metrics. The study reasoned that regions presenting clear semantic information would be more sensitive to the presence of impairments, which may be judged more annoying by users as they hinder the content recognition. The authors therefore proposed to extract the object-like regions, and weight them based on how likely the region is actually containing an object. They would then extract local descriptors for evaluating quality from the top-N regions.

In this work, we look deeper at the role that semantics plays in image quality assessment. Our rationale relies on the widely accepted definition of vision by Marr [135]: vision is the process that allows “to know what is where by looking”. As such, vision involves two mechanisms: the filtering and organizing of visual stimuli (perception), and the understanding and interpreting of these stimuli through recognition of their content [136]. The earliest form of interpretation of visual content is semantic categorization, which consists of recognizing (associating a semantic category to) every element in the field of view (e.g., "creek" or "forest" in the top-left picture in Figure 4.1). In vision studies, semantics refers to meaningful entities that people recognize as content of an image. These entities are usually categorized based on scenes (e.g., landscape, cityscape, indoor, outdoor), or objects (e.g., chair, table, person, face).

It is known that early categorization involves basic and at most superordinate semantic categories [137, 138], which are resolved within the first 500 ms of vision [139]. Most of the information is actually already processed within the first fixation (~100 ms, [62]). Such a rapid response is motivated by evolutionary mechanisms, and is at the basis of every other cognitive process related to vision. When observing impaired images,

however, semantic categories are more difficult to be resolved [140]. The HVS needs to rely on context (i.e. other elements in the visual field) to determine the semantic category of, e.g., blurred objects. This extra step (1) slows down the recognition process, and (2) reduces the confidence on the estimated semantic category. In turn, this may compromise later cognitive processes, such as task performance or decision making. Hence, visual annoyance may be a reaction to this hindrance, and may depend on the entity of the hindrance as well as on the semantic category of the content to be recognized. Some categories may be more urgent to be recognized, e.g. because of evolutionary reasons (it is known, for example, that human faces and outdoor scenes are recognized faster [62]). Images representing these categories may tolerate a different amount of impairment than others, thereby influencing the final quality assessment of the user.

It is important to remark here that the influence of semantic categories on visual quality should not be confused with the perception of utility or usefulness of an image [141, 43]. Image utility is defined as the image usefulness as a surrogate for its reference, and so relates with the amount of information that a user can still draw from an image despite any impairment present. The idea that image usefulness can influence image quality perception has been exploited in some work on no-reference image quality assessment such as in [134], although there are studies that argue the relationship between utility and quality perception is not straightforward [141]. Instead of looking at the usefulness of an image content, we look at users' *internal* bias toward the content category, and show in this chapter the difference between the two and their respective relationship with quality perception.

In our previous research, we conducted a psychophysical experiment to verify whether the semantic content of an image (i.e., its scene and/or object content category) influences users' perception of quality [61]. Our findings suggest that this is the case. Using JPEG impaired images, we found that users are more critical of image quality for certain semantic categories than others. The semantic categories we used in our study are *indoor, outdoor natural* and *outdoor manmade* for scene categories, and *inanimate* and *animate* for object categories. In [142], we then showed initial results that adding object category features to perceptual quality features significantly improves the performance of existing no-reference image quality metrics (NR-IQMs) on two well-known image quality datasets. Based on these studies, in this work we look into improving NR-IQMs by injecting semantic content information in their computation.

In this chapter, we extend our previous work to include (1) different types of impairments and (2) scene category information in NR-IQM. As a first step, we collect subjective data of image quality for a set of images showing high variance in semantic content.

Having verified the validity of the collected data, we then use it as ground truth to train our semantic-aware blind image quality metric. The latter is based on the joint usage of perceptual quality features (either from Natural Scene statistics [33], or directly learned from images [143]), and semantic category features. We then show the added value of semantic information in image quality assessment, and finally propose an analysis of the interplay between semantics, visual quality and visual utility.

Our contribution through this chapter can be summarized as follows.

1. We introduce a *new image quality dataset comprising a wide range of semantic categories*. In the field of image quality research, several publicly available datasets exist. However, most (if not all) of these datasets do not cover the different semantic categories extensively or uniformly. To open more possibilities of research on visual quality and semantics, we set up an image quality dataset which spans a wider and more uniform range of semantic categories than the existing datasets.

2. We show how *using scene and object information in NR-IQMs improves their performance across impairments and image quality datasets.* We perform experiments to analyze how different types of semantic category features would be beneficial to use in improving NR-IQM. We also compare the performance of adding semantic features to improve NR-IQMs on different impairments and image quality datasets.

This chapter is organized as follows. In the following section, we review existing work on blind image quality assessment, creation of subjective image quality datasets, and automatic methods for categorizing images semantically. In Section 4.3, we introduce our new dataset, SA-IQ, detailing the data collection, reliability and analysis to prove that semantic categories do influence image quality perception. In Section 4.4, we describe the experiments proposing our semantic-aware objective metrics, based on the addition of semantic features to the perceptual quality ones. In addition, in Section 4.5, we compare the relationship of semantic categories with image utility and image quality. We conclude our chapter in Section 4.6.

4.2. RELATED WORK

4.2.1. NO-REFERENCE IMAGE QUALITY ASSESSMENT

Blind or No-reference image quality metrics aim at predicting perceived image quality without the use of a reference image. Many algorithms have been developed to perform this task, and usually fall into one of two categories: impairment-specific or gen-

eral purpose NR-IQMs. As the name suggests, impairment-specific NR-IQMs rely on prior knowledge of the type of impairment present in the test image. Targeting one type of impairment at a time, these metrics can exploit the characteristics of the particular impairment and how the HVS perceives it to design features that convey information on the strength and annoyance of such impairments. Examples of these metrics include those for assessing blockiness in images [16, 144], blur [36, 128], and ringing [17].

General purpose NR-IQMs deal with multiple impairment types, and do not rely on prior information on the type of impairment present in a test image. This of course allows for a wider applicability of the metrics, but also requires a more complex design of the quality assessment problem. To develop these metrics, usually a set of features is selected that can discriminate between different impairment types and strengths, followed by a mapping (pooling) of those features into a range of quality scores that matches human perception as closely as possible [145].

Handcrafted features are often used to develop general purpose NR-IQMs, one of the most common being natural scene statistics (NSS), although other types of features have also been proposed, such as the recent free-energy based features [146, 147]. NSS assume that pristine natural images have regular statistical properties which are disrupted when the image is impaired. Capturing this disruption can reveal the extent to which impairments are visible (and thus annoying) in the image. To do so, typically the image is transformed to a domain (e.g. DCT or wavelet), that better captures frequency or spatial changes due to impairments. The transform coefficients are then fit to a predefined distribution, and the fitting coefficients are taken as the NSS features.

Different NSS-based NR-IQMs have used various image representations to extract image statistical properties. In [33], for example, the NSS features were computed from the subband coefficients of an image's wavelet transform. Beside fitting a generalized Gaussian distribution to the subband coefficients, some correlation measures on the coefficients were also used in extracting the features. The study aimed at predicting the quality of images impaired by either JPEG or JPEG 2000 compression, white noise, Gaussian blur, or a Rayleigh fading channel. Saad et al. [34] computed NSS features with a similar procedure, but in the DCT domain. Mittal et al. [35] worked out NSS features in the spatial domain instead. They fitted a generalized Gaussian distribution on the image's normalized luminance values and their pairwise products along different orientations. In this case, the parameters of the fit were used directly as features. Another study in [148] took the Gradient Map (GM) of an image, and filtered it using Laplacian of Gaussian (LOG) filters. The GM and LOG channels of the image were then used to compute statistical features for the quality prediction task.

Lately, the IQM community has also picked up on the tendency of using learned, rather than handcrafted (e.g., NSS and free energy-based) features. A popular approach is to first learn (in an unsupervised way) a dictionary or codebook of image descriptors from a set of images. Using another set of images, the codebook will then be used as the basis for extracting features to learn a prediction model. To extract these features, an encoding step is performed on the image descriptors, followed by a pooling step. The study in [149] used this approach. The codebook was built based on normalized image patches and K-means clustering. To extract features for training and testing the model, a soft-assignment encoding was then performed, followed by max-pooling on the training and testing images. In [150], image patches underwent Gabor filtering before being used as descriptors to build the codebook. Hard assignment encoding was then performed, after which average pooling was used to extract the image features. To limit the computational burden yield by the large size of codebooks, a more recent study [143] proposed using a small sized codebook, built using K-means clustering based on normalized image patches. Smaller sized codebook usually decreases the prediction performance, and so to compensate for that, the study proposes to calculate the differences of high order statistics (mean, covariance and co-skewness) between the image patches and corresponding clusters as additional features.

Finally, the research on NR-IQMs has also recently started looking at features learned through convolutional neural networks (CNNs). CNNs [151] are multilayer neural networks which contain at least one convolutional layer. The network structure already includes parts that extract features from input images and a regression part to output a prediction for the corresponding input. The training process of this network not only optimizes the prediction model, but also the layers responsible for extracting representative features for the problem at hand. The study in [38] is an example of NR-IQMs using this approach, which brings promising results. However, one should be aware that, when learning features especially through CNNs, their interpretability is mostly lost. The high dimensionality of learnable CNN parameters also makes those features to be prone to overfitting of the training data, which is especially a risk when the size of data is small, as in the case of Image Quality Assessment databases (see more details in sec. 4.2.2).

The NR-IQMs described earlier, which are based on features representing perceptual changes in an image due to the presence of impairments, have higher interpretability and can still obtain acceptable accuracy. In this chapter, we aim at improving accuracy while maintaining interpretability. Therefore, we focus on this category of metrics and on enabling them to incorporate features that account for semantic content understanding.

Table 4.1: Properties of Several Publicly Available Image Quality Datasets

Dataset	Number of Images	Number of Reference Images	Impairment Types * Levels	Semantic Categories (of Reference Images)	
				Scene	Object
TID2013 [152]	3000	25	24 * 5	21 Outdoors 3 Indoors 1 N/A	7 Animate 14 Inanimate 1 N/A
CSIQ [153]	900	30	6 * 5	30 Outdoors 0 Indoors	13 Animate 17 Inanimate
LIVE [123]	982	29	5 * 6-8	28 Outdoors 1 Indoor	8 Animate 20 Inanimate
MMSPG HDR with JPEG XT [154]	240	20	3 * 4	12 Outdoors 6 Indoors 2 N/A	4 Animate 14 Inanimate 2 N/A
IRCCyN-IVC on Toyama [124]	224	14	2 * 7	14 Outdoors 0 Indoors	3 Animate 11 Inanimate
UFRJ Blurred Image DS [155]	585	N/A	N/A	412 Outdoor 173 Indoors	198 Animate 387 Inanimate
ChallengeDB [156]	1163	N/A	N/A	759 Outdoor 403 Indoors	321 Animate 842 Inanimate
SA-IQ	474	79	2 * 3	39 Outdoors 40 Indoors	25 Animate 54 Inanimate

4.2.2. SUBJECTIVE IMAGE QUALITY DATASETS

Over the years, the IQM community has developed a number of datasets for metric training and benchmarking. Such datasets usually consist of a set of reference (pristine) images, and a larger set of impaired images derived from the pristine ones. Impaired images are typically obtained by injecting different types of impairments (e.g., JPEG compression or blur) at different levels of strength. Each image is then associated with a subjective quality score, usually obtained from a subjective study conducted with a number of users. Individual user judgments of Quality are averaged per image across users into Mean Opinion Scores, which represent the quantity to be predicted by Image Quality Metrics.

Most Subjective Image Quality datasets are structured to have a large variance in terms of types and level of impairments, as well as perceptual characteristics of the reference images, such as Spatial Information or Colorfulness [157]. On the other hand, richness in semantic content of the reference images is often disregarded, nor information is provided on categories of objects and scenes represented there. This limits the understanding and assessment of image quality as it excludes users' higher-level interpretation of image content in their evaluation of quality. Table 4.1 gives an overview of the semantic diversity covered by several well-known and publicly available image datasets. The

semantic categorization follows that proposed by Li et al. in their work related to pre-attentional image content recognition [62] (note that these categories were not provided with the datasets and were manually annotated by the authors of this chapter).

From the table, we can see that most datasets do not have a balanced number of scene or object categories to allow for further investigation of the relationship between semantic categories and image quality. Two datasets are quite diverse in their semantic content: the UFRJ Blurred Image dataset [155], and the Wild Image Quality Challenge dataset [156]. On the other hand, these datasets lack structured information on the impairment types and levels of impairments present in the images. The images were collected "in the wild", meaning that they were collected in typical real-world settings with a complex mixture of multiple impairments, instead of being constructed in the lab by creating well-modeled impairments on pristine reference images.

These datasets were created to simulate the way impairments typically appear in consumer images. An impaired image in these datasets thus does not correspond to any reference image, and there is no clear framework to refer to in order to obtain information about how the impairments were added to the images. This makes it difficult to systematically look into the interplay between image semantics, impairments, and perceived quality.

In this work we propose a new dataset rich in semantic content diversity. We look at 79 reference images with contents covering different object and scene categories. These images are further impaired to obtain blur and JPEG compression artifacts at different levels. The proposed dataset SA-IQ can be seen as the last entry in Table 4.1, and we explain details of how the dataset was constructed in Section 4.3.

4.2.3. IMAGE SEMANTICS RECOGNITION

One of the most challenging problem in the field of computer vision has long been that of recognizing the semantic content of an image. A lot of effort has been put by the research community to improve image scene and object recognition performances: creating larger datasets [158], designing better features, and training more robust machines [159]. In the past five years, wider availability of high-performance computation machines and labelled data has allowed for the rise of Convolutional Neural Networks (CNNs) [151], and resulted in vast progress in the field of image semantic recognition.

One of the pioneering attempts of deploying CNNs for object recognition was AlexNet by Krizhevsky et al. [160]. Based on five convolutional and three fully connected layers, the AlexNet processes 224x224 images to map them into a 1000-dimensional vector, the elements of which represent the probability values that the input image belongs to any

of a thousand predefined object categories. Since AlexNet, current state-of-the-art systems include VGG [161], and GoogleNet [162]. For a more comprehensive overview of state-of-the-art systems, readers are referred to [159].

Along with object recognition, scene recognition has also had its share of rapid development with the advent of CNNs. One recently proposed trained CNN for scene recognition is the Places-CNN [163, 164]. This CNN is trained on the Places image dataset, which contains 2.5 millions images with a scene category label. 205 scene categories are defined in this dataset. The original Places-CNN was trained using similar architecture as the Alexnet mentioned above. Further improvements of the original Places-CNN were obtained by training on the VGG and GoogleNet architectures [164].

The implementation we use in this chapter is the PlacesVGG. The architecture has 13 convolutional layers, with four pooling layers among them, and a fifth pooling layer after the last convolutional layer. Three fully connected layers follow afterwards. The network outputs a 205-dimensional vector with elements representing the probability that the input image belongs to any of the 205 scene categories.

4.3. SEMANTIC-AWARE IMAGE QUALITY (SA-IQ) DATASET

As mentioned in Section 4.2, most publicly available image quality datasets do not cover a wide range of semantic categories. This limitation does not allow us to look deeper into how users evaluate image quality in relation with their interpretation of the semantic content category. For this reason, we created a new image quality dataset with not only a wider range of semantic categories included in it, but also a more even distribution of these categories. We describe our proposed dataset in the following subsections.

4.3.1. STIMULI

We selected 79 images that were 1024x768 in size from the LabelMe image annotation dataset [165]. The images were selected such that there was a balanced number of images belonging to each of the scene categories indoor, outdoor natural, and outdoor manmade, and within each scene category, enough number of animate and inanimate objects. Animate objects include humans and animals, whereas inanimate objects include objects in nature (e.g., body of water, trees, hill, sky) and objects that are manmade (e.g., buildings, cars, roads).

To have an unbiased annotation of the image categories, we asked five users to categorize the image scenes and objects. They were shown the pristine or unimpaired version of the images, and asked to assign the image to either of the three scene categories

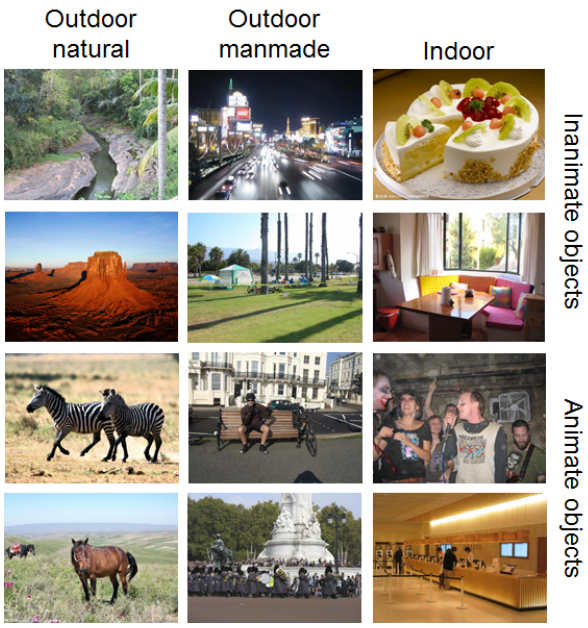


Figure 4.1: Examples of images in the SA-IQ dataset; the dataset contains images with indoor, outdoor natural, and outdoor manmade scenes, as well as animate and inanimate objects.

and either of the two object categories. The images were presented one at a time, and we did not restrict the time for users to view each image. Each image was then assigned the scene and object category which had the majority vote from the five users. In the end, we have 39 indoor images, 19 outdoor natural images, and 21 outdoor man-made images. In terms of object categories, we have 25 images with animate objects and 54 with inanimate objects. Figure 4.1 shows examples of the images in the dataset.

Image texture and luminance analysis. A possible concern in structuring a subjective quality dataset based on semantic, rather than perceptual, properties of the reference images is that certain semantic categories could include a majority of images with specific perceptual characteristics, and be more or less prone to visual alterations caused by the presence of impairments. For example, outdoor images may have higher luminance than indoor ones, and risk incurring luminance masking of artifacts. If that were the case, outdoor images would mask impairments better, thereby resulting in higher quality than indoor ones; this difference, though, would not be due to semantics.

Texture and luminance are two perceptual properties that are known to influence and possibly mask impairment visibility [166, 167]. We therefore verified that the images included in the dataset had similar texture and luminance levels across semantic

categories. Although this does not make our image set bias-free with respect to other possible perceptual characteristics, as luminance and texture play a major role in the visibility of artifact (and, consequently, on perceptual quality) [166, 167], this ensures that we rule out possible major effects of potential biases on our results so that we can ascribe differences in perceptual quality (in our study) to differences in semantics.

We used a modified version of Law's texture energy filter based on [16] to measure texture in horizontal and vertical directions. For each image, we computed the average mean and standard deviation of texture measures in both horizontal and vertical directions. Similarly, we used a weighted low-pass filter based on [16] to measure luminance in horizontal and vertical directions. We then calculated the average mean and standard deviation in both directions as our image luminance measure.

We compared the luminance and texture values of the images in the different scene categories using a one-way ANOVA. To compare the values across the different object categories, we used a T-Test. Our analysis showed that there is no significant difference in luminance or texture among the indoor, outdoor natural, and outdoor manmade images ($p>0.05$). Similarly, no significant difference was found for the two perceptual characteristics among the images belonging to animate or inanimate object categories ($p>0.05$). Hence, we can conclude that perceptual properties of the images are uniform across semantic categories.

Impairments. We impaired the 79 reference images with two different types of impairments, namely JPEG compression and Gaussian blur. We chose these two impairment types, as they are typically found in practical applications [168]. Moreover, most image quality assessment studies typically include these two impairment types, giving us the possibility to easily compare our results with previous studies. Of course, other types of impairments may be added in further studies. The impairments were introduced as follows.

1. *JPEG compression.* We impaired the original images through Matlab's implementation of JPEG compression. We set the image quality parameter Q to 30 and 15, to obtain images with visible artifacts of medium and high strength, respectively.

2. *Gaussian blur.* We applied Gaussian blur to the original images using Matlab's function *imgaussfilt*. To obtain images with visible artifacts of medium and high strength, the standard deviation parameter was set to 1.5 and 6, respectively. As for the choice of parameters for our JPEG compression, we also considered the parameters for our Gaussian blur to represent medium and low quality images.

Eventually, we obtained 316 impaired images. JPEG and blur images were then evaluated

in two separate subjective experiments.

4.3.2. SUBJECTIVE QUALITY ASSESSMENT OF JPEG IMAGES

To collect subjective quality scores for the JPEG compressed images, we conducted an experiment in a laboratory setting. 80 naive participants (28 of them were females) evaluated each 60 images. The 60 images were selected randomly from the whole set of 237 images (79 reference + 158 impaired), such that no image content would be seen twice by a participant, and at the end of the test rounds, we would obtain 20 ratings for each image.

The environmental conditions (*e.g.,* lighting and viewing distance) followed those recommended by the ITU in [52]. Images were shown in full resolution on a 23" Samsung display. At the beginning of each experiment session, participants went through a short training to familiarize themselves with the task and experiment interface. Participants were then shown the test images one at a time, in a randomized order, to avoid fatigue or learning effects in the responses. There was no viewing time restriction. Participants could indicate that they were ready to score the image by clicking on a button; this would make a discrete 5-point Absolute Category Rating (ACR) quality scale appear, on which they could express their judgment of perceived quality.

4.3.3. SUBJECTIVE QUALITY ASSESSMENT OF BLUR IMAGES

For the images impaired with Gaussian blur, we decided to conduct the experiments in a crowdsourcing environment. Crowdsourcing platforms such as AMTurk[1], Micorworkers [2] and Crowdflower [3] have become an interesting alternative environment to perform subjective tests as it is more cost and time-friendly compared with its lab counterpart. A consistent body of research has shown that crowdsourcing-based subjective testing can yield reliable results, as long as a number of precautions are taken to ensure that the scoring task is properly understood and carried out properly [56]. For example, evaluation sessions should be short (no longer than 5 minutes) and control questions (honey pots) should be included in the task to monitor the reliability of the execution.

We used Microworkers as the platform to recruit participants for our test. We randomly divided our 237 images into 5 groups consisting of 45-57 images each, such that we could set up 5 tasks/campaigns on Microworkers. Each campaign would take 10-15 minutes to complete. A user on Microworkers could only participate in one campaign

[1] http://mturk.com
[2] http://microworkers.com
[3] http://crowdflower.com

out of the five, and would be paid $0.40 for completing the campaign. To avoid misunderstanding of the task, and since our experiment was presented in English, we constrained our participants to those coming from countries with fluency in English. The aim here was to prevent users from misinterpreting the task instructions, which is known to impact task performance ([43, 169, 56]). Users were directed to our test page through a link in Microworkers. We obtained 337 participations over all of the campaigns.

Protocol. When a Microworkers user chose our task, s/he was directed to our test page, and shown instructions explaining the aim of the test (to rate image quality), and how to perform evaluations. To minimize the risk of users misunderstanding their task, we were careful to provide detailed instructions and training for our users. In the first part of our training session (as recommended by [56]), we gave a definition of what we intended as impaired images in the experiment (i.e., images with blur impairments). Example images of the worst and best quality that could be expected in the experiment were provided. Afterwards, participants were asked to rate an example image to get acquainted with the rating interface. The test started afterwards. Images were shown at random order, along with the rating scale at the bottom of the screen.

We used a continuous rating scale with 5-point ACR labels in this experiment. In [63], it was shown that both discrete 5-point ACR and continuous scale with 5-point ACR labels in crowdsourcing experiments would yield results with comparable reliability. We decided to use the continuous scale in this experiment, to give users more flexibility to move the rating scale. The continuous scale range was [0..100]. In our analysis, we will normalize the resulting mean opinion scores (MOS) into the range [1..5] using a linear normalization, so that we can easily compare the results on blurred images with those on JPEG images.

To help filter out unreliable participants, we included two control questions in the middle of the experiment. For these control questions, we would show a high quality image with a rating scale below it. After the user rates that image, a control question would appear, asking the users to indicate what they saw in the last image. A set of four options were given from which the users could select an answer.

4.3.4. DATA OVERVIEW AND RELIABILITY ANALYSIS

For the lab experiment on the JPEG images, we ended up with a total of 4618 ratings for the whole 237 images in the dataset after performing an outlier detection. One user was indicated as an outlier, and was thus removed for subsequent analysis. After this step, as customary, individual scores were pooled into Mean Opinion Scores (MOS) across participants per image, resulting in 237 MOS now provided along with the images.

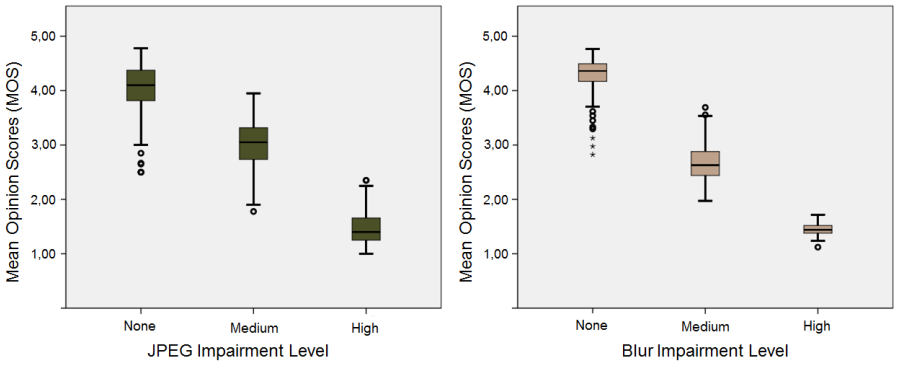


Figure 4.2: Overview of MOS across impairments for the two impairment types: Blur and JPEG compression

For the crowdsourcing experiments on blurred images, we first filtered out unreliable users based on incorrect answers to the content questions in the experiment, and incomplete task executions. We also performed outlier detection on the filtered data. From the 337 total responses that we received across all campaigns, we removed almost half of them due to incorrect answers to content questions, and failure to complete the whole task in one campaign. We did not find any outliers from the filtered data. In the end, we had 179 users whose responses were considered in our data analysis, with on average 37 individual scores per image. These were further pooled in MOS as described above. Figure 4.2 shows the collected MOS values across all impairment levels for the two impairment types: JPEG compression and Blur.

Given the diversity in the data collection method, and the concerns in terms of faithfulness of the evaluations obtained in crowdsourcing, we performed a reliability analysis. Our aim was to establish whether the obtained MOS were estimated with sufficient confidence, i.e., whether different participants expressed sufficiently similar evaluations for the same image. To do so, based on our and other previous work [142, 74], we chose the following measures to compare data reliability:

1. *SOS hypothesis alpha.* The SOS hypothesis was proposed in [98], and models the extent to which the standard deviation of opinion scores (SOS) changes with the mean opinion scores (MOS) values. This change is represented through a parameter α. A higher value of α would indicate higher disagreement among user scores. The hypothesis for an image i is expressed as in Eq. 4.1 below.

$$SOS^2(i) = \alpha(-MOS^2(i) + (V_1 + V_K)MOS(i) - V_1 V_K), \tag{4.1}$$

Table 4.2: SOS hypothesis alpha and average confidence interval (CI) across datasets

Dataset	Rating Methodology	Number of Ratings per Image	Experiment Environment	SOS Hypothesis Alpha	Average CI
SA-IQ (JPEG images)	discrete 5-point ACR scale	19-20	Lab	0.200	0.316
SA-IQ (Blur images)	continuous with 5-point ACR labels	37 on average	Crowdsourcing	0.2473	0.3182
CSIQ	multistimulus comparison by positioning a set of images on a scale	n/a	Lab	0.065	n/a
IRCCyN-IVC on Toyama	discrete 5-point ACR scale	27	Lab	0.1715	0.1680
UFRJ Blurred Image DS	continuous 5-point ACR scale	10-20	Lab	0.1680	0.5011
MMSPG HDR with JPEG XT	DSIS 1 [52], 5-grade impairment scale	24	Lab	0.201	0.273
ChallengeDB	continuous with 5-point ACR labels	175 on average	Crowdsourcing	0.1878	2.85 (100-point scale)
TID2013	tristimulus comparison	>30	Lab and Crowdsourcing	0.001	n/a

where V_1 and V_K indicate, respectively, the lowest and highest end of a rating scale.

2. *Average 95% confidence interval.* We calculate the average confidence interval over all images in a dataset to indicate user's average agreement on their ratings across the images. The confidence interval of an image i, rated by N users, is given as follows.

$$CI(i) = 1.96 * \frac{SOS(i)}{\sqrt{N}} \tag{4.2}$$

Table 4.2 gives a comparison of SOS hypothesis alpha and average CI values across different image quality studies and datasets. We also note in the table the different experiment setups used in the studies to construct the datasets. From the table, we see that the highest user agreement is obtained in studies that use comparison methods (i.e. double stimulus [52]) as their rating methodology. This was not a feasible option for us, as comparison methods on quite a large number of images as we have would be very costly. Nevertheless, our laboratory and crowdsourcing studies obtained highly comparable reliability measures. Moreover, our studies showed comparable reliability to that of other studies that also employ single stimulus rating methodology, and have the number of ratings per image proportionate to ours (i.e. the datasets IRCCyN-IVC on Toyama, UFRJ Blurred Image DS, and MMSPG HDR with JPEG XT as shown in Table 4.2).

4.3.5. EFFECT OF SEMANTICS ON VISUAL QUALITY

Having established that our collected data is reliable, we proceeded to check how semantic categories influence visual quality ratings at different levels and types of impair-

Table 4.3: Comparison of p-values for semantic category variables obtained through GLMM fitting

Impairment Type	**Independent Variables to Predict Visual Quality Ratings**	***p-value***
JPEG	Scene category	p=0.00
	Object category	p=0.00
	Scene category and object category (interaction of the two)	p=0.00
Blur	Scene category	p=0.015
	Object category	p=0.086
	Scene category and object category (interaction of the two)	p=0.00

ments. Perception studies have looked into the relation of scene versus objects with respect to human interpretation of image content. Questions such as whether users recognize scenes or objects first when looking at images have been asked and explained. In [62], it was found that even in pre-attentive stages, users do not have the tendency to recognize scenes or objects one faster than the other. Both are processed simultaneously to form an interpretation of the image content. Here, we attempt to check if one holds more significance than the other in influencing the user assessment of image quality.

Figures 4.3 and 4.4 show bar plots of the mean opinion scores (MOS) across impairment levels and semantic categories for JPEG and blurred images, respectively. From the plots, we see that images with no perceptible impairments are rated similarly in both cases: indoor images are rated more critically than outdoor images, and images with animate objects are rated more critically than those with inanimate objects. From the figures, we see that in the case of JPEG compressed images, this tendency of being more critical towards indoor images and images with animate objects continues for images with lower quality. However, the reverse seems to happen in the case of blurred images. It seems that with the presence of blur impairments, indoor images and images with animate objects are rated higher than other scene and object categories.

To check how semantic categories influence visual quality ratings, we fit a Generalized Linear Mixed Model (GLMM) to Visual Quality ratings, where semantic categories (scene and object) and impairment levels act as fixed factors, and users are considered as a random factor. Due to the different rating scale used to evaluate the two different impairment types, the model for JPEG images uses a multinomial distribution with logit link function, while that for blurred images uses a linear distribution with an identity link function. We use the following notation to describe the output of our statistical analysis. Next to each independent variable that we looked into, we indicate the degrees of freedom (*df1, df2*), the F-statistic evaluating the goodness of the model's fit (F), and the p-value representing the probability that the variable is not relevant to the model (p). A

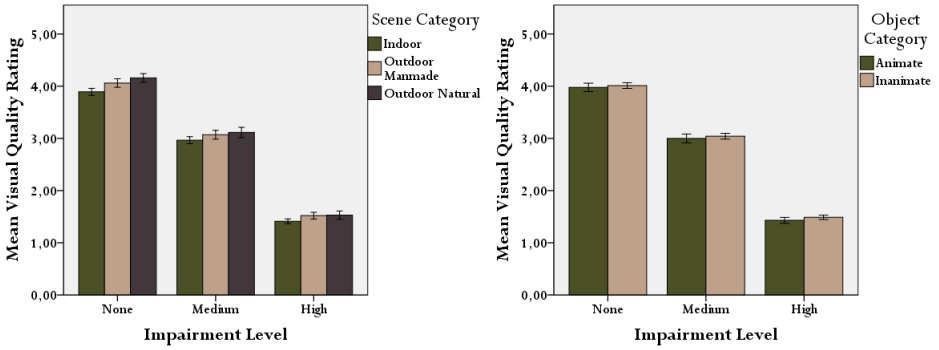


Figure 4.3: Bar plots of mean visual quality rating of JPEG compressed images across impairment levels and scene categories (right), and object categories (left)

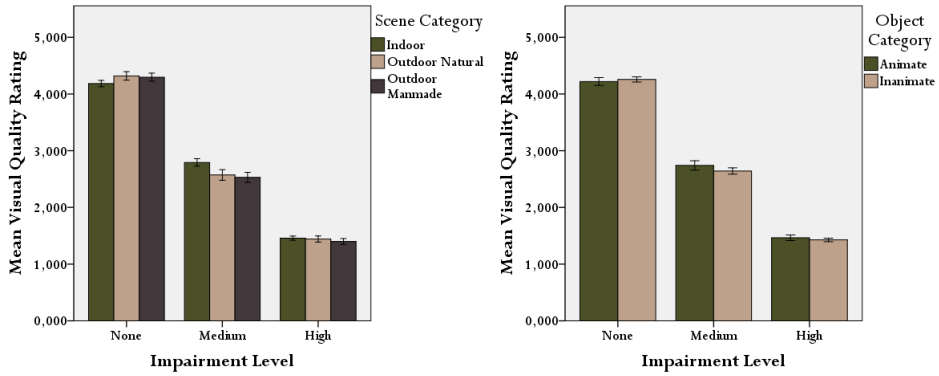


Figure 4.4: Bar plots of mean visual quality rating of blurred images across impairment levels and scene categories (left), and object categories (right)

p-value that is less than or equal to 0.05 indicates a statistically significant influence of a variable to predicting visual quality ratings.

For images with JPEG impairments, we find that all three independent variables, as well as the interaction of the three of them significantly influence user rating of visual quality (impairment level: *df1=2, df2=4.657, F=1193.54, p=0.00*; scene category: *df1=2, df2=4.657, F=28.35, p=0.00*; object category: *df1=1, df2=4.657, F=13.35, p=0.00*; impairment level*scene category*object category: *df1=6, df2=4.657, F=18.28, p=0.00*). This shows us that in judging images with JPEG compression impairments, users are significantly influenced by both scene and object category content.

Interestingly, for blurred images, a different conclusion is found. When we consider both scene and object categories, our model shows that scene category and impairment level has a significant effect on visual quality ratings, while object category only signif-

icantly influences visual quality rating in interaction with scene category and impairment level (impairment level: *df1=2, df2=8.717, F=1880.8, p=0.00*; scene category: *df1=2, df2=8,717, F=4.18, p=0.01*; scene category*impairment level: *df1=4, df2=8.717, F=9.74, p=0.00*; impairment level*scene category*object category: *df1=6, df2=4.657, F=6.722, p=0.00*). Unlike images with JPEG compression impairments, the visual quality rating of images with blur impairments are more significantly influenced by their scene category content than their object category content. For a clear overview of the *p-values* for the different (semantic category) independent variables, a summary is given in Table 4.3.

4.4. IMPROVING NR-IQMS USING SEMANTIC CATEGORY FEATURES

In this section, we show how the performance of no-reference image quality metrics can significantly improve when taking semantic category information into consideration. We do this by concatenating features that represent image semantic category (as extracted, for example, by large convolutional networks trained to detect objects and scenes in images) to perceptual quality features. Figure 4.5 illustrates this idea. A no-reference image quality metric (NR-IQM) typically consists of two building blocks [145]. The first is a feature extraction module, which produces a set of features that represent the image, as well as any artifacts present in it. The next block is the prediction or pooling module, which translates the set of features from the previous block into a quality score Q. In the following subsections, we compare the performance of image quality prediction when using only well-known perceptual quality features (condition 1 in Figure 4.5), with that of using a combination of perceptual quality features and semantic category features (condition 2 in the figure).

4.4.1. PERCEPTUAL AND SEMANTIC FEATURES FOR PREDICTION

We used the following perceptual quality features in our experiment:

1. *NSS features.*

   As mentioned in Section 4.2, NSS features are hand-crafted, and designed based on the assumption that the presence of impairments in images disrupts the regularity of an image's statistical properties. We used three different kinds of NSS features in our experiment, *BLIINDS* [34], *BRISQUE* [35], and *GM-LOG* [148]. These three metrics were chosen such that we would have NSS features extracted in different domains (DCT, spatial, and GM-LOG channels, respectively).

2. *Learned features.*

   We also chose to perform our experiment using learned features (codebook-based features). As these features are learned directly from image patches, it is possible that the features themselves already have semantic information embedded. It is therefore interesting to check how our approach would add to this type of metrics. We used HOSA features [143] to represent learned features in this chapter.

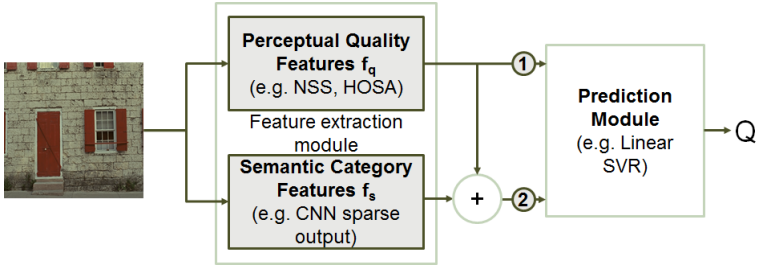


Figure 4.5: Block diagram of no-reference image quality metrics (NR-IQM): (1) using only perceptual quality features, and (2) using both perceptual quality features and semantic category features

To extract *semantic category features*, we fed the test images to the AlexNet [160] to obtain object category features, and to PlacesVGG [164] to obtain scene category features. We used the output of the last softmax layer of each CNN as our semantic category features. This led to a 1000-dimensional vector resulting from AlexNet, and a 205-dimensional vector resulting from PlacesVGG. Each element k in these vectors represents the probability that the corresponding image content depicts the k-th semantic category (scene or object). Each of these semantic category feature vectors would then be appended to the one containing the perceptual quality features. Adding object category features would result in an additional 1000-dimensional feature vector to the perceptual quality feature vector, while adding scene category features would result in an additional 205-dimensional feature vector to the perceptual quality features. Consequently, adding both to evaluate the benefit of considering jointly scene and object information in the IQM, would increase the feature count of 1205.

In Figure 4.6, we show heatmaps of the 1000-object category probability values that were output by AlexNet for two images with different levels of JPEG compression impairment. From the image, we can observe that most of the probability values of the 1000 object categories are very small. Given that quality prediction is a regression problem, we decided to use a sparse representation of these semantic feature vectors to improve on computational complexity. With a sparse representation, the number of non-zero multiplications to be performed by our regression model is significantly smaller, thereby

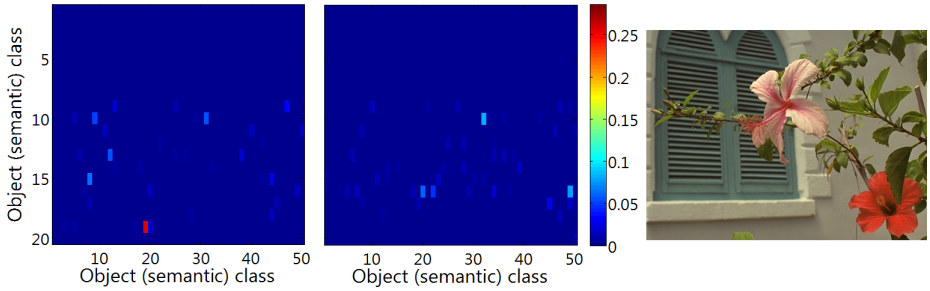


Figure 4.6: Heat map of probability values for the 1000 semantic classes output by AlexNet for two impaired images (with JPEG compression) taken from the TID2013 dataset, and the corresponding reference image on the right.

reducing the computation time. To make the semantic feature vector sparse, we set to zero the values of all but the top-N semantic categories in each vector.

In our previous study [142], we compared the performance of using only the top-10, 20, and 50 probability values in the object feature vector in addition to perceptual quality features. Our results showed no significant difference in performance among the three choices of N for top-N object category features. Given this result, we proceeded with using the top-20 object category features in subsequent experiments. In the next subsection, we investigate whether these results also hold for scene category features.

4.4.2. AUGMENTING NR-IQM WITH SEMANTICS

To investigate the added value of using semantic category information in NR-IQM, we first compared metrics with and without using semantic information in a simplified setting. We first concatenated the sparsified semantic feature vectors with *10, 20* and *50* top-N scene semantic features to the NSS and HOSA features described in the previous section. Then, we fed the perceptual + semantic feature vector to a prediction module as depicted in Figure 4.5. For the sake of comparison, we also added a condition with *N=0*, corresponding to not adding semantic features. This condition represents, for this specific test, our baseline.

With reference to Figure 4.5, we used the same prediction module: a Support Vector Regression (SVR) with a linear kernel. This means that here we discarded the prediction modules used in the original studies proposing the perceptual quality features (i.e. BLINDS uses a bayesian probabilistic inference module [34], BRISQUE and GM-LOG use linear SVR with RBF kernel [35, 148], white HOSA uses a linear kernel SVR [143]). This step is necessary to isolate the benefit that adding semantic information brings in terms of prediction accuracy: using different learning methods to implement the prediction

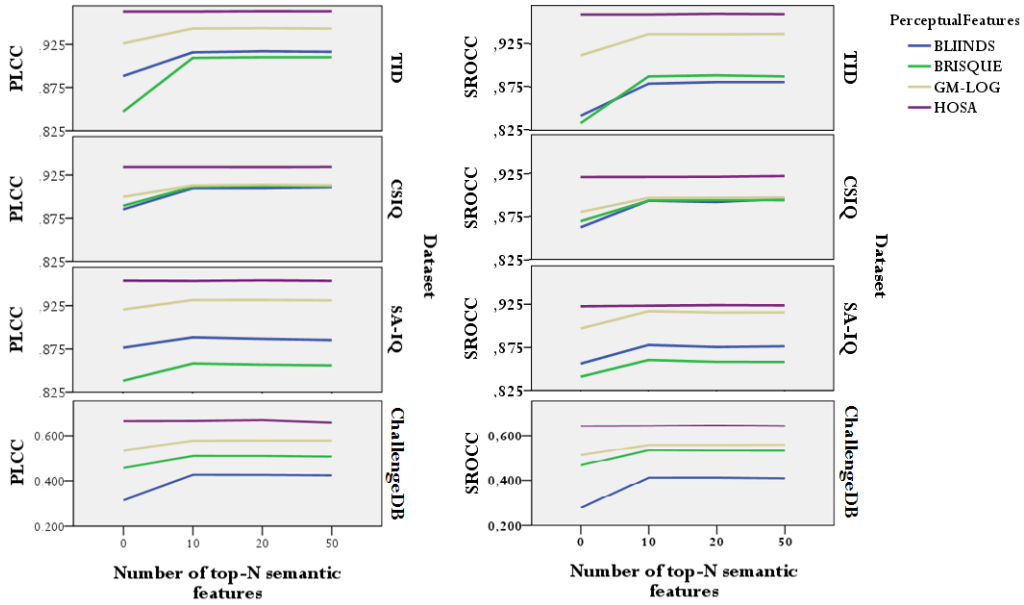


Figure 4.7: Impact of the number of top-N semantic categories condsidered in the IQM, in terms of Pearson and Spearman Correlation Coefficients (PLCC and SROCC respectively), between teh IQM prediction and the subjetive quality scores of different datasets. When the number of semantic features is 0, no semantic information is attached to the perceptual features, and the metric is calculated purely on perceptual feature information.

module would be a confounding factor for our result here.

We performed our experiments on four datasets, TID2013, CSIQ, our own SA-IQ, and ChallengeDB. The subjective scores of these datasets were collected in different experiment setups, e.g. display resolution, impairment types and viewing distance, such that our experiment results are not limited to images viewed in one particular setting. The TID2013 dataset [152] and CSIQ dataset [153] originally contains images with 5 to 24 different types of image impairments. As most perceptual quality metrics (including those used in this chapter) are constructed to evaluate images impaired with JPEG, JPEG2000 compression, additive white noise, and Gaussian blur, we limited our experiments to images with these impairments only.

The ChallengeDB dataset contains images with impairments present “in the wild” (typical real world consumption of images), and we included this dataset in our experiments to see how our approach would perform on said impairment condition. We used all the images in the ChallengeDB dataset in our experiment. We ran 1000-fold cross validation to train the SVR, with data partitioned into an 80%:20% training and testing set. The resulting median Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coeffient (SROCC) values between subjective and predicted quality

scores are reported in Figure 4.7 for performance evaluation.

In our previous work [142], we observed that the addition of object category features in combination with NSS perceptual quality features (BLIINDS, BRISQUE, and GM-LOG) improved the performance of quality prediction. These improvements in both PLCC and SROCC were statistically significant (T-test, $p < 0.05$). However, using object category features in combination with learned features (in this case, HOSA), did not bring significant added value. A similar result can be seen for the case of combining scene category features with perceptual quality features. In Figure 4.7, we see that for the NSS perceptual quality features, prediction performance increased with the addition of scene semantic categories. Conducting T-tests on the resulting PLCC and SROCC values showed that the improvements were statistically significant ($p < 0.05$). On the other hand, combining scene category features with HOSA features did not contribute to significant performance improvement. The average PLCC and SROCC values for the TID2013 dataset without scene features, for example, were 0.962 and 0.959, respectively, while the values when using scene features were 0.963 and 0.959, respectively.

A possible reason for the lack of improvement of the HOSA-based metric is that, unlike the handcrafted NSS features that specifically capture impairment visibility, HOSA features are learned directly from image patches. The features learned in this way may also capture semantic information, beside the impairment characteristics. Thus, the addition of semantic category features to these features may be redundant. Despite this observation, it is worth noting that the addition of semantic categories (either object or scene) could bring NSS-based models' performances close to that of HOSA while keeping the input dimensionality and thus model complexity lower (HOSA uses 14700 features, whereas NSS models use less than 100).

From the figure, we also notice that prediction performance did not change significantly among the *N=10, 20* and *50* for top-N scene features (further confirmed using one-way ANOVA, giving *p=0.05*). This applies for all four datasets and four perceptual quality metrics used in the experiment, and is aligned with our previous study on object category features [142].

4.4.3. FULL-STACK COMPARISON

In the previous section, we used a uniform prediction module (i.e. linear kernel SVR) across combinations of perceptual and semantic features to isolate the effect of adding semantic information on the performance of IQM. Referring once again to Figure 4.5, most image quality metrics in literature are optimized using a specific prediction module. For example, BLIINDS uses a Bayesian inference model, while BRISQUE and GM-

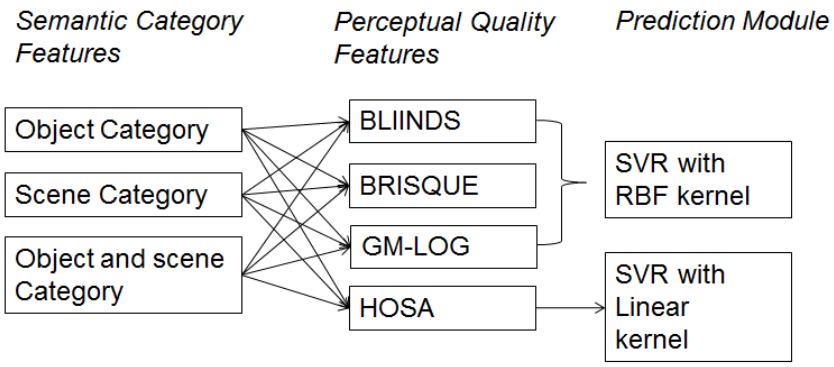


Figure 4.8: Features and prediction module combinations for blackbox comparison

LOG use SVR with an RBF kernel, and HOSA uses a linear kernel. In this subsection, we compare our approach of combining semantic category features with perceptual quality features within the metrics original implementations (i.e. using their proposed perceptual quality features along with their prediction module).

Figure 4.8 shows the semantic category feature combinations that we used in our experiments, along with the prediction module that we use for each perceptual quality feature. There are three types of semantic category features that we looked into: object category features, scene category features, and the combination of both. Each of these were combined with each of the four perceptual quality metrics, and trained using the corresponding learning method as shown in the table. We used RBF kernel SVR as learning method for the combination of semantic features with BLIINDS, BRISQUE and GM-LOG features. For the combination of semantic features with HOSA features, we used linear kernel SVR as our learning method.

As we used optimized prediction modules for each combination of features, we report here the performance of each original NR-IQM also when optimized for each dataset separately. The performance of the NSS metrics optimized for TID2013 and CSIQ that we report here are as per [148], while the HOSA metric performance optimized for the two datasets corresponds to that in [143]. For SA-IQ and ChallengeDB, we used grid search to optimize the SVR parameters of the four metrics. For performance evaluation, again we took the median PLCC and SROCC between the subjective and predicted quality scores across a 1000 folds cross-validation. Figure 4.9 gives an overview of the prediction performance for each feature combination on the four datasets TID2013, CSIQ, SA-IQ and ChallengeDB.

A look into the results on the TID2013 dataset reveals that the addition of semantic

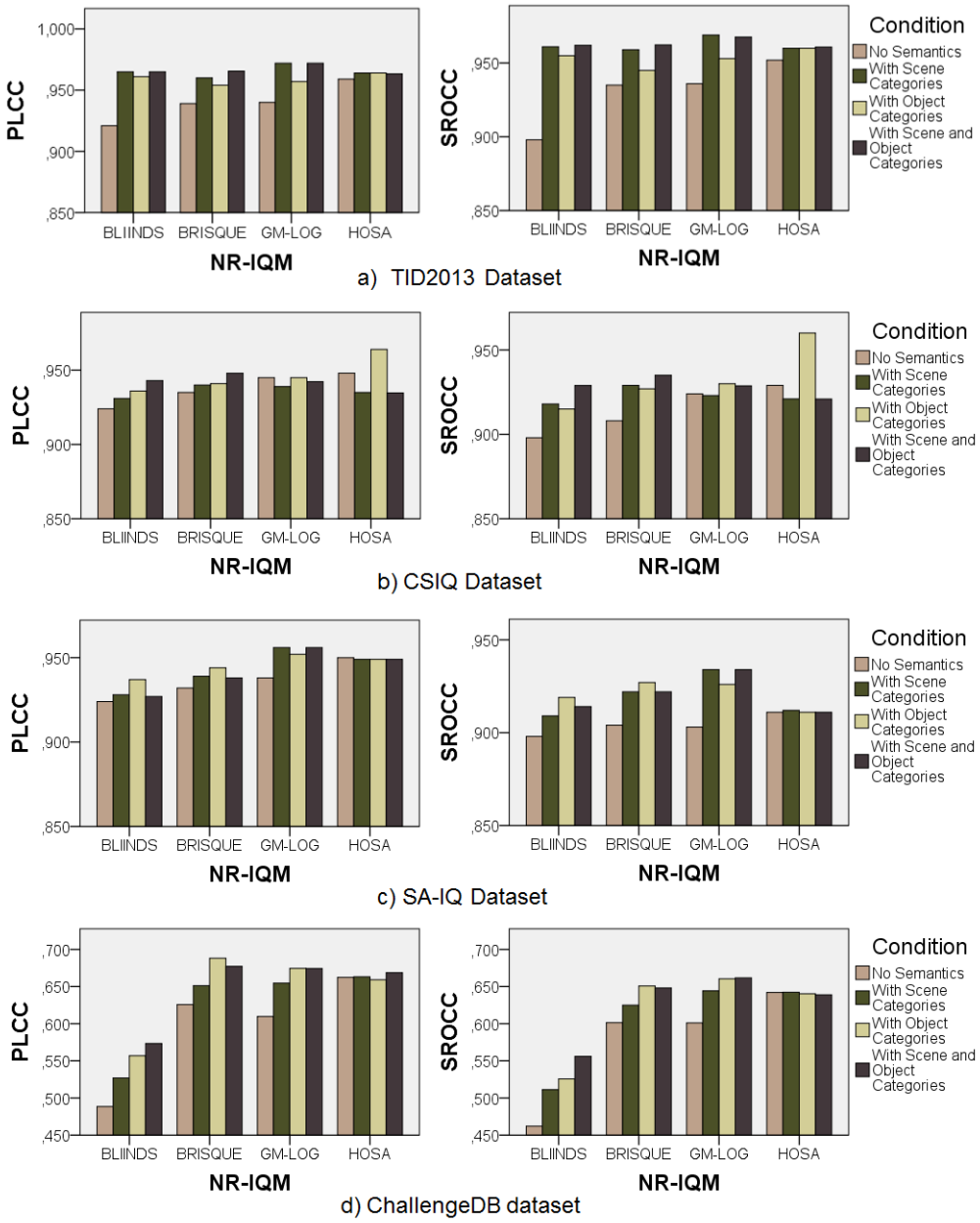


Figure 4.9: Full-stack comparison of the different NR-IQMs and semantic category feature combination on datasets TID2013, CSIQ, SA-IQ, and ChallengeDB

category features generally improved the performance of no-reference image quality as-

sessment. As expected, based on our observation in section 4.4.2, the NSS-based metrics showed larger improvement in predicting quality when combined with semantic category features. Nevertheless, the combination of semantic category features with learned features (HOSA) also improved prediction performance in this case.

Results on the CSIQ dataset showed improvement particularly when the perceptual quality features were combined with object category features. If we refer back to Table 4.1, which gives an overview of semantic categories spanned by the different datasets used in this work, we see that the CSIQ dataset does not have any variance in scene category (all images are outdoor images), whereas there seems to be more diversity in terms of objects. We argue that this could make object category features more discriminative than scene category features.

The figure further shows results on the SA-IQ dataset. We can see that adding semantic features results in a prediction improvement compared with only using NSS features. However, as also observed in Section 4.4, adding semantic features did not improve prediction performance for codebook-based features (*i.e.* HOSA). Furthermore, we also note that adding scene and object category features together did not result in higher prediction performance than when using only scene or only object category features.

Similarly for the ChallengeDB dataset, we observe improvement of quality prediction with the addition of semantic category features across the three NSS-based IQMs. On the other hand, the addition of semantic category features did not improve the performance of learning-based metric, HOSA, similar to our results for the TID2013, and SA-IQ datasets.

As mentioned briefly in Section 4.2, the four datasets that we use in our experiments were constructed through subjective experiments with different experiment setups, including viewing condition and type of impairments. For example, the TID2013 study suggested users to use a viewing distance from the monitor that is comfortable to them [152], while the CSIQ study maintained a fixed viewing distance from the monitor for all its participants [153]. All the datasets use different monitors and display resolutions in their studies. And while the datasets TID2013, CSIQ, and SA-IQ have images with one impairment type per image, the ChallengeDB dataset images contain multiple impairments per image. Considering these differences across the datasets, our results here and in Section 4.2 indicate that our proposed approach to improve NR-IQMs could be applied across multiple impairments and viewing conditions.

Performance with other type of perceptual quality features. So far, our experiment results show that the addition of semantic category features alongside perceptual quality

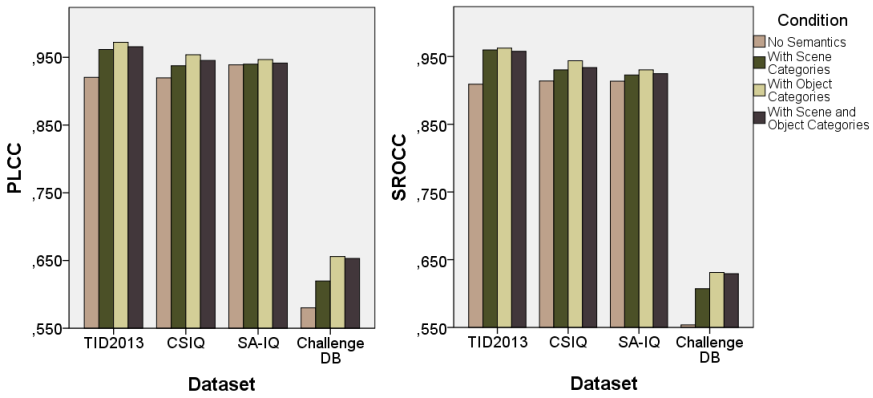


Figure 4.10: Full-stack comparison of the NFERM IQM and semantic category feature combination on datasets TID2013, CSIQ, SA-IQ, and ChallengeDB

features can improve the performance of quality prediction, especially for NR-IQMs with handcrafted (*i.e.* NSS-based) features. We would like to show here that these results still hold for NR-IQMs based on different types of handcrafted features, such as free energy-based features ([146, 147]).

We performed a full-stack comparison using the NFERM metric on the datasets TID, CSIQ, SA-IQ and ChallengeDB. We used grid search to optimize the prediction modules for each combination of features, including when no semantic feature is used. We show our results in figure 4.10, which plots the median PLCC and SROCC between the subjective and predicted quality scores across 1000 folds cross-validation. The figure shows that our previous results for NSS-based NR-IQMs still hold for non NSS-based NR-IQMs such as NFERM, that is, the addition of either scene or object category features, or both, helps improve the performance of blind image quality prediction

4.4.4. PERFORMANCE ON SPECIFIC IMPAIRMENT TYPES

In the previous experiments, we performed our evaluation on datasets consisting of different impairment types: JPEG and JPEG2000 compression, blur, and white noise in the TID2013 and CSIQ datasets, and JPEG and blur in the SA-IQ dataset. As shown through our analysis in section 4.3.5, semantic categories influence the assessment of visual quality in both JPEG compressed and blurred images, but in a different way. It is therefore interesting to look at the prediction performance on different impairment types individually. Our setup for this experiment is similar to that of Section 4.4.3, i.e. the SVR parameters of the NR-IQMs were optimized for evaluating each of the three datasets. The datasets were split into subsets with specific impairment types, and the prediction

models were re-trained for each impairment type. We again refer to [148] for the performance of NSS metrics optimized for TID2013 and CSIQ, and [143] for the HOSA metric performance.

Table 4.4: Comparison of the different NR-IQMs and semantic category features on different impairment types in the SA-IQ dataset

		BLIINDS	BLIINDS+S	BLIINDS+O	BRISQUE	BRISQUE+S	BRISQUE+O
SA-IQ	**JPEG**	0.8717	**0.8941**	**0.8938**	0.885	**0.9086**	**0.9084**
	BLUR	0.8925	**0.9093**	**0.9068**	0.9029	**0.9219**	**0.9222**
TID	**JPEG**	0.8853	**0.9383**	**0.9391**	0.9103	**0.9478**	**0.9530**
	JP2K	0.9118	**0.9591**	**0.9529**	0.9044	**0.9487**	**0.9504**
	BlUR	0.9176	**0.9665**	**0.9696**	0.9059	**0.9635**	**0.9696**
	WN	0.7314	**0.9417**	**0.9409**	0.8603	**0.9524**	**0.9509**
CSIQ	**JPEG**	0.9115	0.9052	**0.9300**	0.9253	**0.9342**	**0.9292**
	JP2K	0.8870	**0.9147**	**0.9416**	0.8934	**0.9056**	**0.9262**
	BlUR	0.9152	0.9003	0.9148	0.9143	0.8781	0.9018
	WN	0.8863	**0.9248**	**0.9246**	0.9310	**0.9398**	**0.9416**

		GM-LOG	GM-LOG+S	GM-LOG+O	HOSA	HOSA+S	HOSA+O
SA-IQ	**JPEG**	0.8843	**0.9218**	**0.9099**	0.9149	0.9140	0.9151
	BLUR	0.9048	**0.9262**	**0.9228**	0.9029	0.9034	0.9030
TID	**JPEG**	0.9338	**0.9478**	**0.9403**	0.9283	0.9288	0.9271
	JP2K	0.9263	**0.9539**	**0.9548**	0.9453	0.9283	0.9265
	BlUR	0.8812	**0.9635**	**0.9604**	0.9538	**0.9604**	0.9562
	WN	0.9068	**0.9513**	**0.9524**	0.9215	0.9273	0.9243
CSIQ	**JPEG**	0.9328	0.8927	0.9220	0.9254	0.9062	0.9071
	JP2K	0.9172	**0.9249**	**0.9316**	0.9244	0.9032	0.9036
	BlUR	0.9070	0.8752	0.8969	0.9266	0.8848	0.9037
	WN	0.9406	0.9342	0.9237	0.9192	**0.9232**	0.9038

Table 4.4 shows the results of our experiments. We report only the SROCC values due to limited space, however we note here that the resulting PLCC values yielded similar conclusions. The bold numbers in the table indicate the conditions in which the prediction performance improved with the addition of semantic category features. From the table, we see that the addition of semantic category features, whether they are scene

or object features, improved significantly the performance of NSS-based no-reference metrics on all impairment types presented for the SA-IQ and TID datasets. However, for the CSIQ dataset, only images with JP2K compression and white noise impairment consistently showed similar improvement. It is interesting to note that the improvement in performance was not significantly different between the addition of object and scene categories. For the codebook-based metric, HOSA, as we have seen in the previous sections, we again observe that the addition of semantic category features did not bring improvement, even for specific impairment types, on any of the three datasets.

4.5. IMAGE UTILITY AND SEMANTIC CATEGORIES

Image quality has often been associated with image usefulness or utility. Nevertheless, studies have shown that perceived utility does not linearly relate to perceived quality [141]. In this section, we show that bias on image content category can influence utility and perceived quality differently, and thus further confirm that an image usefulness cannot always explain perceived image quality. We do this by comparing the relationship between image semantic categories and image utility with the relationship between image semantic categories and image quality. We perform this comparison on our image dataset, SA-IQ.

To perform the comparison, we calculated image utility scores for each image in the dataset. We refer to [170] for image utility metric NICE. The metric calculates image utility based on image contour. For every image, we used an edge detection algorithm (e.g., Canny) to obtain the binary of the test image and its reference, which we denote as B_T and B_R, respectively. We then performed a morphological dilation on the two binary images using a 3x3 plus-sign shaped structural element. We further assumed that the result of this morphological dilation is I_R for the reference image and I_T for the test image. We then obtained the utility score NICE for the image by taking the Hamming distance of I_R and I_T, and dividing it by the number of non-zero elements in B_R, to account for the variability of contours across the reference images. The utility metric NICE gives an estimation of how degraded an image's contours have become due to impairments compared with its reference, and is thus inversely related with image utility.

In Figures 4.11 and 4.12 we show plots of perceived quality mean opinion scores (MOS) against NICE utility scores for JPEG compressed and blurred images in our datasets. If we compare our plots with the perceived utility vs. perceived quality plot found in [141], we can observe that our blurred images span the lower range of image quality and higher range of image quality, in which utility doesn't grow or change with the change

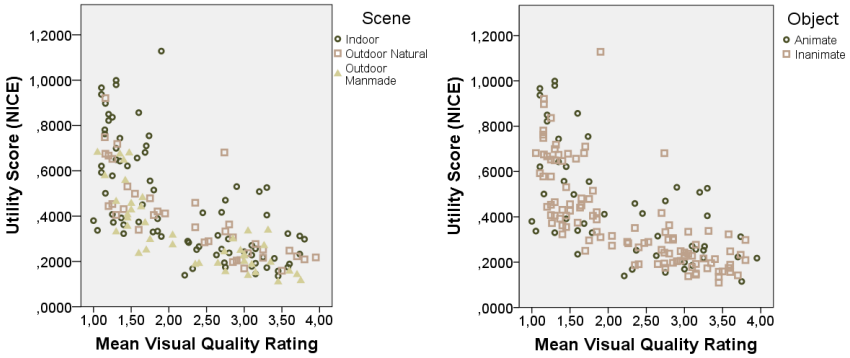


Figure 4.11: Image utility vs. quality scores of JPEG images across semantic categories (left: scene categories, right: object categories

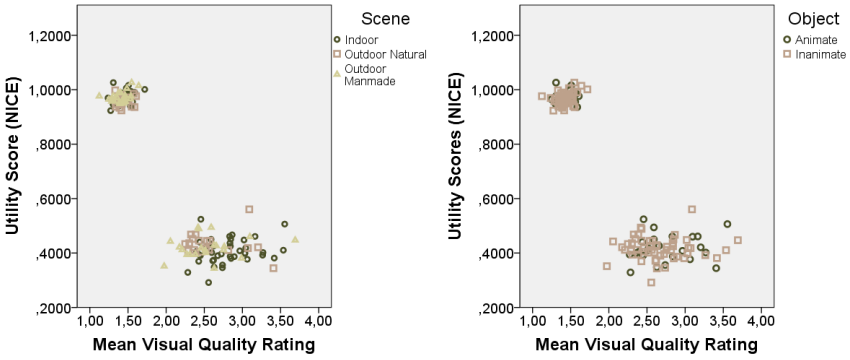


Figure 4.12: Image utility vs. quality scores of blurred images across semantic categories (left: scene categories, right: object categories

of perceived quality. However, our JPEG images seem to span a middle-range quality, in which perceived quality has a linear relationship with perceived utility. In general, we can see that our data represented the different relationships between perceived quality and utility across the range of quality.

We ran K-means on the blurred and JPEG image data, to isolate the different clusters as shown in the plots, and conducted statistical analysis to check how semantic categories influence utility and quality in these clusters. We set the number of clusters k to two for both the blurred and JPEG data. We then performed several one-way ANOVA for each cluster. Specifically, we first conducted one-way ANOVAs with semantic categories (either scene or object categories) as independent variables, and utility as dependent variables. Similarly, we then conducted one-way ANOVAS with quality MOS as dependent variables instead of utility.

Table 4.5: Significance level of semantic categories' influence on image utility and quality across Blurred and JPEG image clusters

Impairment Type	Image Cluster	Semantics on Utility		Semantics on Quality	
		Scene	Object	Scene	Object
BLUR	HQ cluster	p = 0.098	p = 0.971	**p = 0.009**	p = 0.324
	LQ cluster	p = 0.054	p = 0.469	p = 0.177	p = 0.228
JPEG	HQ cluster	**p = 0.03**	**p = 0.049**	p = 0.851	p = 0.866
	LQ cluster	**p = 0.003**	p = 0.219	p = 0.307	p = 0.365

Table 4.5 shows the results of our analysis. We label the two clusters for each image sets as HQ for clusters with images having higher quality range, and LQ for clusters with images having lower quality range. The numbers in bold indicate cases in which semantics has a significant influence on either utility or quality. From the table, we can see that semantic categories influence image utility and quality differently. Moreover, the influence of semantics on utility seems to be more significant in JPEG images than in blurred images.

4.6. CONCLUSION

In this chapter, we showed that an image's semantic category information can be used to improve its quality prediction to align better with human perception. Through subjective experiments, we first observed that an image's scene and object categories influence users' judgment of visual quality for JPEG compressed and blurred images. We then performed experiments on different types of no-reference image quality metrics (NR-IQMs), and showed that blind/no-reference image quality predictions can be improved by incorporating semantic category features into our prediction model. This applied across different image quality datasets representing diverse viewing conditions (e.g. display resolution, viewing distance), and image impairments, including multiple impairments. We also provided a comparison of how semantics influences image utility and image quality, and conclude that semantics has more significant influence on image utility for JPEG images than for blurred images.

Another contribution of this chapter is a new image quality dataset, SA-IQ, consisting of images spanning a wide range of scene and object categories, with subjective scores on JPEG compressed and blurred images. The dataset can be accessed through `http://ii.tudelft.nl/iqlab/SA-IQ.html`. Future work on these findings would include looking into better representations or methods to combine semantic information and perceptual quality features in NR-IQMs.

5

FULL REFERENCE POINT CLOUD QUALITY METRICS BASED ON COLOR DISTRIBUTION

This chapter continues our efforts in incorporating new influencing factors into existing metrics to improve QoE predictions. Current point cloud quality metrics still do not align well with human judgment of point cloud compression quality. In this chapter, we propose a new objective metric for point clouds using the insights that we obtained through subjective experiments in Chapter 3. We propose to use color-based features in our metric, and show that our approach improves state-of-the art point cloud quality metrics.

This chapter is based on: E. Siahaan, J.A. Redi, A. Hanjalic, P. Cesar, *Full-reference quality metrics for point cloud compression based on color distribution*, Under Review, (2018)

5.1. INTRODUCTION

Point cloud has been gaining attention as one of the technologies used for representing 3D scenes and objects in virtual immersive systems (see Figure 5.1). Compared to 3D meshes, which is another approach for representing 3D scenes and objects, point cloud does not require heavy pre-processing (e.g. to compute triangulation), and is more resilient to noise. An indication of point cloud's rise in popularity is the recent international standardization efforts that are being made for its compression [171, 172].

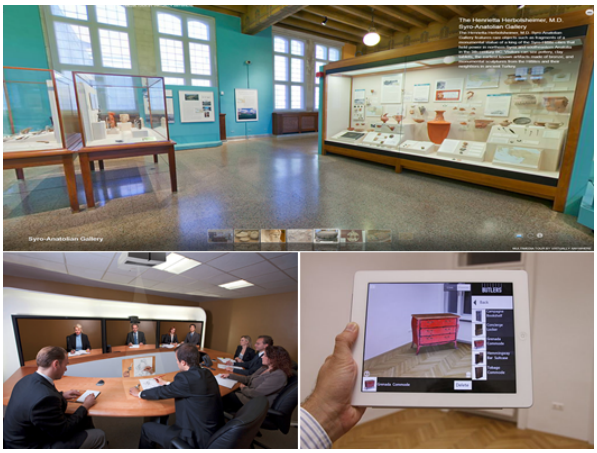


Figure 5.1: Examples of immersive systems (top to bottom, counter-clockwise): virtual museum[1], telepresence system[2], and augmented reality systems[2].

A point cloud consists of points in a three-dimensional space, that together make up the surface of an object. These points are acquired using depth sensors or multiple cameras. Typically, each point in a point cloud would contain its coordinate values in the 3D space, but it could also have additional attributes such as color value, transparency, or normal. Despite its simplicity compared to 3D meshes, point clouds may still demand high resources. For example, a high quality point cloud of a human object can contain up to $9 * 10^5$ points, taking 18 MB of space.

Various point cloud compression technology have been proposed until now ([103, 104, 105, 106, 107]). As is common in visual QoE studies, one of the ways to evaluate compression technology is using quality metrics, to observe whether or not the compressed image or video satisfies consumers' or users' expectations for viewing quality.

[1] Screenshot taken from virtual tour of the Oriental Institute Museum (https://oi.uchicago.edu/virtualtour)
[2] Images taken from Wikimedia Commons.

Not many studies have looked into the quality assessment of point cloud compression ([60, 108, 107]), and as we will describe below, there are still research gaps related to the objective assessment of point cloud quality that aligns with human assessment.

In one study, Alexiou et al. ([60]) performed subjective experiments to compare the quality evaluation of Gaussian noise and octree-pruning (compression-like) impairments on point clouds. The same authors then performed a similar study using a head-mounted display for viewing [108]. Other studies on point cloud quality assessment were performed by Javaheri et al. ([20]) and Zhang et al. [118], though these were to evaluate the quality of point clouds impaired with noise. Noise artifacts differ from compression artifacts, and thus both may affect users' perception of quality in a different way. Compression artifacts create the appearance of fewer points in an object, while noise artifacts result in some points being placed far from their correct position.

The studies above showed that state-of-the-art objective quality metrics do not yield scores that correlate well with human judgment for point cloud compression. However, no metric was proposed to improve the existing ones. Another limitation of the previous studies above is that their focus has mainly been on non-colored geometric point cloud models ([60]-[20]). These point clouds do not represent realistic use cases or application context, for example tele-immersive systems, gaming systems, virtual learning environment, etc.

This chapter aims to fill in the research gaps indicated above by proposing a new objective quality metric for point cloud compression based on a realistic point cloud dataset. We use the insights obtained in Chapter 3 of this thesis to design our new quality metric. In Chapter 3, we performed subjective experiments to explore users' perception of point cloud QoE using a realistic point cloud dataset. Our experiments showed us that users take notice of the clarity of the face and overall structure of the body while evaluating point cloud quality, as well as color alterations or distortions. As state-of-the-art point cloud quality metrics still rely on quantifying geometry/structure distortions only [173, 174], we propose to incorporate color features into our metric. Our results show that our proposed metric is better-aligned with human judgment compared to state-of-the-art metrics.

This chapter is organized as follows. In the next section, we discuss some background work related to our study. In Section 5.3, we describe our proposed objective quality metric, and present experiments on the objective quality metric. In Section 5.4, we discuss the main findings of this chapter and their implications. Finally, section 5.5 presents the conclusion of this chapter.

5.2. BACKGROUND

In this section, we will first present some background on point cloud compression technology. Afterwards, we give an overview of related work on the objective quality assessment of point clouds.

5.2.1. POINT CLOUD COMPRESSION

The use of point clouds to represent 3D models is a good alternative to polygonal meshes as they do not store any geometry or topology specifics. This makes point cloud especially appealing for real-time application scenarios. Nevertheless, millions or billions of points are needed to create a realistic point cloud 3D model. Thus, efficient point cloud compression strategies have been proposed in literature.

Point cloud compression can be categorized as either fixed-rate or progressive, with the latter being useful in adapting to available bandwidth and rendering capabilities. In [104], a framework for progressive point cloud coding was proposed based on octree structure. Their strategy boasts a more efficient bit rearrangement in the octree subdivision, to encode the position, normal and color attributes of points in 3D objects with arbitrary topology. Another study proposed a method to compress time-varying point clouds by exploiting temporal redundancies in point cloud streams [105]. The study used a modified octree data structure, which allows the method to detect and perform differential encoding of spatial changes between temporarily adjacent point clouds. In [106], the authors proposed a motion estimation and compensation algorithm for time-varying point cloud compression, based on a graph representation of the 3D models' geometry and color.

In this chapter, we use the solution proposed by Mekuria et al. [107] to perform the point cloud compression used and evaluated in our study. This work was selected as the anchor for the recent call for proposals for point cloud compression under MPEG (JTC1/SC29/WG11). The study combines octree-based compression schemes with common video coding schemes. To compress a point cloud image or frame, at the first stage, outlier filtering is performed and bounding box is computed for the point cloud. With the resulting bounding box as root, the encoder recursively subdivides the point cloud into eight children, represented as an occupancy code. This results in an octree data structure. Following the approach taken in [105], the occupancy codes are coded using an entropy coder. Finally, the point cloud color attributes are coded by mapping the color attribute from the octree traversal graph to a JPEG image grid.

5.2.2. POINT CLOUD OBJECTIVE QUALITY ASSESSMENT

Quality assessment is essential in the production and delivery of multimedia. Media may go through different processing algorithms that can alter their quality at any point between their acquisition until they are displayed to users. A quality assessment would allow for appropriate decision making or correction for the media at its current stage of production or delivery. Ideally, a quality assessment would rely fully on human feedback (subjective assessment), as humans are ultimately the final consumer of the media. However, human feedback is time consuming and costly. Therefore, a large part of research on quality assessment has looked into finding objective metrics that closely mimic the way humans evaluate quality. This is done by asking a sample of the population (users) to view some media and rate their quality. The scores collected from these users are then utilized to evaluate an objective metric.

So far, existing quality metrics for point clouds measure the similarity of the structure (geometry) between a reference point cloud and a test point cloud. The metrics could be categorized into either point-to-point, or point-to-plane comparison approaches [173, 174]. Point-to-point metrics measure the (symmetric) distance between each point in the degraded point cloud and the corresponding point in the reference point cloud. Point-to-plane measures the distance between a decoded point and the corresponding (estimated) object surface in the reference point cloud. In either point-to-point or point-to-plance approaches, the distances can be measured using the root mean square distance (referred to as $dist_{RMS}$ in this chapter, and Haussdorf distance (referred to as $dist_{Hauss.}$ in this chapter) [173].

Studies on point cloud quality assessment have mostly focused on comparing the performance of state-of-the-art quality metrics with human/subjective assessments. Javaheri et al. looked into evaluating point cloud denoising algorithms on point cloud geometry, and concluded that the point-to-plane metric and the $dist_{RMS}$ seem to correlate better with human judgment than the point-to-point metric or $dist_{Hauss.}$ in evaluating noisy point clouds [20]. Another study performed an interactive subjective study on point cloud geometry impaired with Gaussian noise and octree-pruning to create compression-like impairments [60]. The study showed that state-of-the-art objective metrics correlate well with perceptual quality of noise-impaired point clouds, but fail to predict the perceptual quality of compression-like impairments. The authors then performed another subjective evaluation of point clouds, but in an augmented reality (AR) setup, and came to similar conclusions as before [108]. They also showed that users' judgment for compression-like impairments seems to be influenced by the underlying surface and shape of the point cloud content.

Table 5.1: Comparison of using bin-to-bin distance measure (Euclidean distance) and cross-bin distance measure (QF distance) for the proposed color histogram-based metric. Subscripts h and l indicate hue and luminance, respectively.

Distance measure	$D_{hist(h)}$		$D_{hist(l)}$	
	PLCC	SROCC	PLCC	SROCC
Euclidian distance	0.4264	0.3853	0.9133	0.8764
QF distance	0.9055	0.8915	0.9130	0.8841

Based on the above studies, it is shown that current metrics still correlate poorly with subjective assessment of compressed point clouds. However, no metric has been proposed to improve the state-of-the-art. Moreover, these studies have focused mainly on point cloud geometry, which users may perceive differently from more complex point clouds.

5.3. OBJECTIVE QUALITY ASSESSMENT FOR POINT CLOUD COMPRESSION

In Chapter 3, we performed subjective experiments to explore users' perception of point cloud quality. Our findings in Chapter 3 suggested that users take notice of the clarity of the face and overall structure of the body while evaluating point cloud quality, as well as color alterations or distortions. Considering the use of color point clouds in real-world applications, and given the above observations in our subjective experiments, it is important to also evaluate perceived point cloud quality based on colors. In addition, different point cloud compression algorithms may use different color coding schemes, and thus create different color artifacts. Point cloud quality evaluation based on color has not been explored in previous studies as most of the stimuli used were not colored. For this reason, we look into objectively quantifying the quality of compressed point clouds based on their color properties, and later on show how it improves overall quality prediction together with geometry-based metrics. We consider only point-to-point metrics in our evaluations, as it reduces the need to estimate the object surface.

5.3.1. POINT CLOUD QUALITY METRIC BASED ON COLOR HISTOGRAM AND AUTOCORRELOGRAM

To evaluate the quality of a point cloud, we propose to compare its color distribution information with that of its reference. Luminance, or the perceived brightness of color,

has often been used in image quality metrics to estimate users' perception of color [123]. However, some studies have also suggested the use of hue or chrominance as relevant to quality perception [175]. We perform our experiments using both luminance and hue values of point clouds, and compare the performance of both in predicting point cloud quality. The luminance value is represented by the Y-channel values of the YUV space, while hue is represented by the H-channel values of the HSV space.

Consider a reference point cloud PC_r and its degraded version PC_d. Let $H_{r(l,h)}$ and $H_{d(l,h)}$ be the histogram of either the luminance or hue values of the reference and degraded point clouds, respectively. We measure the quality of the degraded point cloud as the distance between the two histograms $H_{r(l,h)}$ and $H_{d(l,h)}$. Different approaches can be used to compute the difference or distance between two histograms, such as bin-to-bin measurements (e.g. Euclidean distance), or cross-bin measurements (e.g. Quadratic Form/QF distance).

An Euclidean distance between two color histograms is calculated as follows.

$$(Euclidean)D_{hist(l,h)} = \sqrt{\sum^{n}(H_{r(l,h)} - H_{d(l,h)})^2}. \tag{5.1}$$

Meanwhile, the Quadratic Form (QF) distance between two color histograms is calculated as follows.

$$(QF)D_{hist(l,h)} = \sqrt{(H_{r(l,h)} - H_{d(l,h)})^T * A * (H_{d(l,h)} - H_{r(l,h)})}, \tag{5.2}$$

where A represents the similarity matrix which elements are defined as:

$$A(i,j) = 1 - exp(-\alpha * \frac{min(|i-j|, n-|i-j|)}{max(|i-j|, n-|i-j|)}), \tag{5.3}$$

where i, j are the indices of matrix A, n is the number of bins in $H_{r(l,h)}$ and $H_{d(l,h)}$, and $i, j \in n$. As suggested in a study on perceptual color difference [176], we define $\alpha = 32$. Table 5.1 shows a comparison of using the two different distance metrics for our color histogram-based metric.

Measuring the distance of color histograms provides an indication of global alterations in colors due to color coding mechanisms in compression. However, this does not give an indication of the spatial correlation among colors within smaller distances or regions in the point cloud. To check whether or not the use of local spatial correlations would improve the quality estimation, we adopt the concept of color correlogram [177].

A correlogram is represented by a matrix of *nxn* elements, where n is the number of bins used to represent the luminance or hue values. n is usually 256, when color channels are represented by 8 bits. In our experiments, we use *n=256* for both cases. Given a

Table 5.2: Performance comparison using different k-nearest neighbors as distance constraint in autocorrelograms (using Euclidean distance to compare the autocorrelograms of degraded and reference point clouds). Subscripts h and l indicate hue and luminance, respectively.

k	$D_{C(h)}$		$D_{C(l)}$	
	PLCC	SROCC	PLCC	SROCC
4	0.1614	0.2445	**0.817**	**0.7915**
6	0.1156	0.2205	0.7722	0.7522
8	0.0913	0.2214	0.7411	0.7379
10	0.0782	0.2027	0.7162	0.7379

point p in a point cloud, and its neighboring point p_k, an element of the correlogram for point p at index i, j is defined as follows.

$$C_p(i,j) = Pr\left[(c_p \in i, c_{p_k} \in j) \,|\, (|p_k - p| <= d)\right], \tag{5.4}$$

where $Pr[x]$ denotes probability of x, c_{p_k} is the color bin value of point p_k, and point p_k is within distance d to point p. Therefore, the correlogram of a point p with color bin value $c_p \in i$ depicts the probability that a point with color bin value $c_{p_k} \in j$ would be within distance d. We set the distance d to be within the k nearest neighbors of the point p. By its definition, a correlogram is a symmetric matrix.

To obtain the correlogram of a point cloud image PC with N number of points, we sum over the correlograms of every point as follows.

$$Crl_{PC} = \sum_{n=1}^{N} Cp_n \tag{5.5}$$

To compare the quality between two point clouds, we take the autocorrelograms of the point clouds, and calculate their distances. The autocorrelogram of a point cloud PC is the diagonal elements of its correlogram, as follows.

$$AC_{PC} = diag(Crl_{PC}) \tag{5.6}$$

As with the color histogram-based metric, we perform a comparison of using Euclidean distance and QF distance for our autocorrelogram-based metric. We first compare the performance of using different values of k nearest neighbors as shown in Table 5.2 (with the Euclidean distance). Based on this table, we decide to use *k=4* in our metrics. Table 5.3 shows the comparison of performance using Euclidean distance and QF distance on our autocorrelogram-based metric.

Table 5.3: Comparison of using Euclidean distance measure and QF-distance measure for the autocorrelogram-based metric (k=4). Subscripts h and l indicate hue and luminance, respectively.

Distance measure	$D_{C(h)}$		$D_{C(l)}$	
	PLCC	SROCC	PLCC	SROCC
Euclidian distance	0.1614	0.2445	0.817	0.7915
QF distance	**0.8981**	**0.8688**	**0.8654**	**0.8281**

Figure 5.2: Point cloud sequences used in the experiment, from the publicly available 8i Voxelized Full Bodies (8iVFB v2) dataset [116].

5.3.2. METRIC VALIDATION

To evaluate the performance of our metrics, we calculate the objective quality scores for 24 point cloud stimuli that we generated for our subjective experiments in Chapter 3. We use 6 frames from 6 different sequences of full-body humans from [116], and compress each of these frames into 4 different levels of compression using the compression algorithm in [107]. The compression parameters used are Level of Detail (LoD) 10, LoD 9, LoD 8, and LoD 7. LoD 10 means that the compression uses a 10-b octree setting, i.e., 10 quantization bits per direction. Figure 5.2 shows the 6 reference frames of our dataset. After calculating the objective scores of our stimuli, we compare them with the mean opinion scores (MOS) collected from our subjective study in Chapter 3 using the Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation

Coefficient (SROCC).

Tables 5.1 and 5.3 show the performance of our metrics using Euclidean distance and QF distance. Table 5.1 shows that using cross-bin distance (i.e. QF distance) on hue histograms yields a better estimation of quality than using bin-to-bin distance. On the other hand, there is no significant difference between the two distance measures for luminance histograms. Thus, we use the QF distance when comparing our color histogram-based metric with state-of-the-art metrics later on. Table 5.3 also shows that the QF distance gives a better estimation of quality than Euclidean distance for autocorrelograms.

We compare the results of our proposed approaches with currently used point-to-point objective metrics for point clouds. Besides comparing our approach with the metrics for geometric distortion, we decide to use the peak-signal-to-noise ratio (PSNR) of the color channels in YUV-space as a baseline for color distortion metrics. We also calculate the PSNR of the H-channel in HSV space to once again compare the use of hue against luminance in estimating point cloud quality. Table 5.4 shows the PLCC and SROCC values computed for every quality metric against the MOS collected for the 24 stimuli in our dataset.

The table shows that the geometric distance metrics, $dist_{RMS}$ and $dist_{Hauss}$ correlate well with human judgment for our stimuli and setup. The table also shows that the color metrics that we propose correlate better with the subjective scores compared to the state-of-the-art color metrics $PSNR_{(Y,U,V)}$. Hue seems to give lower performance than luminance in estimating point cloud quality, as shown by the PSNR and D_{hist} values. It is worth noting that the $dist_{RMS}$ and our proposed color metrics $D_{hist(l,h)}$ provide higher SROCC values, meaning that they predict monotonicity better (the rank order of the MOS is better preserved in the objective scores, which is a desirable property of a quality metric).

5.3.3. COMBINING GEOMETRY AND COLOR METRIC

In this section, we look into combining geometry-based metric and our best-performing proposed metric to improve point cloud quality estimation. We define a linear combination of the RMS distance metric and luminance histogram metric (using QF distance) as follows.

$$Q_{PC_d} = \alpha * Dist_{RMS} + (1-\alpha) * D_{hist(l)} \tag{5.7}$$

Since the $D_{hist(l)}$ values are very small compared to the $Dist_{RMS}$ values, we first normalized the $D_{hist(l)}$ values to [-1,1]. Figure 5.3 shows the resulting Pearson Linear Cor-

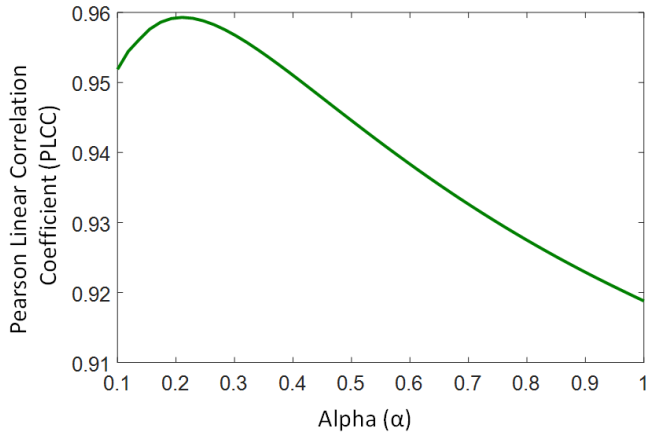


Figure 5.3: Pearson Linear Correlation Coefficient (PLCC) obtained with different values of alpha for the combination of geometry-based metric and luminance histogram-based metric

relation Coefficient (PLCC) values for different values of alpha (α). Based on the figure, we see that combining the geometry-based and luminance histogram based metrics improve the PLCC to up to 0.9596, at $\alpha = 0.2102$. This gives a near 0.4 improvement from 0.9188 and 0.9130 obtained using only $Dist_{RMS}$ and $D_{hist(l)}$, respectively.

5.4. DISCUSSION

Up until now, point cloud quality evaluation has adopted some metrics that have been traditionally used in the 3D graphics domain, such as root mean square distance or the Haussdorf distance metric. These metrics mainly aim at measuring quality based on the geometric surface of an object.

We propose objective metrics based on color, for several reasons. Firstly, in point cloud compression, colors may be coded using different approaches than octree-structure coding, and yield different color artifacts. Adopting geometry-based metric (such as PSNR for the different color channels) may not accurately quantify these color artifacts. Secondly, our subjective study shows that color distortions are an important factor for users' judgment of point cloud quality. We propose to use different color distribution representations to predict the quality of compressed point clouds, and our results correlate well with human judgment. The use of luminance values, especially, consistently yields good correlations with human judgment compared with hue.

In calculating the objective quality scores for the stimuli of our dataset, we find that state-of-the-art geometry-based metrics work quite well in predicting subjective quality

scores. For $Dist_{RMS}$ metric, the PLCC and SROCC values are 0.9188 and 0.8878, respectively. Meanwhile, $Dist_{Hauss}$ gives a PLCC and SROCC of 0.8480 and 0.7945, respectively. These results are contrary to previous studies that show low PLCC and SROCC values between 0.5 and 0.6 for point clouds with compression-like artifacts. One of the main difference between our study and previous studies is our use of complex colored point clouds as stimuli. Moreover, during rendering of our test videos, we assigned bigger point sizes for point clouds with higher levels of compression, to avoid having hollow parts in our objects due to missing points. This could have helped masking the appearance of geometric distortions in our study.

We note here that in our experiment, we created our dataset using 24 point cloud stimuli. This means that this study is more of a first step towards more studies on point cloud quality assessments. We need more point cloud datasets with subjective quality scores, in order to be able to evaluate our metrics on different stimuli.

5.5. CONCLUSION

In this chapter, we propose objective metrics for point cloud images using color distribution information, i.e. color histograms and autocorrelograms. Our proposed metric is based on the insights obtained through our subjective experiments in Chapter 3, which suggest that users consider color alterations or distortions beside the clarity of the face and overall structure of the body when evaluating point cloud quality. We show in this chapter that luminance histograms yield good point cloud quality estimation. Moreover, combining our color-based metric with existing geometry-based metric gives significant improvement in point cloud quality estimation.

Our experiments were conducted using a limited set of stimuli, as there are no publicly available point cloud quality datasets yet. Future studies should perform experiments across different datasets (e.g. different content, as well as compression algorithms) to evaluate the generalizability of our results.

Table 5.4: Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC) of Different Objective Metrics with Subjective Scores Collected in the Experiment. Subscripts h and l indicate hue and luminance, respectively.

	$dist_{RMS}$	$dist_{Hauss.}$	$PSNR_Y$	$PSNR_U$	$PSNR_V$	$PSNR_H$	$D_{hist(l)}$	$D_{hist(h)}$	$D_{C(l)}$	$D_{C(h)}$
PLCC	0.9188	0.8480	0.6085	0.3591	0.1536	0.2458	**0.9130**	**0.9055**	0.8654	0.8981
SROCC	0.8878	0.7945	0.5456	0.2561	0.2003	0.2401	**0.8841**	**0.8915**	0.8281	0.8688

IV

Outlook

6

CONCLUSIONS

In the preceding chapters up to this point, we have addressed various problems related to the subjective assessments and objective metrics for image QoE: from evaluating the reliability and repeatability of aesthetic appeal assessments, via exploring the use of mixed methodology to investigate users' perception of point cloud quality, to incorporating new influencing factors into state-of-the-art image and point cloud quality metrics to improve quality predictions. In this final chapter, we look back on the work presented in the previous chapters and reflect on how they answer our research questions (Section 6.1). We also reflect on the limitations and practical implications of our findings (Section 6.2), and finally discuss potential research directions following up the findings in this thesis (Section 6.3).

6.1. CLOSING THE LOOP

The objective of this thesis was two-fold. First, it aimed to improve state-of-the-art image quality metrics by leveraging QoE influencing factors. Despite various existing work that model the influence of different factors on user QoE, very few have actually proposed to incorporate these factors into existing quality metrics, to improve overall QoE prediction or optimization. Second, this thesis looked into the reliability of subjective methodologies for observing affective QoE variables, or QoE factors of new media, to make sure that the data collected is reliable enough to use in modeling objective quality metrics. This second objective is related to concerns about the reliability of existing subjective methodologies in collecting data related to QoE factors that are highly subjective or new to the research community. Figure 6.1 reiterates the research questions that we presented in Chapter 1, and we summarize our findings for each research question below.

We first addressed the problems related to subjective experiments in Part II of this thesis. Conducting reliable subjective experiments is an important step in building reliable QoE metrics, and it is important to understand in which cases existing methodologies may not be sufficient to collect reliable subjective data. In Chapter 2, we conducted extensive subjective experiments using four different scaling methods, on two different experimental environments, and compared the reliability and repeatability of these different experiment setups in collecting aesthetic appeal evaluations. Our studies revealed that existing rating methods could yield reliable aesthetic appeal assessments, and that aesthetic appeal assessments could be repeatable across experiment environments. This answers our research question RQ-1A (see Figure 6.1).

In Chapter 3, we conducted subjective assessments of point cloud images using a mixed methodology approach. The qualitative approach in our study allowed elicitation of factors users consider when evaluating the quality of point clouds. These factors would not have been discovered if relying on the quantitative approach alone. Our study shows the advantage of using qualitative methods to explore point cloud QoE influencing factors, and answers our research question RQ-1B.

Moving on to Part III of this thesis, we address problems related to objective metrics of visual QoE, looking into incorporating key influencing factors into state-of-the-art quality metrics to obtain better overall visual QoE prediction. In Chapter 4, we performed a subjective assessment to observe whether or not the semantic category of an image would influence its QoE. Our results showed that image scene and object categories influence users' judgment of visual quality for JPEG compressed and blurred im-

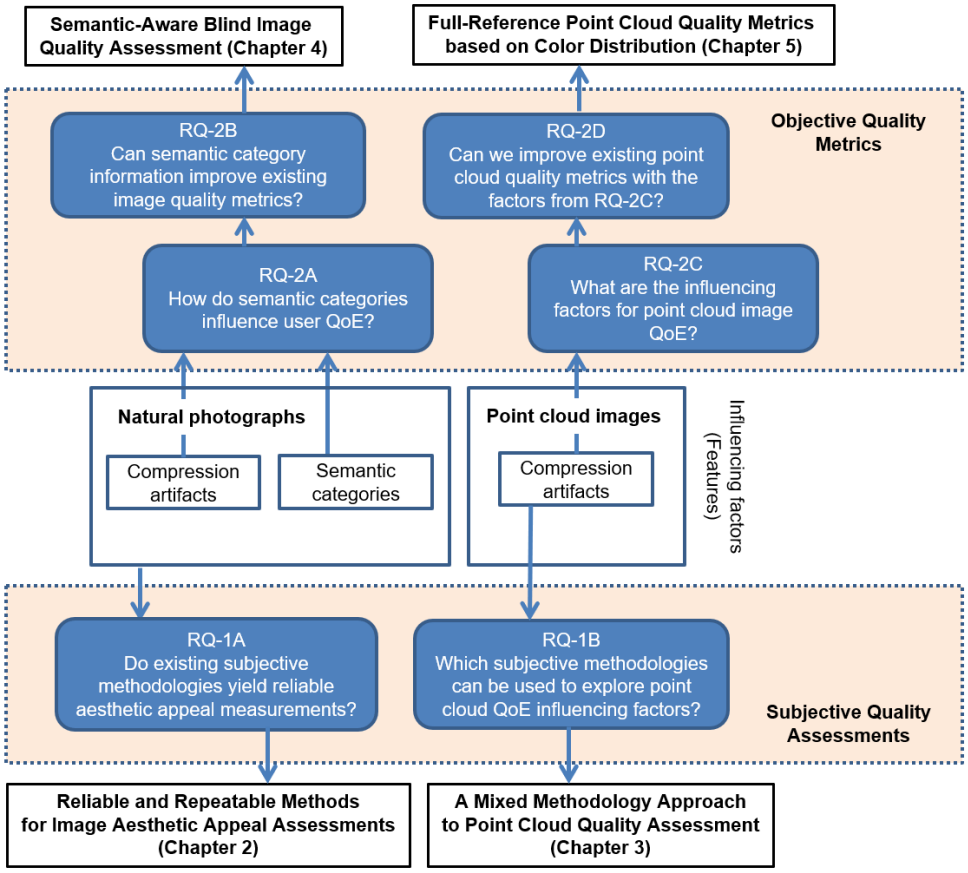


Figure 6.1: Scope and contribution of this thesis. The two orange boxes indicate the two main research scopes of this thesis. Each blue box represents one research question addressed in this thesis. The main contribution of this thesis is indicated with black-bordered boxes.

ages, which answers our research question RQ-2A. Based on these results, we then propose to incorporate semantic category features into existing state-of-the-art no-reference image quality metrics (NR-IQMs). Our experiments show an improvement in quality prediction compared to the state-of-the-art. This answers our research question RQ-2B.

In Chapter 5, we propose a full-reference quality metric for point cloud images based on color distribution features. These color distribution features were proposed based on the insight we obtained in earlier studies: color fidelity is an important factor for users when evaluating point cloud image quality. This answers our research question RQ-2C. Our experiments on the full-reference metrics show that the combination of color distribution features with state-of-the-art geometry features outperforms state-of-the-art in predicting point cloud quality. This answers our research question RQ-2D.

In the following section, we will discuss the limitations and practical implications of our findings.

6.2. DISCUSSION

Visual QoE studies aim at building reliable predictors of image/video quality. To accomplish this, reliable subjective data is needed, as well as additional features/factors beside visual artifact features. This has been supported by previous work in the field [12, 41, 39, 44]. Our work in this thesis is one step further in the effort of performing reliable subjective QoE studies, and building QoE metrics.

Our extensive studies in Chapter 2 show that image aesthetic appeal assessments can be conducted reliably and repeatedly with existing subjective methods. Although the chapter focuses on aesthetic appeal evaluations, the reliability measures used in the chapter can be applied to verify the reliability of subjective data of other affective variables, such as interestingness or enjoyment. Granted, the findings are still limited to the rating methods and environment setup that we used, i.e. discrete 5-point ACR scale, continuous scale with 5-point ACR labels, continuous scale with visual anchors, and binary scales in controlled lab and crowdsourcing environments. However, we believe that these are representative of the methods typically used within the community [52, 56, 57, 80, 77]. Part of the rating scales used in our studies have been recommended for traditional visual quality assessments in international standards [52, 53]. Our study shows how these scales are used by users in assessing affective variables. The international standards for subjective assessments of visual stimuli are established for lab environments. Recently, however, efforts have been initiated to append to existing recommendations some notes on performing experiments in a crowdsourcing environment

[178]. We believe that our study could contribute insights on this as well, especially in evaluating the reliability of results and comparisons with lab experiments.

Our work in Chapter 2 suggests that reliability measures can be used to compare data from different subjective assessments/datasets. Nevertheless, there are more discussions related to the reliability of subjective assessments that could be of interest within the QoE field, especially on how reliability measures would affect decisions in building objective QoE metrics. So far, most literature in QoE uses average scores (MOS) to build objective QoE metrics. However, these scores do not show us the bigger picture of whether or not users highly agreed with each other. Optimizing for a system with high user agreement should be different from optimizing for a system with low user agreement, even though both cases might have the same ground truth MOS. Reliability measures can and should be used to create smarter objective metrics in this regard. One way of using reliability measures in this case could be through deciding on a threshold of reliability when using different datasets to train objective metrics. Another example could be to also train objective metrics such that they learn and predict the reliability of a score instead of the average score only.

In chapters 3, 4 and 5, we conducted subjective experiments to understand whether or not certain factors influence user QoE. We performed these with different methodologies and setup (i.e. crowdsourcing, mixed methods, controlled lab setup). We then used the insights from these subjective assessments to propose additional features to extend and improve on existing QoE metrics. It is frequent to find in literature studies where additional features for QoE metrics are designed without involving users or humans in the design process. As a consequence, the role of subjective quality assessments is sometimes overlooked, and they often appear solely as a final task to evaluate the performance of a quality metric. This practice limits our understanding of human perception on visual QoE to mere correlation values with different metrics. If our systems are designed to give better predictions for humans, we should be able to trace back the success or failures of our systems to humans themselves. This can be done by carefully considering subjective quality assessments even in the steps leading to building a quality metric, especially when there is little to no prior work on the factors related to user QoE for a particular media type or application context.

In Chapter 4, we chose to use scene and object categories to represent media content as an influencing factor to image QoE. Our semantic features were extracted from existing convolutional neural networks (CNNs). A natural next step would be to also learn the perceptual quality features using CNNs. Another interesting direction would be to look into the relation of semantic category (i.e. media content) with other factors, such

as aesthetics or immersiveness, and incorporating these factors in an overall QoE metric. Our semantic-aware quality metric in chapter 4 can be implemented for any image collection application, for example to optimize how an image collection is ordered based on quality. As we mentioned in the beginning of this thesis, the idea behind our findings could be extended for videos as well. However, in this case of semantic-aware quality metric, it may be more useful to incorporate semantic category features that are more relevant for videos, e.g. genre-related instead of scene/object-related.

Finally, in chapters 3 and 5, we dealt with one example of new visual media, i.e. point clouds. As mentioned at the end of these chapters, our findings are still limited to the particular dataset that was available for us to use, which is, to date, one of the few publicly available. This is not a surprising problem with new types of media. To be more confident in generalizing research findings on these media, we will need to perform experiments on more datasets. We have taken some initial steps in this direction, as we participated in the formal subjective evaluations of the submissions for the call for proposals for point cloud compression within MPEG [179]. Nevertheless, one important step forward in visual QoE research for immersive media would be to create and share more QoE datasets of these media. Additionally, there should be awareness of the different contexts in which these media could be consumed and evaluated. For example, some media can be consumed interactively and future applications may have two or more types of media combined. This calls for variety in the type of subjective assessments conducted and datasets shared. Our metric for point clouds in Chapter 5 aims to evaluate compression technologies. However, since it is a full-reference metric, it may be used to evaluate other algorithms that alter the structure or color appearance of point clouds.

6.3. OUTLOOK AND RECOMMENDATIONS

Finally, we would like to discuss some potential research directions following up the findings in this thesis.

One fundamental direction that could be looked into is the potential of incorporating reliability measures in QoE predictions. The QoE field has used average user scores as ground truth to build metrics, but this discards the information of how user's disagree/agree in the scores. Further work could look into the merit of training a QoE predictor with reliability measures as well as the quality score, such that the predictors can provide a level of confidence along with their predicted quality score. Additionally, future work could look into leveraging reliability measures to indicate cases in which a

personalized QoE prediction would be more advantageous. In chapters 3 and 4, we gave examples of reliability measures of subjective assessments across different application contexts. This gives a rough idea of potentially providing personalized QoE optimization in applications such as gaming or web streaming, in contrast with cases/applications that have higher user agreement such as video or image compression.

Another fruitful direction in the field is the QoE of immersive media. In this thesis, we chose to work on point cloud QoE. However, future applications seem to move toward implementing mixed media (e.g. point clouds on top of omni-directional videos[1]). Future work should look into exploring how users perceive the QoE of mixed media: do users perceive the different media as one, or would users be more sensitive to the visual quality of one media compared to the other? One initial step would be to perform a study in which we control for the quality of two different media types (e.g. one as foreground and the other as background), and ask users to evaluate the quality of both media individually, as well as the overall quality of the viewing experience. The answer to these questions would be useful to design the optimization of applications with mixed media. Moreover, many applications of immersive media aspire to provide users with a sense of immersion or presence, besides high visual QoE. Looking into the relationship of these variables would be a fundamental research direction related to immersive media. A natural first step would be to determine a sensible application context, and explore the tangible aspects of the experience that users associate with immersion, presence, and quality. This could then be followed up with experiments that control for the different aspects, and ask users to evaluate immersion, presence and quality.

[1] http://vrtogether.eu/

BIBLIOGRAPHY

[1] Deloitte LLP, "The Deloitte consumer review, the growing power of consumers," 2014, Accessed: Feb 2018. [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/consumer-business/consumer-review-8-the-growing-power-of-consumers.pdf

[2] Conviva, "How consumers judge their viewing experience: a 2015 consumer survey report," Accessed: Feb 2018. [Online]. Available: https://www.conviva.com/research/consumer-survey-report-2015-how-consumers-judge-their-viewing-experience/

[3] Ofcom, "The consumer experience 2015," 2016, Accessed: Feb 2018. [Online]. Available: https://www.ofcom.org.uk/__data/assets/pdf_file/0012/51105/cer_2015_final.pdf

[4] P. Le Callet, S. Möller, A. Perkis *et al.*, "Qualinet white paper on definitions of quality of experience," *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)*, vol. 3, 2012.

[5] Cisco, "Cisco visual networking index: Forecast and methodology, 2016–2021," 2017, Accessed: Feb 2018. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf

[6] E. Perret, "Here's how many digital photos will be taken in 2017," 2016, Accessed: Feb 2018. [Online]. Available: https://mylio.com/true-stories/tech-today/how-many-digital-photos-will-be-taken-2017-repost

[7] J. Edwards, "Planet Selfie: We're now posting a staggering 1.8 billion photos every day," 2014, Accessed: Feb 2018. [Online]. Available: http://www.businessinsider.com/were-now-posting-a-staggering-18-billion-photos-to-social-media-every-day-2014-5

[8] S. S. Hemami and A. R. Reibman, "No-reference image and video quality estimation: applications and human-motivated design," *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 469–481, 2010.

[9] Mux, "2017 Video streaming perceptions report," 2017, Accessed: Feb 2018. [Online]. Available: http://connect.mux.com/2017-streaming-perceptions-report

[10] S. R. Gulliver and G. Ghinea, "Defining user perception of distributed multimedia quality," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 2, no. 4, pp. 241–257, 2006.

[11] S. H. Jumisko, V. P. Ilvonen, and K. A. Vaananen-Vainio-Mattila, "Effect of TV content in subjective assessment of video quality on mobile devices," in *Multimedia on Mobile Devices*, vol. 5684. International Society for Optics and Photonics, 2005, pp. 243–255.

[12] J. A. Redi, "Visual quality beyond artifact visibility," in *Proc. of SPIE-IS&T Electronic Imaging, vol. 8651*, San Francisco, USA, Feb 2013.

[13] J. A. Redi, Y. Zhu, H. De Ridder, and I. Heynderickx, "How passive image viewers became active multimedia users," in *Visual Signal Quality Assessment*. Springer, 2015, pp. 31–72.

[14] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE transactions on broadcasting*, vol. 57, no. 2, pp. 165–182, 2011.

[15] D. M. Chandler, "Seven challenges in image quality assessment: past, present, and future research," *ISRN Signal Processing*, vol. 2013, 2013.

[16] H. Liu and I. Heynderickx, "A perceptually relevant no-reference blockiness metric based on local image characteristics," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, p. 263540, 2009.

[17] H. Liu, N. Klomp, and I. Heynderickx, "A no-reference metric for perceived ringing artifacts in images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 4, pp. 529–539, 2010.

[18] "Perceptual blur and ringing metrics: application to JPEG2000, author=Marziliano, Pina and Dufaux, Frederic and Winkler, Stefan and Ebrahimi, Touradj, journal=Signal processing: Image communication, volume=19, number=2, pages=163–172, year=2004, publisher=Elsevier."

[19] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *International journal of computer vision*, vol. 74, no. 1, pp. 59–73, 2007.

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

[21] F. Boulos, B. Parrein, P. Le Callet, and D. Hands, "Perceptual effects of packet loss on H. 264/AVC encoded videos," in *Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM-09*, 2009.

[22] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[23] M. Kirchner and R. Bohme, "Hiding traces of resampling in digital images," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 4, pp. 582–592, 2008.

[24] L. Goldmann, F. De Simone, F. Dufaux, T. Ebrahimi, R. Tanner, and M. Lattuada, "Impact of video transcoding artifacts on the subjective quality," in *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*. IEEE, 2010, pp. 52–57.

[25] M. Čadík, M. Wimmer, L. Neumann, and A. Artusi, "Evaluation of hdr tone mapping methods using essential perceptual attributes," *Computers & Graphics*, vol. 32, no. 3, pp. 330–349, 2008.

[26] H. Pan, X.-F. Feng, and S. Daly, "Lcd motion blur modeling and analysis," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 2. IEEE, 2005, pp. II–21.

[27] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: a survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, 2011.

[28] M. Shahid, A. Rossholm, B. Lövström, and H.-J. Zepernick, "No-reference image and video quality assessment: a classification and review of recent approaches," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, p. 40, 2014.

[29] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Transactions on communications*, vol. 43, no. 12, pp. 2959–2965, 1995.

[30] S. Winkler, "Perceptual distortion metric for digital color video," in *Human Vision and Electronic Imaging IV*, vol. 3644. International Society for Optics and Photonics, 1999, pp. 175–185.

[31] A. B. Watson, Q. J. Hu, and J. F. McGowan, "Digital video quality metric based on human vision," *Journal of Electronic imaging*, vol. 10, no. 1, pp. 20–30, 2001.

[32] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[33] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: from natural scene statistics to perceptual quality," *IEEE transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[34] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012.

[35] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[36] P. V. Vu and D. M. Chandler, "A fast wavelet-based algorithm for global and local image sharpness estimation," *IEEE Signal Processing Letters*, vol. 19, no. 7, pp. 423–426, 2012.

[37] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 6, pp. 1275–1286, 2015.

[38] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[39] J. Xue and C. W. Chen, "Mobile video perception: New insights and adaptation strategies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 3, pp. 390–401, 2014.

[40] Y. Zhu, I. Heynderickx, and J. A. Redi, "Understanding the role of social context and user factors in video quality of experience," *Computers in Human Behavior*, vol. 49, pp. 412–426, 2015.

[41] Y. Zhu, A. Hanjalic, and J. A. Redi, "QoE prediction for enriched assessment of individual video viewing experience," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 801–810.

[42] F. Pereira, "Sensations, perceptions and emotions towards quality of experience evaluation for consumer electronics video adaptations," in *Proc. of Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2005, p. 910.

[43] H. Ridder and S. Endrikhovski, "33.1: Invited paper: image quality is fun: reflections on fidelity, usefulness and naturalness," in *SID Symposium Digest of Technical Papers*, vol. 33, no. 1. Wiley Online Library, 2002, pp. 986–989.

[44] H. Alers, J. Redi, H. Liu, and I. Heynderickx, "Studying the effect of optimizing image quality in salient regions at the expense of background content," *Journal of Electronic Imaging*, vol. 22, no. 4, 2013.

[45] L.-C. Hsieh, W. H. Hsu, and H.-C. Wang, "Investigating and predicting social and visual image interestingness on social media by crowdsourcing," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4309–4313.

[46] C.-H. Demarty, M. V. Sjöberg, B. Ionescu, T.-T. Do, H. Wang, N. Q. Duong, F. Lefebvre *et al.*, "Mediaeval 2016 predicting media interestingness task," in *MediaEval 2016 Multimedia Benchmark Workshop Working Notes Proceedings of the MediaEval 2016 Workshop*, 2016.

[47] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (Csur)*, vol. 40, no. 2, p. 5, 2008.

[48] R. Halonen, S. Westman, and P. Oittinen, "Naturalness and interestingness of test images for visual quality evaluation," in *Image Quality and System Performance VIII*, vol. 7867. International Society for Optics and Photonics, 2011, p. 78670Z.

[49] P. J. Seuntiëns, I. E. Heynderickx, W. A. IJsselsteijn, P. M. van den Avoort, J. Berentsen, I. J. Dalm, M. T. Lambooij, and W. Oosting, "Viewing experience and naturalness of 3d images," *Three-Dimensional TV, Video, and Display IV*, 2005.

[50] A. Singla, S. Fremerey, W. Robitza, and A. Raake, "Measuring and comparing qoe and simulator sickness of omnidirectional videos in different head mounted displays," in *Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on*. IEEE, 2017, pp. 1–6.

[51] M. Garau, M. Slater, V. Vinayagamoorthy, A. Brogni, A. Steed, and M. A. Sasse, "The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2003, pp. 529–536.

[52] I. REC, "Bt. 500-12," *Methodology for the subjective assessment of the quality of television pictures*, 2009.

[53] I. T. U. (T), "Recommendation ITU-T P. 910," *Subjective video quality assessment methods for multimedia applications*, 2008.

[54] M. H. Pinson, L. Janowski, R. Pépion, Q. Huynh-Thu, C. Schmidmer, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, "The influence of subjects and environment on audiovisual subjective tests: An international study," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 640–651, 2012.

[55] N. Staelens, S. Moens, W. Van den Broeck, I. Marien, B. Vermeulen, P. Lambert, R. Van de Walle, and P. Demeester, "Assessing quality of experience of iptv and video on demand services in real-life environments," *IEEE Transactions on broadcasting*, vol. 56, no. 4, pp. 458–466, 2010.

[56] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2014.

[57] J. A. Redi, T. Hoßfeld, P. Korshunov, F. Mazza, I. Povoa, and C. Keimel, "Crowdsourcing-based multimedia subjective evaluations: a case study on image recognizability and aesthetic appeal," in *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*. ACM, 2013, pp. 29–34.

[58] W. A. Mansilla, A. Perkis, and T. Ebrahimi, "Implicit experiences as a determinant of perceptual quality and aesthetic appreciation," in *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 2011, pp. 153–162.

[59] R. Schatz, A. Sackl, C. Timmerer, and B. Gardlo, "Towards subjective quality of experience assessment for omnidirectional video streaming," in *Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on*. IEEE, 2017, pp. 1–6.

[60] E. Alexiou and T. Ebrahimi, "On subjective and objective quality evaluation of point cloud geometry," in *9th International Conference on Quality of Multimedia Experience (QoMEX 2017)*, no. EPFL-CONF-228273. IEEE, 2017.

[61] E. Siahaan, A. Hanjalic, and J. A. Redi, "Does visual quality depend on semantics? A study on the relationship between impairment annoyance and image semantics at early attentive stages," *Electronic Imaging*, vol. 2016, no. 16, pp. 1–9, 2016.

[62] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, "What do we perceive in a glance of a real-world scene?" *Journal of vision*, vol. 7, no. 1, pp. 10–10, 2007.

[63] E. Siahaan, A. Hanjalic, and J. Redi, "A reliable methodology to collect ground truth data of image aesthetic appeal," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1338–1350, 2016.

[64] M. Tkalčič, A. Odić, A. Košir, and J. F. Tasič, "Impact of implicit and explicit affective labeling on a recommender system's performance," in *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 2011, pp. 342–354.

[65] K. Yadati, H. Katti, and M. Kankanhalli, "Cavva: Computational affective video-in-video advertising," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 15–23, 2014.

[66] W. Yin, T. Mei, C. W. Chen, and S. Li, "Socialized mobile photography: Learning to photograph with social context via mobile devices," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 184–200, 2014.

[67] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 271–280.

[68] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *European Conference on Computer Vision*. Springer, 2006, pp. 288–301.

[69] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, 2011.

[70] C. Li and T. Chen, "Aesthetic visual quality assessment of paintings," *IEEE Journal of selected topics in Signal Processing*, vol. 3, no. 2, pp. 236–252, 2009.

[71] S. Rudinac, M. Larson, and A. Hanjalic, "Learning crowdsourced user preferences for visual summarization of image collections," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1231–1243, 2013.

[72] P. Engeldrum, *Psychometric Scaling, A Toolkit for Imaging Systems Development*. Imcotek Press, Winchester, USA, 2000.

[73] H. de Ridder, "Cognitive issues in image quality measurement," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 47–55, 2001.

[74] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," *IEEE Transactions on Broadcasting*, vol. 57, no. 1, pp. 1–14, 2011.

[75] J. Redi, H. Liu, H. Alers, R. Zunino, and I. Heynderickx, "Comparing subjective image quality measurement methods for the creation of public databases," in *Image Quality and System Performance VII*, vol. 7529. International Society for Optics and Photonics, 2010, p. 752903.

[76] J. A. Redi and I. Heynderickx, "Image integrity and aesthetics: towards a more encompassing definition of visual quality," in *Human Vision and Electronic Imaging XVII*, vol. 8291. International Society for Optics and Photonics, 2012, p. 829115.

[77] A. Agrawal, V. Premachandran, and R. Kakarala, "Rating image aesthetics using a crowd sourcing approach," in *Pacific-Rim Symposium on Image and Video Technology*. Springer, 2013, pp. 24–32.

[78] T. S. Sachs, R. Kakarala, S. L. Castleman, and D. Rajan, "A data-driven approach to understanding skill in photographic composition," in *Asian Conference on Computer Vision*. Springer, 2010, pp. 112–121.

[79] T. Zhang, H. T. Nefs, J. Redi, and I. Heynderickx, "The aesthetic appeal of depth of field in photographs," in *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*. IEEE, 2014, pp. 81–86.

[80] C. D. Cerosaletti and A. C. Loui, "Measuring the perceived aesthetic quality of photographic images," in *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*. IEEE, 2009, pp. 47–52.

[81] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2408–2415.

[82] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2206–2213.

[83] L. L. Thurstone, "A law of comparative judgment." *Psychological review*, vol. 34, no. 4, p. 273, 1927.

[84] G. A. Gescheider, *Psychophysics: the fundamentals*. Psychology Press, 2013.

[85] B. Keelan, *Handbook of image quality: characterization and prediction*. CRC Press, 2002.

[86] J. A. Roufs, "Perceptual image quality: Concept and measurement," *Philips Journal of Research*, vol. 47, no. 1, pp. 35–62, 1992.

[87] S. Winkler, "On the properties of subjective ratings in video quality experiments," in *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*. IEEE, 2009, pp. 139–144.

[88] C.-C. Wu, K.-T. Chen, Y.-C. Chang, and C.-L. Lei, "Crowdsourcing multimedia qoe evaluation: A trusted framework," *IEEE transactions on multimedia*, vol. 15, no. 5, pp. 1121–1137, 2013.

[89] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," in *Computer Graphics Forum*, vol. 31, no. 8. Wiley Online Library, 2012, pp. 2478–2491.

[90] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video," in *Quality of multimedia experience (QoMEX), 2010 second international workshop on*. IEEE, 2010, pp. 82–87.

[91] A. M. Van Dijk, J.-B. Martens, and A. B. Watson, "Quality asessment of coded images using numerical category scaling," in *Advanced Image and Video Communications and Storage Technologies*, vol. 2451. International Society for Optics and Photonics, 1995, pp. 90–102.

[92] W.-T. Chu, Y.-K. Chen, and K.-T. Chen, "Size does matter: How image size affects aesthetic perception?" in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 53–62.

[93] A. L. J. Stout and S. Ocko, "Jene, a lightweight evolutionary art package for java," 2009, Accessed: 2014. [Online]. Available: https://code.google.com/p/jene/

[94] K. Sims, *Artificial evolution for computer graphics*. ACM, 1991, vol. 25, no. 4.

[95] E. A. Vessel and N. Rubin, "Beauty and the beholder: highly individual taste for abstract, but not real-world images," *Journal of vision*, vol. 10, no. 2, pp. 18–18, 2010.

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

[97] K. A. Hallgren, "Computing inter-rater reliability for observational data: an overview and tutorial," *Tutorials in quantitative methods for psychology*, vol. 8, no. 1, p. 23, 2012.

[98] T. Hoβfeld, R. Schatz, and S. Egger, "Sos: The mos is not enough!" in *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*. IEEE, 2011, pp. 131–136.

[99] L. Janowski and M. Pinson, "Subject bias: Introducing a theoretical user model," in *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*. IEEE, 2014, pp. 251–256.

[100] E. Siahaan, J. A. Redi, and A. Hanjalic, "Beauty is in the scale of the beholder: Comparison of methodologies for the subjective assessment of image aesthetic appeal," in *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*. IEEE, 2014, pp. 245–250.

[101] K. Sijtsma, "On the use, the misuse, and the very limited usefulness of cronbach's alpha," *Psychometrika*, vol. 74, no. 1, p. 107, 2009.

[102] M. Lombard, F. Biocca, J. Freeman, W. IJsselsteijn, and R. J. Schaevitz, *Immersed in Media: Telepresence Theory, Measurement & Technology*. Springer Publishing Company, Inc., 2015.

[103] R. Schnabel and R. Klein, "Octree-based point-cloud compression." *Spbg*, vol. 6, pp. 111–120, 2006.

[104] Y. Huang, J. Peng, C.-C. J. Kuo, and M. Gopi, "A generic scheme for progressive point cloud coding," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 2, pp. 440–453, 2008.

[105] J. Kammerl, N. Blodow, R. B. Rusu, S. Gedikli, M. Beetz, and E. Steinbach, "Real-time compression of point cloud streams," in *2012 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2012, pp. 778–785.

[106] D. Thanou, P. A. Chou, and P. Frossard, "Graph-based compression of dynamic 3d point cloud sequences," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1765–1778, 2016.

[107] R. Mekuria, K. Blom, and P. Cesar, "Design, implementation, and evaluation of a point cloud codec for tele-immersive video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 828–842, 2017.

[108] E. Alexiou, E. Upenik, and T. Ebrahimi, "Towards subjective quality assessment of point cloud imaging in augmented reality," in *19th International Workshop on Multimedia Signal Processing*, no. EPFL-CONF-230115. IEEE, 2017.

[109] E. Alexiou and T. Ebrahimi, "On the performance of metrics to predict quality in point cloud representations," in *Applications of Digital Image Processing XL*, vol. 10396. International Society for Optics and Photonics, 2017.

[110] B. Jackson, "3D scanning expert explains how point clouds are making a better virtual reality," 2017. [Online]. Available: https://3dprintingindustry.com/news/3d-scanning-expert-explains-point-clouds-making-better-virtual-reality-105373/

[111] "Intel FreeD Technology," 2017. [Online]. Available: https://www.intel.com/content/www/us/en/sports/technology/intel-freed-360-replay-technology.html

[112] S. Zancajo-Blazquez, S. Lagueela-Lopez, D. Gonzalez-Aguilera, and J. Martinez-Sanchez, "Segmentation of indoor mapping point clouds applied to crime scenes reconstruction," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 7, pp. 1350–1358, 2015.

[113] M. Whitty, S. Cossell, K. S. Dang, J. Guivant, and J. Katupitiya, "Autonomous navigation using a real-time 3d point cloud," in *2010 Australasian Conference on Robotics and Automation*, 2010, pp. 1–3.

[114] S. Spina, K. Debattista, K. Bugeja, and A. Chalmers, "Point cloud segmentation for cultural heritage sites," in *VAST: International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage*. The Eurographics Association, 2011.

[115] S. O. E. Charles Loop, Qin Cai and P. A. Chou, "Microsoft voxelized upper bodies - a voxelized point cloud dataset," last accessed: December 2017. [Online]. Available: https://jpeg.org/plenodb/pc/microsoft/

[116] T. M. Eugene dÉon, Bob Harrison and P. A. Chou, "8i voxelized full bodies - a voxelized point cloud dataset," ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG), Tech. Rep. WG11M40059/WG1M74006, January 2017.

[117] G. de Tratamiento de Imágenes Universidad Politécnica de Madrid, "Gti-upm point-cloud data set," last accessed: December 2017. [Online]. Available: https://jpeg.org/plenodb/pc/upm/

[118] J. Zhang, W. Huang, X. Zhu, and J.-N. Hwang, "A subjective quality evaluation for 3d point cloud models," in *2014 International Conference on Audio, Language and Image Processing (ICALIP)*. IEEE, 2014, pp. 827–831.

[119] M. Cadoret, S. Lê *et al.*, "The sorted napping: A new holistic approach in sensory evaluation," *Journal of Sensory Studies*, vol. 25, no. 5, pp. 637–658, 2010.

[120] S. Lê, T. Lê, and M. Cadoret, "Napping and sorted napping as a sensory profiling technique," in *Rapid sensory profiling techniques. Applications in New Product Development and Consumer Research*. Woodhead Publishing Cambridge, 2015, pp. 197–213.

[121] D. Strohmeier, S. Jumisko-Pyykkö, and K. Kunze, "Open profiling of quality: a mixed method approach to understanding multimodal quality perception," *Advances in multimedia*, vol. 2010, 2010.

[122] D. Strohmeier, K. Kunzem, K. Göbel, and J. Liebetrau, "Evaluation of differences in quality of experience features for test stimuli of good-only and bad-only overall audio visual quality," in *Image Quality and System Performance X*, vol. 8653, 2013.

[123] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on image processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[124] S. Tourancheau, F. Autrusseau, Z. P. Sazzad, and Y. Horita, "Impact of subjective dataset on the performance of image quality metrics," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. IEEE, 2008, pp. 365–368.

[125] T. Brandao, L. Roque, and M. P. Queluz, "Quality assessment of h. 264/avc encoded video," in *Proc. of conference on telecommunications-ConfTele*, Sta. Maria da Feira, Portugal, 2009.

[126] I. Viola, M. Řeřábek, and T. Ebrahimi, "Impact of interactivity on the assessment of quality of experience for light field content," in *9th International Conference on Quality of Multimedia Experience (QoMEX 2017)*. IEEE, 2017, pp. 1–6.

[127] Conviva, "Viewer experience report," 2015, Accessed: May 2017. [Online]. Available: http://www.conviva.com/conviva-viewer-experience-report/vxr-2015/

[128] K. Gu, G. Zhai, W. Lin, X. Yang, and W. Zhang, "No-reference image sharpness assessment in autoregressive parameter space," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 3218–3231, 2015.

[129] S. C. Guntuku, J. T. Zhou, S. Roy, W. Lin, and I. W. Tsang, "Understanding deep representations learned in modeling users likes," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3762–3774, 2016.

[130] U. Engelke, R. Pepion, P. L. Callet, and H.-J. Zepernick, "Linking distortion perception and visual saliency in H. 264/AVC coded video containing packet loss," in *Visual Communications and Image Processing*. SPIE, 2010.

[131] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, "The application of visual saliency models in objective image quality assessment: A statistical evaluation," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 6, pp. 1266–1278, 2016.

[132] K. Gu, L. Li, H. Lu, X. Min, and W. Lin, "A fast reliable image quality predictor by fusing micro-and macro-structures," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 5, pp. 3903–3912, 2017.

[133] D. Temel and G. AlRegib, "Resift: Reliability-weighted sift-based image quality assessment," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2047–2051.

[134] P. Zhang, W. Zhou, L. Wu, and H. Li, "Som: Semantic obviousness metric for image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2394–2402.

[135] D. Marr, "Vision: A computational approach," 1982.

[136] S. Edelman, S. Dickinson, A. Leonardis, B. Schiele, and M. Tarr, "On what it means to see, and what we can do about it," *Object Categorization: Computer and Human Vision Perspectives*, pp. 69–86, 2009.

[137] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem, "Basic objects in natural categories," *Cognitive psychology*, vol. 8, no. 3, pp. 382–439, 1976.

[138] A. Rorissa and H. Iyer, "Theories of cognition and image categorization: What category labels reveal about basic level theory," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 9, pp. 1383–1392, 2008.

[139] I. Biederman, R. C. Teitelbaum, and R. J. Mezzanotte, "Scene perception: a failure to find a benefit from prior expectancy or familiarity." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 9, no. 3, p. 411, 1983.

[140] A. Torralba, K. P. Murphy, and W. T. Freeman, "Using the forest to see the trees: exploiting context for visual object detection and localization," *Communications of the ACM*, vol. 53, no. 3, pp. 107–114, 2010.

[141] D. M. Rouse, R. Pepion, S. S. Hemami, and P. Le Callet, "Image utility assessment and a relationship with image quality assessment," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2009, pp. 724 010–724 010.

[142] E. Siahaan, A. Hanjalic, and J. A. Redi, "Augmenting blind image quality assessment using image semantics," in *2016 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2016, pp. 307–312.

[143] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, 2016.

[144] S. Ryu and K. Sohn, "Blind blockiness measure based on marginal distribution of wavelet coefficient and saliency," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1874–1878.

[145] P. Gastaldo, R. Zunino, and J. Redi, "Supporting visual quality assessment with machine learning," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, pp. 1–15, 2013.

[146] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 50–63, 2015.

[147] K. Gu, J. Zhou, J. Qiao, G. Zhai, W. Lin, and A. Bovik, "No-reference quality assessment of screen content pictures," *IEEE Transactions on Image Processing*, 2017.

[148] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4850–4862, 2014.

[149] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1098–1105.

[150] P. Ye and D. Doermann, "No-reference image quality assessment using visual codebooks," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3129–3138, 2012.

[151] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, 1995.

[152] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Image database tid2013: peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015.

[153] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, 2010.

[154] P. Korshunov, P. Hanhart, T. Richter, A. Artusi, R. Mantiuk, and T. Ebrahimi, "Subjective quality assessment database of hdr images compressed with jpeg xt," in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*. IEEE, 2015, pp. 1–6.

[155] A. Ciancio, A. L. N. T. da Costa, E. A. da Silva, A. Said, R. Samadani, and P. Obrador, "No-reference blur assessment of digital pictures based on multifeature classifiers," *IEEE Transactions on image processing*, vol. 20, no. 1, pp. 64–75, 2011.

[156] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016.

[157] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616–625, 2012.

[158] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: a large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[159] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[160] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[161] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[162] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[163] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495.

[164] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[165] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[166] T. N. Pappas, R. J. Safranek, and J. Chen, "Perceptual criteria for image quality evaluation," *Handbook of image and video processing*, pp. 669–684, 2000.

[167] C.-H. Chou and Y.-C. Li, "A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile," *IEEE Transactions on circuits and systems for video technology*, vol. 5, no. 6, pp. 467–476, 1995.

[168] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "Tid2008-a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

[169] B. L. Jones and P. R. McManus, "Graphic scaling of qualitative terms," *SMPTE journal*, vol. 95, no. 11, pp. 1166–1171, 1986.

[170] D. M. Rouse, S. S. Hemami, R. Pépion, and P. Le Callet, "Estimating the usefulness of distorted natural images using an image contour degradation measure," *JOSA A*, vol. 28, no. 2, pp. 157–188, 2011.

[171] R. Mekuria and L. Bivolarsky, "Overview of the mpeg activity on point cloud compression," in *Data Compression Conference (DCC), 2016*. IEEE, 2016, pp. 620–620.

[172] R. Mekuria and P. Cesar, "Mp3dg-pcc, open source software framework for implementation and evaluation of point cloud compression," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 1222–1226.

[173] R. Mekuria, Z. Li, C. Tulvan, and P. Chou, "Evaluation criteria for point cloud compression," *ISO/IEC MPEG*, no. 16332, 2016.

[174] D. Tian, H. Ochimizu, C. Feng, R. Cohen, and A. Vetro, "Evaluation metrics for point cloud compression," *ISO/IEC JTC m74008, Geneva, Switzerland*, 2017.

[175] F. De Simone, F. Dufaux, T. Ebrahimi, C. Delogu, and V. Baroncini, "A subjective study of the influence of color information on visual quality assessment of high resolution pictures," in *Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM-09)*, no. MMSPL-CONF-2009-001, 2009.

[176] D. Lee and K. N. Plataniotis, "Perceptual color difference assessment using histogram distance on hue histogram descriptor," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 546–550.

[177] J. A. Redi, P. Gastaldo, I. Heynderickx, and R. Zunino, "Color distribution information for the reduced-reference assessment of perceived image quality," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 12, pp. 1757–1769, 2010.

[178] I. T. U. (T), "Recommendation ITU-T P. 912 (03/2016)," *Subjective video quality assessment methods for recognition tasks*, 2016.

[179] V. Baroncini, P. Cesar Garcia, E. Siahaan, I. Reimat, and S. Subramanyam, "Report of the formal subjective assessment test of the submission received in response to the call for proposals for point cloud compression, contribution to mpeg, october 2017," 2017.

ACKNOWLEDGEMENTS

No one exists in a vacuum, nor does her work. The same goes for myself and this thesis. Were it not for the guidance, support, and help of the people around me, my PhD years would not have been as fruitful as they are.

I would like to firstly thank my promotor and copromotor, Alan and Judith, for inviting me to Delft to work on this PhD project almost five years ago. To Alan, thank you for being a supportive promotor since the beginning; I truly enjoyed our high-level discussions, and appreciate the freedom you gave to do my research and grow professionally.

To Judith, who acted as the daily supervisor of my PhD: thank you for being an inspiration to me with your ethics and passion, thank you for our interesting discussions on work and life in general, and thank you for believing in me and my work.

I thank my committee members, whom I've had the pleasure to know and learn from even before the defence of this thesis. Patrick Le Callet, whom I met through QoMex, SPIE, and the IEEE-EURASIP Signal Processing summer school; Alexander Raake through QoMex; Ingrid Heynderickx through your visits to the TU and the Qualinet summer school; Huib de Ridder through SPIE and our encounters at the Pi-Lab meetings; Catholijn Jonker through my time at the Interactive Intelligence (II) group.

Special thanks also to Pablo Cesar, my supervisor at CWI, Amsterdam. Thank you for the opportunity to work with, and learn from you and the group at CWI. Thank you for your positivity and openness, and for all the advice you gave through our discussions.

I would like to thank the faculty members of the Multimedia Computing (MMC) Group, Huijian, Cynthia, and Martha, for your input and support that I have received throughout my time with the group.

I am grateful for the administrative and technical support of Saskia Peters, Ruud de Jong, Bart Vastenouw, and Anita Hoogmoed.

I would like to also extend my thanks to the Pi-Lab members. I'm glad for all the insights I received from you on conducting perception studies, and that we were able to help each other on our research.

A big part of my PhD involved user studies, and I am thankful for everyone who willingfully participated and contributed to the studies. A special thanks to Karina Angelia and David Herlaar who participated in every single one of my experiments since the be-

ginning!

A PhD is often described as a lonely journey, and though I understand why it appears to be so, I thankfully did not feel like it was that way. Thanks to my colleagues during my time at the II group and MMC group of the TU Delft, and at the Distributed Interactive Systems (DIS) group at CWI in Amsterdam: Reyhan Aydogan, Tingting Zhang, Iulia Lefter, Chris Rozemuller, Vanessa Vakili, Tim Barslaag, Christina Katsimerou, Alex Kayal, Karthik Yadati, Raynor Vliegendhart, Alessio Bazzica, Babak Loni, Yi Zhu, Xiuxiu Zhan, Jaehun Kim, Manel Slokom, Jie Li, Marwin Schmitt, Yiping Kong, Shishir Subarmanyam, Nacho Reimat, Simon Gunkel, Kees Blom, Fons Kuijk, Jack Jansen, Thomas Roggla. Thank you for the discussions, the friendly banters, the coffee breaks, the game/movie/dinner nights, and all the help and support you have given to me.

I'd also like to thank Elmar and his group, the Computer Graphics and Visualization (CGV) group of the TU Delft, for always welcoming me at their Friday beers.

My friends at the ISC Delft Choir kept me grounded and happy whenever we spent our many times together. Constanca, Catarina, Silvia, Claudia Aguilar, Gijs, Claudia Latorre, Pavel, Francesco, Mariana, Mohan, Mithun, Elsa, Ashish, Agnelo, and all the members who are now far from Delft. If sound could travel through these pages, please know that I am singing to you these words: *"Who can say if I've been changed for the better, but because I knew you, I have been changed for good."*

Thank you to Fr. Avin and Rev. Waltraut for their support through the ISC Delft community during my time in Delft.

I am grateful to have had wonderful roommates, travel buddies, and Indonesian friends who made my life outside of work so stress-free: Anne, Catarina Santos, Ella, Elizabeth, Tri, Tesa, Tara, Karina, Jessica, and Andhika.

My family has been my motivation to do well in life for as long as I can remember. My Dad once said to me that "friends will come and will leave, but family will always be a constant in who you are". Thank you Bapak, Mama, Swarna, Boas, for being a positive constant for me: a sure source of support, motivation, and sense of identity.

I also thank my extended family in Indonesia.

I am thankful for my parents-in-law, Axel and Bärbel, for always showing their support and kindness from the first time we met.

I am grateful to have met my now-husband, Pablo Bauszat, during my PhD. Thank you, Pablo, for always believing in me, and making me believe in myself.

Lastly, but never the least. I thank Jesus Christ, the author and finisher of my faith.

About the Author

Ernestasia Siahaan was born in Medan, Indonesia, on 16 November 1987. She completed her high school education there (2002-2006), and spent one year as an exchange student in Belgium with the American Field Service (AFS) program. In 2010, she graduated as Bachelor of Engineering in Informatics Engineering from Bandung Institute of Technology, Indonesia. During her bachelor studies, she interned as a software engineer at the Information Technology Department of Bank Indonesia. After her graduation, she spent one year working as a research assistant at the Graphics and Artificial Intelligence Lab of Bandung Institute of Technology, while working full-time as a teaching staff for the Informatics Engineering Department of Harapan Bangsa Institute of Technology in Bandung, Indonesia.

In 2011, Ernestasia was awarded the Taiwan Government Scholarship by the Ministry of Education (MoE) of Taiwan to conduct her master's study at Taiwan's National Central University (NCU). She graduated in 2013 as a Master of Science in Computer Science and Information Engineering. She then worked as a research assistant at the Media Systems Lab of NCU.

Ernestasia came to Delft in The Netherlands in December 2013 to pursue her PhD degree under the supervision of Prof. Dr. Alan Hanjalic and Dr. Judith Redi at Delft University of Technology. During her PhD studies, she was awarded Best Poster Award at the 2016 IEEE-EURASIP Summer School on Signal Processing. From September 2016, she became a visiting researcher at the Distributed & Interactive Systems (DIS) Group in Centrum voor Wiskunde en Informatica (CWI), a research center for Mathematics and Informatics in Amsterdam. From December 2017, she continued working at CWI as a post-doctoral researcher. In summer 2018, she moved to Zurich in Switzerland, and is now building her career there.

Propositions

accompanying the dissertation

VISUAL QUALITY OF EXPERIENCE

A METRIC-DRIVEN PERSPECTIVE

by

Ernestasia SIAHAAN

1. Objective metrics are only as objective as the subjective ground truth. (this thesis)
2. A human-centric system needs to involve humans since its inception, instead of only during its evaluation. (this thesis)
3. To move forward the research on the QoE of immersive media is to move away from traditional, standard methodologies of subjective visual QoE assessments. (this thesis)
4. Devising reliable QoE metrics requires interpretability of the QoE influencing factors these metrics take into account.
5. The common practice of averaging people out should be discouraged in research as much as it is discouraged in society.
6. Political correctness is the placebo to society's problems with diversity.
7. Equality cannot be reached without positive discrimination.
8. It is better for PhD students to not do multi-disciplinary research.
9. For democracy to work, we first need to banish ignorance.
10. We need more karaoke booths in the Netherlands.

These propositions are regarded as opposable and defendable, and have been approved as such by the promotor prof. dr. A. Hanjalic.