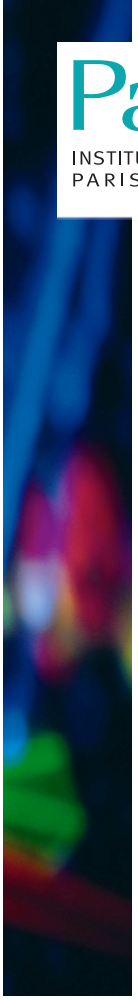
Assessment and analysis of high dynamic range video quality

Emin Zerman



ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2018-ENST-0003

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal et Images »

présentée et soutenue publiquement par

Emin ZERMAN

le 19 Janvier 2018

Évaluation et analyse de la qualité vidéo à haute gamme dynamique

Assessment and analysis of high dynamic range video quality

Directeur de thèse : **Frédéric DUFAUX**
Co-encadrement de la thèse : **Giuseppe VALENZISE**

Jury

M. Rémi COZOT, Maître de Conférences HDR, IRISA, Université de Rennes 1	Rapporteur
M. Erik REINHARD, Distinguished Scientist, Technicolor Research and Innovation	Rapporteur
M. Patrick LE CALLET, Professeur, LS2N, Polytech Nantes/Université de Nantes	Examinateur
Mme. Gözde BOZDAĞI AKAR, Professeur, Middle East Technical University	Examinateur

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

T H È S E

Abstract

In the last decade, high dynamic range (HDR) image and video technology gained a lot of attention, especially within the multimedia community. Recent technological advancements made the acquisition, compression, and reproduction of HDR content easier, and that led to the commercialization of HDR displays and popularization of HDR content. In this context, measuring the quality of HDR content plays a fundamental role in improving the content distribution chain as well as individual parts of it, such as compression and display. However, HDR visual quality assessment presents new challenges with respect to the standard dynamic range (SDR) case. In this thesis, we identify some of these challenges and suggest solutions to these problems.

The first challenge is the new conditions introduced by the reproduction of HDR content, e.g. the increase in brightness and contrast. Even though accurate reproduction is not necessary for most of the practical cases, accurate estimation of the emitted luminance is necessary for the objective HDR quality assessment metrics. In order to understand the effects of display rendering on the quality perception, an accurate HDR frame reproduction algorithm was developed, and a subjective experiment was conducted to analyze the impact of different display renderings on subjective and objective HDR quality evaluation. Additionally, in order to understand the impact of color with the increased brightness of the HDR displays, the effects of different color spaces on the HDR video compression performance were also analyzed in another subjective study.

Another challenge is to estimate the quality of HDR content objectively, using computers and algorithms. In order to address this challenge, the thesis proceeds with the performance evaluation of full-reference (FR) HDR image quality metrics. HDR images have a larger brightness range and higher contrast values. Since most of the image quality metrics are developed for SDR images, they need to be adapted in order to estimate the quality of HDR images. Different adaptation methods were used for SDR metrics, and they were compared with the existing image quality metrics developed exclusively for HDR images. Moreover, we propose a new method for the evaluation of metric discriminability based on a novel classification approach.

Motivated by the need to fuse several different quality databases, in the third part of the thesis, we compare subjective quality scores acquired by using different subjective test methodologies. Subjective quality assessment is regarded as the most effective and reliable

way of obtaining “ground-truth” quality scores for the selected stimuli, and the obtained mean opinion scores (MOS) are the values to which generally objective metrics are trained to match. In fact, strong discrepancies can easily be notified across databases when different multimedia quality databases are considered. In order to understand the relationship between the quality values acquired using different methodologies, the relationship between MOS values and pairwise comparisons (PC) scaling results were compared. For this purpose, a series of experiments were conducted using double stimulus impairment scale (DSIS) and pairwise comparisons subjective methodologies. We propose to include cross-content comparisons in the PC experiments in order to improve scaling performance and reduce cross-content variance as well as confidence intervals. The scaled PC scores can also be used for subjective multimedia quality assessment scenarios other than HDR.

Keywords: High dynamic range, objective quality assessment, image quality metrics, display rendering, subjective quality assessment.

Abstract

Au cours de la dernière décennie, la technologie de l'image et de la vidéo à haute gamme dynamique (High dynamic range - HDR) a attiré beaucoup d'attention, en particulier dans la communauté multimédia. Les progrés technologiques récents ont facilité l'acquisition, la compression et la reproduction du contenu HDR, ce qui a mené à la commercialisation des écrans HDR et à la popularisation du contenu HDR. Dans ce contexte, la mesure de la qualité du contenu HDR joue un rôle fondamental dans l'amélioration de la chaîne de distribution du contenu ainsi que des opérations qui la composent, telles que la compression et l'affichage. Cependant, l'évaluation de la qualité visuelle HDR présente de nouveaux défis par rapport au contenu à gamme dynamique standard (Standard dynamic range - SDR). Dans cette thèse, nous identifions certains de ces défis et suggérons des solutions à ces problèmes.

Le premier défi concerne les nouvelles conditions introduites par la reproduction du contenu HDR, par ex. l'augmentation de la luminosité et du contraste. Même si une reproduction exacte de la luminance d'une scène n'est pas nécessaire pour la plupart des cas pratiques, une estimation précise de la luminance émise est cependant nécessaire pour les mesures d'évaluation objectives de la qualité HDR. Afin de comprendre les effets du rendu d'affichage sur la perception de la qualité, un algorithme permettant de réproduire très précisement une image HDR a été développé et une expérience subjective a été menée pour analyser l'impact de différents rendus sur l'évaluation subjective et objective de la qualité HDR. En outre, afin de comprendre l'impact de la couleur avec la luminosité accrue des écrans HDR, les effets des différents espaces de couleurs sur les performances de compression vidéo HDR ont également été analysés dans une autre étude subjective.

Un autre défi consiste à estimer objectivement la qualité du contenu HDR, en utilisant des ordinateurs et des algorithmes. Afin de relever ce défi, la thèse procède à l'évaluation des performances des métriques de qualité d'image HDR avec référence (Full reference - FR). Les images HDR ont une plus grande plage de luminosité et des valeurs de contraste plus élevées. Étant donné que la plupart des métriques de qualité d'image sont développées pour les images SDR, elles doivent être adaptées afin d'estimer la qualité des images HDR. Différentes méthodes d'adaptation ont été utilisées pour les mesures SDR, et elles ont été comparées avec les métriques de qualité d'image existantes développées exclusivement pour les images HDR. De plus, nous proposons une nouvelle méthode d'évaluation des métriques

objectives basée sur une nouvelle approche de classification.

Motivée par la nécessité de fusionner plusieurs bases de données de qualité, dans la troisième partie de la thèse, nous comparons les scores de qualité subjectifs acquis en utilisant différentes méthodologies de test subjectives. L'évaluation subjective de la qualité est considérée comme le moyen le plus efficace et le plus fiable d'obtenir des scores de qualité «vérité-terrain» pour les stimuli sélectionnés, et les scores moyens d'opinion (Mean opinion scores - MOS) obtenus sont les valeurs auxquelles les métriques objectives sont entraînées pour correspondre. En fait, de fortes divergences peuvent facilement être rencontrés lorsque différentes bases de données de qualité multimédia sont considérées. Afin de comprendre la relation entre les valeurs de qualité acquises à l'aide de différentes méthodologies, la relation entre les valeurs MOS et les résultats des comparaisons par paires rééchellonés (Pairwise comparisons - PC) a été comparée. Á cette fin, une série d'expériences ont été menées entre les méthodologies double stimulus impairment scale (DSIS) et des comparaisons par paires. Nous proposons d'inclure des comparaisons inter-contenu dans les expériences PC afin d'améliorer les performances de rééchelonnement et de réduire la variance inter-contenu ainsi que les intervalles de confiance. Les scores de PC rééchellonés peuvent également être utilisés pour des scénarios subjectifs d'évaluation de la qualité multimédia autres que le HDR.

Mots clés: Haute gamme dynamique, évaluation objective de la qualité, métriques de qualité d'image, rendu d'affichage, évaluation subjective de la qualité.

Table of Contents

List of Figures

List of Tables

Introduction

Context and Motivation

The human visual system (HVS) is able to perceive a much wider range of colors and luminous intensities present in our environment than the traditional standard dynamic range (SDR) imaging systems can capture and reproduce. High dynamic range (HDR) technology attempts to overcome these limitations of SDR imaging systems and to enhance user experience. Thanks to the advancements in the imaging and display technologies in the last decade, we are now able to capture, store, transmit, and display images and videos in a more realistic manner [BADC11, DLCMM16]. Being able to reproduce HDR scenes accelerated the standardization efforts for HDR image and video compression [Ric13, LFH15, HRE16] as parts of end-to-end HDR content delivery chain. In order to ensure that compression is done with the highest quality possible, quality assessment is necessary for HDR images and videos. However, the enhanced brightness and contrast of HDR introduce new conditions and constraints to quality assessment problem.

This thesis focuses on the assessment and analysis of the high dynamic range image and video. Image and video quality assessment problem is a widely studied problem in the signal processing community [SSB06, SSBC10a, PJI$^+$15] for the case of SDR. However, these works have a number of limitations. Human perception of light is not proportional to the physical magnitude of the light. In order to account for this non-linearity, the image pixel values are processed using a power law curve, called *gamma correction function* [ITU11], for SDR displays. After this operation, the SDR pixel values become perceptually linear where the change in the magnitude will correspond to a proportional change in the perception. Thus, the objective SDR quality assessment methods assume that the image pixels are perceptually uniform. This is not the case for the HDR images, as HDR images generally store physical luminance values, in cd/m^2, or pixel values which are proportional to the physical luminance values. Similarly, subjective HDR quality assessment is expected to be different since the level and the ratio of brightness are different. For a proper assessment, these new conditions have to be taken into account.

Although estimation and assessment of the quality of the video are essential for many other applications, image and video compression is considered as the main source of distortion throughout the thesis, as it is the most practical and realistic scenario. Based

on these considerations, we ask the following question: ***What are the parameters that affect the estimation of full-reference objective quality and the perception of subjective quality for the case of HDR image and video compression?***

Attempting to answer this question, we first identify two main aspects that may impact both the objective and subjective quality assessment of HDR image and video:

- Although the luminance range of SDR can be made perceptually uniform using the gamma correction function, the sRGB gamma correction function does not work for the brighter and darker luminance values which are introduced by HDR [AMS08]. For this purpose, a perceptually uniform (PU) encoding of HDR luminance values was proposed by Aydın et al. [AMS08]. Using PU encoding, HDR pixel values can be represented as perceptually linear values, and objective quality assessment methods yield better results compared to using physical luminance (i.e. photometric) values [VDSLD14]. Nevertheless, in [VDSLD14], the display model is simulated, and the exact values of the emitted luminance are not known. Objective quality metrics designed for HDR images [NMDSLC15] and videos [NDSLC15] require emitted luminance values for estimation of quality scores. However, the impact of the knowledge of emitted luminance values and the effects of different display renderings on HDR quality assessment have not been studied yet. Additionally, even though some researchers analyzed the human visual system in terms of brightness and contrast perception for a wider luminance range, the effects of different display renderings on the perceived quality of complex images are not thoroughly analyzed in literature. Therefore, we try to answer the following question: *How does the HDR display rendering affect the HDR quality assessment, both subjectively and objectively?*

- Most of the SDR objective quality metrics work only on the luminance channel [DVKG$^+$00, WB02, WSB03, WBSS04, SBDV05, SB06, CH07], and the effect of color is generally overlooked in the SDR quality assessment, especially for image and video compression scenarios. However, the increased luminance in HDR conditions can change the way we perceive the quality, and color may influence the perceptual quality due to some aspects of color appearance phenomena, e.g. the Hunt effect, the Bezold-Brücke hue shift, etc. [Fai13]. As we consider compression as our main distortion throughout the thesis, we try to analyze and understand the impact of color on compression and thus ask the following question: *What are the effects of the color space transformation and the related color specific distortions on the HDR quality assessment?*

As we analyze the effects of these two aspects and evaluate the quality metrics in the following chapters, we notice that the subjective evaluation results, i.e. mean opinion scores (MOS), are found to have different ranges for the considered subjectively annotated quality databases. This difference is due to several environmental and experiment-related factors

such as the the training session conducted before the experiment, the aim of the subjective experiment, the range of the distortions, etc. Even though the objective quality of the stimuli are the same, the subjective quality score of a stimulus can be different for different databases. This observation has important results for the subjective quality assessment.

As 'quality' is subjective by its definition, most of the objective quality assessment algorithms use MOS values as ground-truth and find the necessary parameters for their algorithms using these MOS values. In order to use these databases for either the evaluation or development of objective metrics, MOS values need to be aligned. This way, subjective quality scores of two stimuli with the same objective quality scores would be similar. In order to address this issue, we try to answer the following question: *How can we better define a quality scale that would not be affected by the environmental factors and which subjective quality assessment methodology should we use?*

Throughout the thesis, we aim to answer these questions and understand the underlying factors which affect the HDR quality assessment, with a series of subjective experiments and extensive analyses.

Contributions

The following contributions are presented in this thesis. In addition to the contributions discussed in this section, our collaborations throughout France, Spain, and Turkey have brought about some other contributions to HDR image and video coding and perception of brightness in HDR. The complete list of publications is presented in the Publications section (Chapter 7) of the thesis.

- We propose an accurate HDR frame rendering model after detailed characterization of the SIM2 HDR47 display in Chapter 2. The proposed model is based on an iterative scaling algorithm. Experimental results show that the proposed algorithm is able to both reconstruct the HDR images with their intended luminance values and estimate the emitted luminance values accurately. This contribution is explained in detail in the following paper:

  > Emin Zerman, Giuseppe Valenzise, and Frédéric Dufaux, "A Dual Modulation Algorithm for Accurate Reproduction of High Dynamic Range Video", *IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, Bordeaux, France, July 2016.

- We study the effects of display rendering on both human visual perception of compression quality and objective HDR image quality assessment in Chapter 2. Results are analyzed both quantitatively and qualitatively and show that using a simple model of the display response is both necessary and sufficient for objective HDR quality assessment. The details are presented in the following paper:

  > Emin Zerman, Giuseppe Valenzise, Francesca De Simone, Francesco Banterle, Frédéric Dufaux, "Effects of Display Rendering on HDR Image Quality Assessment", *SPIE*

Optical Engineering+ Applications, Applications of Digital Image Processing XXXVIII, San Diego, CA, USA, August 2015.

- We study the effects of color space transformation on the performance of HDR video compression in Chapter 3. These effects are analyzed both subjectively and objectively. Experimental results show that the color space does not have a significant effect on compression performance and luminance-only metrics can predict the results at least as efficiently as the color metrics. This study is presented in the following paper:

  Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, Rafał Mantiuk, Frédéric Dufaux, "Effect of Color Space on High Dynamic Range Video Compression Performance", *9th International Conference on Quality of Multimedia Experience (QoMEX)*, Erfurt, Germany, June 2017.

- We present the results of an extensive evaluation of full-reference HDR image quality metrics in Chapter 4. This evaluation was done using 25 different quality metrics some of which were SDR image quality metrics employed after a pixel encoding step. In total, 690 compressed HDR images, which constitute the largest compressed HDR image dataset to the best of our knowledge, are used in this evaluation. We find that the MOS values coming from different databases need to be aligned, and SDR metrics can perform similar to HDR metrics if pixel values are made perceptually linear.

- We propose a novel method for the evaluation of objective quality metric discriminability in Chapter 4. This proposed method is used with other commonly used statistical analysis methods to evaluate the objective metrics. Both the details of this method and the evaluation results mentioned in the previous item are explained in the following article:

  Emin Zerman, Giuseppe Valenzise, and Frédéric Dufaux, "An Extensive Performance Evaluation of Full-Reference HDR Image Quality Metrics", *Quality and User Experience*, volume 2, April 2017.

- In order to gather more robust subjective scores and to eliminate the need for alignment that was found to be necessary in Chapter 4, we propose to use pairwise comparison scaling results as the subjective quality assessment scores in Chapter 5. Furthermore, we propose to add cross-content pairs to the standard pairwise comparisons subjective quality assessment methodology, which is found to reduce the error accumulation during preference matrix scaling and the confidence intervals of the resulting quality scores. To this end, a comparison of two different subjective methodologies is made. The details are described in the following paper:

  Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, Rafał Mantiuk, Frédéric Dufaux, "The Relation Between MOS and Pairwise Compariosons and the Importance of Cross-Content Comparisons", *IS&T/SPIE Electronic Imaging, Human Vision and Electronic Imaging XXII*, San Francisco, California, USA, January 2018.

In order to better organize the thesis, these contributions are presented within five main chapters which are discussed in the section below.

Structure of the thesis

The thesis consists of five chapters as follows:

- **Chapter 1** provides a clear picture of the image and video quality assessment methods proposed for the standard dynamic range conditions. The details of high dynamic range imaging and the parts of HDR content delivery chain, from acquisition to display, are also discussed in this chapter. Finally, the state-of-the-art in both subjective and objective HDR quality assessment research is explained.

- In order to understand the new distortions that may be induced by the HDR displays, **Chapter 2** discusses and analyzes the effects of display rendering on the HDR image quality assessment. For this purpose, we develop an HDR frame reproduction algorithm after thorough analysis and characterization of the SIM2 HDR47 display. The experimental validation of the proposed rendering method is performed in order to understand its accuracy for reproduction and estimation of emitted luminance. We also analyze the responses of the built-in rendering method of SIM2 display and the developed frame reproduction algorithm, and we compare the effects of these two rendering methods.

- **Chapter 3** analyzes the effects of color space on HDR video compression performance and quality perception, which is generally neglected in the case of SDR. To this end, we design and conduct a subjective experiment with pairwise comparisons methodology, where we compress HDR video sequences with different color space transformations. For this experiment, we select the compressed video sequences to be just noticeably different to capture the minute differences. Then, the performance of compression is measured both objectively and subjectively.

- Based on the findings of the preceding chapters, in **Chapter 4**, the existing full-reference HDR image quality metrics are evaluated. We collect different types of contents and distortions together using four different subjectively annotated HDR image quality databases, in addition to a new database which we create in this chapter. These five different databases are merged by aligning their subjective scores, and a larger set of compressed HDR images is created. In addition to the commonly employed statistical analysis methods, we propose a new method to evaluate metric discriminability, which is based on a classification approach. The objective quality metrics are analyzed using both the statistical methods and the proposed discriminability analysis method.

- **Chapter 5** explains the methods of scaling pairwise comparisons data and the relation between two subjective quality scores: mean opinion scores (MOS) and pairwise comparisons (PC) scaling results. The PC scaling is expected to conceive a "universal scale" on which several experiment results might be compared and fused without a need for alignment of subjective scores as done in Chapter 4. Moreover, we propose to extend the standard pairwise comparisons methodology by including cross-content comparisons. The impact of including cross-content comparisons is assessed subjectively and the findings are presented in this chapter.

The thesis ends with a summary of the experiments conducted, findings, proposed methods, and the results, as well as a number of directions for the future work in this field.

Chapter 1

Background and State of the Art

Contents

The multimedia quality assessment problem is important for many different areas of application. Although it is previously studied for standard dynamic range systems, the increased brightness, contrast, and color range introduced by the high dynamic range technology bring about new conditions and challenges.

This chapter discusses the previous studies on subjective and objective quality assessment, stages of the HDR content delivery from acquisition to display, and the state-of-the-art of HDR quality assessment.

1.1 Subjective Quality Assessment

Although the word “quality” has a meaning of “a distinguishing attribute” or “peculiar and essential character” [qua17b], in the context of multimedia quality assessment, its

definition is closer to "how good or bad something is" [qua17a]. Indeed, in the white paper of COST Action Qualinet, the word "quality" is defined as: "[Quality] [i]s the outcome of an individual's comparison and judgment process. It includes perception, reflection about the perception, and the description of the outcome. In contrast to definitions which see quality as 'qualitas', i.e. a set of inherent characteristics, we consider quality in terms of the evaluated excellence or goodness ... " [LCMP13]. Human perception, and the perceived quality, is relative and subjective. Therefore, the subjective quality assessment is the most reliable and effective way to understand and analyze the multimedia quality.

Perceived quality of the images and videos depend on several factors. These factors may be simple features such as brightness, contrast, pixel resolution, sharpness, etc. or may also be more complex such as the content, aesthetics, color grading, etc. Although it is rather easy to understand the effects of these simple features (as they are used in many subjective studies, see *Subjective Assessment of Image and Video Quality* part in this section), other complex factors such as the artistic intent and aesthetics [MB11] are more challenging they are found to be 'highly subjective' in its nature. In fact, in their images or movies, artists often include noise, reduce the contrast, and change the color tones in order to convey a certain feeling or tone. Ideally, these specific modifications must be preserved during processing [BCMD17b]. However, most applications distort these complex features, e.g. removing intended noise or increasing the contrast, and can lead to a higher quality. A similar observation was made in the subjective quality scores of TID [PIL+13] color image database, where some of the distorted stimuli had higher opinion scores. The target application, or objective, is also important to understand and solve the quality assessment problem. For some applications such as medical imaging, remote sensing, or astronomical imaging the intelligibility can be more important compared to "evaluated excellence or goodness". In some other target applications such as surveillance, it is important that the quality of the media should be both intelligible and good quality at the same time [KO05]. In this thesis, we limit our scope as the multimedia compression scenario and related quality assessment techniques without focusing to a specific target application.

In subjective quality assessment experiments, the multimedia content (*stimuli*) are presented to a group of people (*subjects*), and people are asked to rate or rank the stimuli according to the perceived quality of the stimuli. While some subjective psychovisual experiments use perceptual measurements to understand and analyze the human visual perception of a specific attribute, most subjective quality experiments try to measure the overall impression or overall quality of the presented multimedia content. Depending on the purpose of the experiment and the research question, the responses can be collected using either *direct scaling* (mostly using an interval scale) or *indirect scaling* (e.g. difference threshold, pairwise comparisons, etc.) methods [DS12]. Direct scaling methods ask viewers to determine the quality of the stimuli directly using a categorical or numerical interval scale which can be discrete or continuous. Indirect scaling methods, on the other hand, ask viewers to rank the presented stimuli according to their preferences, or the viewers

are asked to increase or decrease certain parameters until they notice a difference in order to find the threshold or *just noticeable difference* (JND). Most commonly used subjective quality assessment methodologies are discussed below.

Subjective Quality Assessment Methodologies

In order to ensure that the subjective experiment results are collected properly and are relevant, many standards or recommendations were published as guidelines for multimedia or video quality assessment [ITU08, ITU12b, EBU03]. These standards thoroughly describe the requirements for the subjective experiments such as the environmental set-up of the experiment, the procedure and the methodology of the experiment, material selection, etc. Methodologies can be generally classified as single-stimulus, double-stimulus, and comparison methods.

Single-stimulus methods present only one stimulus at a time and ask viewers to rate the quality of the presented stimulus. Some examples include single-stimulus with multiple repetitions (SSMR) [ITU12b], absolute category rating (ACR) [ITU08], and single-stimulus continuous quality evaluation (SSCQE) [ITU12b]. In ACR (or SSMR), the presentation of stimuli and voting are sequential, whereas the voting is continuous in SSCQE and is done alongside representation of the stimulus (generally video) in real time.

Double-stimulus methods present both the reference and the distorted stimulus to the viewers before rating. Viewers may be asked to rate the distortion or rate both of the stimuli. The presentation of the reference and distorted stimulus can be sequential or simultaneously in side-by-side fashion depending on the variant of the methodology. Some examples include double stimulus impairment scale (DSIS) [ITU12b], degradation category rating (DCR) [ITU08], double stimulus continuous quality scale (DSCQS) [ITU12b], and simultaneous double stimulus for continuous evaluation (SDSCE) [ITU12b]. Similar to the case of SSCQE, the voting is continuous in SDSCE and is done simultaneously with the representation.

Comparison methods present viewers two or more stimuli and ask them to compare the presented stimuli. The presented stimuli can be voted with several levels (from 'Much worse' to 'Much better') to indicate preference or relation to the other stimulus [ITU12b]. Alternatively, the subjects may be asked to prefer one of the stimuli (two alternative forced choice) or, additionally, subjects may be allowed to select the "Same" option (three alternative forced choice). The comparison methods are called pairwise comparisons (PC) methods when there are only two stimuli compared at a time. Comparison methods are more suitable to the cases where the visual difference between two stimuli is small. Although they are generally used to understand the preference between two processing methods, they are also used for other purposes such as finding JND points [LJH+15, JLH+16] or estimation of quality scores [Thu27, BT52, LDSE11] (PC scaling is further discussed in Section 5.1).

The subjective assessment of multimedia video quality (SAMVIQ) [EBU03] is an alternative method which is a mixture of single-stimulus, double-stimulus and comparison methods. Viewers are presented a number of stimuli, and they can select a stimulus to rate. Viewers can select the stimuli in their order of preference, go back, compare, and correct their votes. This unique property of SAMVIQ makes it a multi-stimulus methodology.

The acquired subjective data is analyzed generally by finding the mean opinion score (MOS) and standard deviation for the presented stimuli [ITU12b, ITU08, ITU12c]. Depending on the methodology, MOS values are found either by taking the mean of the opinion scores (e.g. ACR), or taking the mean after subtracting the opinion scores from each other (e.g. DSCQS). The weaknesses and strengths of these different methodologies were compared by many researchers as discussed more in the next section.

Comparison of Subjective Quality Assessment Methodologies

There has been a substantial amount of work comparing different methodologies for the subjective quality assessment. In [PW03a], Pinson and Wolf compared single-stimulus and double-stimulus continuous quality evaluation methods (SSCQE and DSCQS) and found that the quality estimates are comparable to one another. In [THOT10], ACR, DSIS, DSCQS, and SAMVIQ were compared. The authors found no significant difference between the compared methods. The compared methods were also ranked for the assessment times and the ease of evaluation. It was found that from fastest to slowest, the ranking was ACR, DSIS, SAMVIQ, and DSCQS. The ease of evaluation analysis yielded a similar result with the exception that ACR with 11-point scale was the hardest to evaluate whereas ACR with 5-point scale was the easiest. SAMVIQ and ACR were further compared in [RPLCH10], and SAMVIQ was found to require fewer subjects and longer time compared to ACR. In the study of Mantiuk et al. [MTM12], four different subjective methods were compared: single-stimulus categorical rating (absolute category rating with hidden reference (ACR-HR)), double-stimulus categorical rating, forced-choice pairwise comparison, and pairwise similarity judgments. No significant difference was found between double-stimulus and single-stimulus methods, in agreement with the previous studies. The forced-choice pairwise comparison method was found to be the most accurate and requiring the least experimental effort amongst the four compared methods.

The methodology of a subjective experiment depends on the intent and research problem. Although direct rating methods are able to obtain quality scores directly, ranking methods such as pairwise comparison offer additional preference information.

Subjective Assessment of Image and Video Quality

Subjective assessment is used for many applications of computer graphics, computer vision, and multimedia signal processing, such as perception of visual artifacts in image rendering [VLD07, VCL+11], image editing and computer vision [CFL+15, XPCH17], view con-

version and segmentation [ASAU12, CHA+14], perception of artifacts like blur [MDWE02, CDLN07] and blocking [KMFH05, MK05], and compression and multimedia applications [ABA05, BLCCC09, SJPK+10, SSBC10a, HRDSE12, CCPP15]. In addition to its usage for assessing subjective quality of certain applications, a number of subjectively-annotated databases were created after subjective assessment studies [Win12]. Some of these databases focused on image quality, and they are widely used: Laboratory of Image and Video Engineering (LIVE) Image Quality Database [WBSS04, SSB06], Computational and Subjective Image Quality (CSIQ) [LC10], and Tampere Image Database (TID) [PLZ+09, PJI+15]. Some others focused on video quality. This group includes Video Quality Experts Group (VQEG) FR-TV Phase I [VQE00, RLCW00], VQEG HDTV [PS10], IRCCyN/IVC 1080i [PPLC08a, PPLC08b], IRCCyN/IVC SD RoI [FBC09, BCPLC09], PoliMI-EPFL Video Quality [DSNT+09, DSTN+10], Poly NYU [OZW10], and LIVE Video Quality [SSBC10b, SSBC10a] databases.

There is a very strong link between the subjective studies and objective quality assessment methods. In fact, most of these databases are used as 'ground-truth' data for objective quality assessment metrics. Therefore, it is very important to understand the relationship between the perception of quality and the representative quality score.

1.2 Objective Quality Assessment

Despite their accuracy, subjective quality assessment experiments are expensive to conduct in terms of time and resources. The quality of the presented stimuli can also be estimated with the help of algorithms and computers, and it is called objective quality assessment. Although not as precise as subjective quality assessment methods, objective quality assessment methods are much faster and crucial for many applications.

Objective quality metrics are classified into three categories according to the availability of the undistorted reference. Full-reference (FR) metrics require the reference in order to estimate the objective quality of the stimulus. Reduced-reference (RR) metrics need only a part of the reference information such as edges, areas, etc. No-reference (NR) metrics do not require any reference information. In this thesis, we only consider full-reference objective quality metrics as it is the most commonly used type of objective quality assessment metric within the compression framework.

For the case of SDR, plenty of objective quality assessment metrics were proposed. In the following subsections, we discuss some of the most popular metrics for image and video quality used in this thesis. For a thorough analysis of objective quality metrics, interested readers can refer to [Win05], [WB06], and [LJK11].

1.2.1 Image Quality

Full-reference image quality metrics estimate the quality of an image with respect to its reference. Most of these quality metrics work only on the luminance channel of the images. The most popular image quality metrics are discussed below in four classes: simple arithmetic, structural, information-theoretic, and color difference metrics.

Simple Arithmetic Difference Metrics

Mean squared error (MSE) and peak signal-to-noise ratio (PSNR) are two most commonly used simple arithmetic difference metrics. MSE is calculated as:

$$MSE = \frac{1}{width \times height} \sum_{i=1}^{width} \sum_{j=1}^{height} \left(I_{\text{Reference}}(i,j) - I_{\text{Test}}(i,j)\right)^2 \tag{1.1}$$

where i is the column index and j is the row index, $I_{\text{Reference}}$ and I_{Test} are the reference and test images with image *width* and *height*, respectively. Although MSE is not well correlated with human perception and its use is a topic of debate [WB09], it is still widely used in many applications where a simple and quick image difference metric is required.

PSNR is also popular especially for very specific applications where the type and magnitude of the distortion are within some range, such as image and video compression. However, it has the same drawbacks, as it is a logarithmic representation of MSE:

$$PSNR = 10 \log \left(\frac{{max_I}^2}{MSE} \right) = 20 \log \left(\frac{max_I}{\sqrt{MSE}} \right) \tag{1.2}$$

where max_I is the maximum pixel value of the image I.

Structural Similarity Metrics

Universal quality index (UQI) [WB02], structural similarity index (SSIM) [WBSS04], and multiscale SSIM (MSSIM) [WSB03] can be classified together as they all measure structural similarity. In order to estimate objective quality, UQI uses simple statistical parameters of the image pixel value such as mean, variance and covariance of the pixel value between two images. These parameters are arranged to find the correlation of pixel values between the images, as well as the change in image luminance and image contrast. To find a quality estimate, these terms are multiplied.

SSIM is an extension of UQI metric. Similar to the UQI, mean signifies luminance, variance signifies contrast, and covariance signifies the structural similarity between the reference and the test image. Additionally, some constant variables are included in order to regularize the quality score estimated by the metric. MSSIM further extends SSIM by making computations on multiple scales of image. It calculates the contrast and structure comparison parameters for different scales of the image and multiplies them with the

luminance comparison parameter.

Information-Theoretic Metrics

The information fidelity criterion (IFC) [SBDV05], visual information fidelity (VIF) [SB06], and its pixel-based version (VIFp) use information theoretic approaches. These metrics analyze the natural scene statistics in order to estimate the quality of the given stimulus. For this purpose, IFC estimates the quality scores by calculating the sum of the conditional mutual information $I(C^N; D^N|s^N)$, which is computed considering a source, C^N, and distortion model, D^N, for different wavelet decomposition subbands. VIF also includes a human visual system (HVS) model in addition to the source and distortion models in order to include the uncertainties in the human visual perception of images, and this uncertainty is modeled as an additive Gaussian noise.

Color Difference Metrics

In order to find the differences between colors, the International Commission on Illumination (Commission Internationale de l'Éclairage - CIE) developed CIE1976 (ΔE^*_{ab}) [CIE86] color difference metric which finds the Euclidean distance between two colors using $L^*a^*b^*$ color space. It assumes that $L^*a^*b^*$ color space is perceptually uniform, where the unity Euclidean distance would correspond to the same amount of perceptual difference. In order to address the nonlinearity of $L^*a^*b^*$ color space, this metric was improved with the developments of CIE94 (ΔE_{94}) [CIE95] and CIE DeltaE 2000 (ΔE_{00}) [LCR01].

Additionally, ΔE_{00} is further extended to be computed on S-CIELAB instead of CIELAB [ZW97]. In their paper, Zhang and Wandell proposed to separate color image into luminance and color opponent (R-G and B-Y) channels and filter spatially before the color conversion to CIELAB. Then, the CIE ΔE_{00} is computed on spatially filtered S-CIELAB color space, and we denote this spatial extension as ΔE^S_{00}. According to a recent study [OJKP16], ΔE_{00} and ΔE^S_{00} are found as the best performing color difference metrics.

1.2.2 Video Quality

Objective video quality assessment is harder compared to image quality assessment because of the need to consider the temporal variations of the video. The perception of video is also different compared to the case of image due to the response of human visual system and its properties such as eye fixation duration and visual short-term memory. Even though special care is required for objective video quality assessment, image quality algorithms are also commonly used by pooling the frame-by-frame results. Averaging pixel values over the frames is one of the most popular pooling methods used for this purpose.

Among many other works for full-reference video quality assessment [WM08, MB11], the two most commonly used video quality metrics are 'video quality metric' (VQM) [PW04] and motion-based video integrity evaluation (MOVIE) [SB10].

VQM computes seven parameters through several filtering operations. These parameters are collected to analyze different aspects of video quality such as the differences in edge information (blurring, edge sharpening, or edge enhancement via `si_loss` and `si_gain`), shift in edge orientation (either for edges lost with blurring or edges created as a result of blocking via `hv_loss` and `hv_gain`), color impairments (via `chroma_spread` and `chroma_extreme`), and temporal impairments (by multiplying contrast information and temporal information via `ct_ati_gain`). These parameters are then combined linearly to find the VQM quality score.

MOVIE uses separable Gabor filterbanks to filter the video sequences. In order to find the spatial error, the mean squared normalized error (error between the filtered video sequences is normalized with a division operation) for each sub-band of the filtered video sequence is found. This per sub-band error is combined with another error term where Gaussian filtered video is used, and these error terms are pooled together to find Spatial MOVIE index. The Temporal MOVIE index is found by pooling the temporal distortion, which is computed using the optical flow information of the reference video and Gabor filtered video sequences, for each frame. The overall MOVIE quality scores are found by multiplying the Spatial MOVIE and Temporal MOVIE indices.

1.2.3 Evaluation of Quality Metric Performance

In order to measure the performance of an objective quality metric, several methods are used [ITU04a, ITU12c, ITU04d]. These methods can be divided into two categories. The first category includes the most commonly used statistical evaluation methods, and the second category includes alternative methods for the evaluation of objective quality metric performance.

Statistical Evaluation Methods

Statistical evaluation methods are used by an overwhelming majority of studies on quality assessment research. These methods are described in the ITU-T Recommendation P.1401 [ITU12c] in detail. The objective quality metric results, using either a linear, polynomial, or non-linear function, are fitted to subjective quality scores, and these fitted (i.e. predicted) quality scores are used to compute the statistical evaluation metrics described below. According to the ITU Recommendations [ITU12b, ITU04a, ITU04d], the performance of objective quality metrics is characterized by the following three attributes:

- Prediction accuracy
- Prediction monotonicity
- Prediction consistency

Prediction accuracy is computed by two evaluation metrics: Pearson correlation coefficient (PCC) and root mean squared error (RMSE). PCC finds the correlation between two sets of data, which are the subjective quality score and the predicted quality score for metric evaluation problem, and it can be interpreted as the linearity of the relation. PCC is calculated as:

$$PCC = \frac{1}{n-1} \frac{\sum_{i=1}^{n}(S_{subj,i} - \widehat{S}_{subj,i})(S_{pred,i} - \widehat{S}_{pred,i})}{\sigma_{S_{subj,i}} \sigma_{S_{pred,i}}} \tag{1.3}$$

where i is the stimulus index, n is the number of stimuli, $S_{subj,i}$ and $S_{pred,i}$ are the subjective and predicted quality scores, respectively, $\sigma_{S_{subj,i}}$ and $\sigma_{S_{pred,i}}$ are the standard deviation for subjective and predicted quality scores, and $\widehat{S}_{subj,i}$ indicates the mean of $S_{subj,i}$. RMSE is the absolute prediction error, and it can be interpreted as it assumes linearity to find the accuracy of the given predicted (objective) quality scores compared to the subjective quality scores. RMSE is calculated as:

$$RMSE = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (S_{subj,i} - S_{pred,i})^2} \tag{1.4}$$

Ideally, the relationship between subjective and predicted quality results should be monotonic. That is, the predicted quality scores should increase when there is an increase in subjective quality scores. Spearman rank-order correlation coefficient (SROCC) is commonly used to find the prediction monotonicity. As it does not assume any relationship between the predicted and subjective quality scores, it is also invariant to the non-linear fitting. SROCC is calculated as:

$$SROCC = \frac{\sum_{i=1}^{n}(R_{subj,i} - \widehat{R}_{subj,i})(R_{pred,i} - \widehat{R}_{pred,i})}{\sqrt{\sum_{i=1}^{n}(R_{subj,i} - \widehat{R}_{subj,i})^2} \times \sqrt{\sum_{i=1}^{n}(R_{pred,i} - \widehat{R}_{pred,i})^2}} \tag{1.5}$$

where $R_{subj,i}$ and $R_{pred,i}$ are the ranks of $S_{subj,i}$ and $S_{pred,i}$, respectively, and $\widehat{R}_{subj,i}$ and $\widehat{R}_{pred,i}$ are the mean ranks of $S_{subj,i}$ and $S_{pred,i}$.

Prediction consistency is another attribute that needs to be checked in order to determine the performance of an objective quality metric. In addition to the other attributes, it is important to understand how consistent the prediction results are. For prediction consistency, outlier detection (OR) is used. It is calculated as:

$$OR = \frac{\text{Total number of outliers}}{n} = \frac{\sum_{i=1}^{n} f_{outlier}(i)}{n} \tag{1.6}$$

where $f_{outlier}$ is the count function which counts the outliers:

$$f_{outlier}(i) = \begin{cases} 1 & \text{if } |S_{subj,i} - S_{pred,i}| > 1.96 \times \sigma_{S_{subj,i}} \\ 0 & \text{otherwise} \end{cases} \tag{1.7}$$

In order to understand the performance of the objective quality metrics, the numerical results of the PCC, RMSE, SROCC, and OR are reported in many studies [SSBC10a, PJI+15, HBP+15, HRE16]. Without any other information, the differences may not be clear and may be misleading in certain cases. Therefore, it is important to understand the significance of these differences. In ITU-T Recommentation P.1401 [ITU12c], guidelines for the significance analysis of PCC, RMSE, SROCC, and OR are given. In addition to the numerical results of the statistical evaluation methods, significance analysis results are also necessary to understand the impact of the difference between two objective quality metrics.

Alternative Evaluation Methods

The statistical evaluation methods (except RMSE* [ITU12c]) assume the subjective quality scores as deterministic 'ground-truth' data. However, the subjective quality scores are probabilistic as they are collected from a sample set of human population. The alternative evaluation methods described in this part treat subjective quality scores as random variables, and therefore, they present a better understanding in evaluating quality metrics.

Brill et al. [BLC+04] proposed a method to analyze the performance of the objective quality metric and find the minimum significant objective quality score difference, also called *"resolving power"*.

The steps of this algorithm are briefly described in the following: First, the objective quality scores (S_{obj}) are converted to a common scale (S_{pred}) by fitting the scores to a 4^{th} order polynomial function, and the quality difference $\Delta S = S_{pred,k} - S_{pred,l}$ is found. To find the probability of significance to ΔS, subjective results are used in a one-tailed z-test. The probability of significance is then found by sweeping the threshold for ΔS value and the $\overline{\Delta S}$ value corresponding to the 95% significance probability, $p = 95\%$, is selected. Using the selected quality difference value, the classification rates are found as:

- False Tie, i.e. $S_{subj,k} \not\equiv S_{subj,l}$ and $S_{pred,k} \equiv S_{pred,l}$
- False Differentiation, i.e. $S_{subj,k} \equiv S_{subj,l}$ and $S_{pred,k} \not\equiv S_{pred,l}$
- False Ranking, i.e. $S_{subj,k} < S_{subj,l}$ and $S_{pred,k} > S_{pred,l}$ or vice-versa
- Correct Decision

where $k \neq l$, the $\equiv$ sign represents the statistical equivalence (i.e. $S_{pred,k} \equiv S_{pred,l}$ if $S_{pred,k} - S_{pred,l} < \overline{\Delta S}$), and $S_{pred,k}$ is the objective quality metric prediction of MOS for data point k.

This method is used in several other studies [PW08, Bar09, HŘE15, NVH16] and standardized in ITU Recommendations [ITU04c, ITU04b]. Nevertheless, it has some drawbacks. It needs polynomial fitting, and fitting may not be successful in some cases. Also, the resolving power is defined as a function of the objective metric result itself, i.e. $RP(S_{obj}) = |f^{-1}\left(f(S_{obj}) + \overline{\Delta S}\right) - S_{obj}|$, where $RP(\cdot)$ is the resolving power function and $f(\cdot)$ is the fitting function. Thus, it is not trivial to analyze a variable function for a large number of metrics. Instead, it enables graphical comparison of the classification performance of the objective metric.

Another method for evaluating objective metric performance was proposed by Krasula et al. [KFLCK16]. Similar to the method of Brill et al. [BLC+04, ITU04b], z-scores are calculated for subjective quality scores. The statistical equivalence (or statistically significant difference) of these subjective quality scores is determined by checking whether the probability of two stimuli being different from each other is greater than 95%. This probability is calculated using the cumulative distribution function (CDF) of the normal distribution. The results indicate whether two stimuli are equivalent and which one is better if they are different. Without any fitting or preprocessing, the objective score differences are calculated. These objective metric differences are then used to find the receiver operating characteristics (ROC) for each objective quality metric. The ROC values and correct classification rates are then used to determine if an objective quality metric is better or worse than, or equal to, the other objective quality metric. In addition to the different/similar and better/worse analyses, a statistical significance method was proposed in this work.

In parallel to the work of Krasula et al. [KFLCK16], we also developed an alternative metric evaluation method similar to their method. The similarities and differences are discussed in Section 4.3.3 where we also explain our method.

The quality assessment methods described above are used for SDR display systems; however, HDR changes some of the assumptions of SDR technology fundamentally. Therefore, the HDR imaging and content delivery is discussed in the next section to provide an understanding of HDR technology.

1.3 HDR Imaging and Content Delivery

The human visual system is capable of perceiving much larger range of brightness, contrast, and color compared to those SDR systems can offer. Legacy SDR displays can display up to 250 cd/m^2 and have contrast ratio of $\sim 1000:1$[1] whereas humans can see up to 9 magnitudes of luminance range (contrast ratio $\sim 10^9:1$) [Hoe07] with a simultaneous dynamic range higher than the SDR display (when the eyes adapt to a certain range – see Figure 1.1).

[1] A standard Dell 20 (E2016H) monitor was taken as reference. `http://www.dell.com/en-ie/shop/dell-20-monitor-e2016h/apd/210-afpd/monitors-monitor-accessories`. Accessed online: 02/12/2017

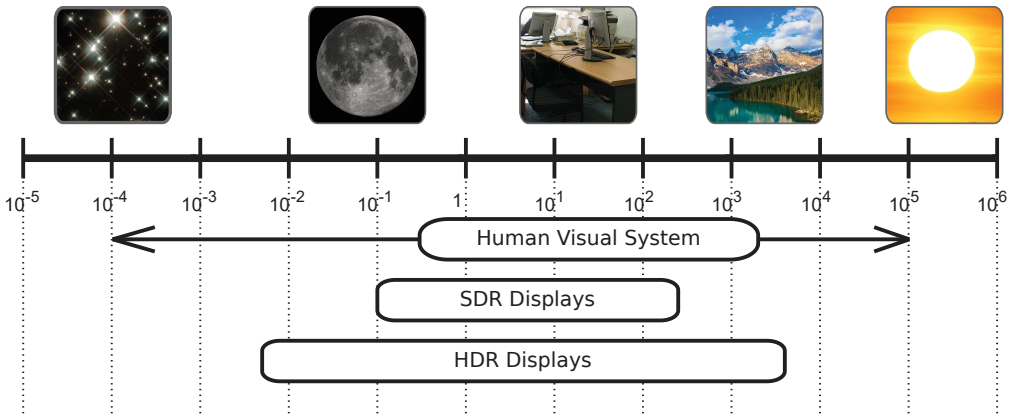


Figure 1.1 – Comparison of SDR and HDR display systems with respect to the real world luminance values and to the capabilities of human visual system. (Values of the scale are in cd/m^2.)

Until recently, the SDR display framework was the limiting factor for SDR imaging and content delivery. Because of the displays, all captured content were converted to 8-bit integer pixels even though the camera sensors are able to capture a wider range. The compression systems were also adapted to the 8-bit (or 24-bit with 3 color channels) structure of images and videos.

With the advancements in HDR displays, HDR technology removes some of these limitations and lets us to capture, store, transmit, and reproduce images and videos with a larger range of luminance and color. In order to understand the challenges and the limitations of HDR quality assessment, we need to understand the core concepts of HDR imaging and content delivery. Therefore, in this section, we discuss the main points of HDR frame acquisition and storage, HDR image and video compression, reproduction and display of HDR content.

1.3.1 Acquisition and Storage

Although many camera sensors have the capability of capturing higher dynamic range content, the frame acquisition pipeline [RSYD05] produces *"display-referred"* (also called *"output-referred"*) content. That is, the images or video frames are captured and processed for a specific set of output devices (in this case, SDR displays). Using the techniques discussed in this subsection, most of the HDR imaging techniques/devices aims to capture *"scene-referred"* content [MKMS07], which would have the values proportional to the physical luminance values of the scene itself.

Acquisition

In order to capture the luminance values of the scene, both the darker parts and the brighter parts of the scene should be acquired correctly. For this purpose, different techniques are

used. HDR frame acquisition techniques can be divided into three [UHK16] according to the method used to capture the natural scene:

- *Temporal methods*

Temporal methods capture multiple images with different exposures in order to get the information from both dark and bright parts of the scene. Several approaches exist for traditional temporal bracketing [MP95, DM97, MN99, GN03, RBS03, LGYS04]. A detailed explanation for these methods can be found in [GS16]. These methods are easy to use but take time to capture an HDR image. Therefore, they are susceptible to different sources of disturbances such as misalignment and ghosting. As these methods require time, the camera can move during the capture and the captured images may need to be aligned [War03, TM07]. Similarly, ghosting effects may occur if moving objects are present in the scene, and de-ghosting may be required to ameliorate the image [ZBW11, GKTT13, KAR16].

- *Spatial methods*

In order to reduce the time required, the combination of different exposures can be done spatially. These spatial methods decrease the time required to capture an HDR image in exchange for increased noise and decreased spatial resolution. Nayar and Mitsunaga [NM00] proposed high dynamic range image capture using spatially varying sensors. Using neutral density filters, these sensors are able to capture the scene in different levels of exposures. The acquired image is then processed to yield the HDR image. Rolling shutter is also used to capture HDR images. In their work [GHMN10], Gu et al. proposed two different methods for this purpose. One of these methods uses auto exposure by finding the exposure times of each row adaptively. Their other method uses specific readout and exposure scheme which reads three consecutive lines in different exposure times in order to generate HDR image and handle motion blur.

- *Optical methods*

Optical methods change the amount of light directed to a sensor or camera using either an optical device called beamsplitter or a semi-transparent mirror. This way, the image captured by the sensor receiving a small percentage of the light can be considered as short exposure image and the image captured by the sensor receiving a larger percentage of the light can be considered as long exposure image. This method was employed using two cameras during the capture of Stuttgart HDR video dataset [FGE+14]. The beamsplitters can also be used within the camera [AA04], and multiple sensors can be used to receive different percentages of light [TKTS11].

In addition to these methods, another very important and historically prior HDR content acquisition method is the generation of HDR content using computer graphics software such as 3D modeling and rendering [War94, Pau02]. These computer generated images simulate the light absorbed, diffused, and reflected by the objects and create a realistic scene.

Storage

HDR images are either generated by computer graphics rendering software or captured by HDR acquisition methods discussed above. For further processing (e.g. post-processing, compression, transmission), the scene-referred pixel values need to be stored in floating point numbers. However, using floating point numbers for each pixel is costly (96 bits-per-pixel if single-precision is used). Three different pixel encoding and storage formats are commonly used for this purpose: RGBE, TIFF, and EXR. These file formats are used for HDR images and video frames alike, as there is no other specific file format for HDR video at the time of writing this thesis.

As proposed in [War91, War94], the Radiance RGBE format (which uses `.hdr` file extension) represents each pixel with 32 bits. As its name implies, these bits are divided between red, green, blue, and exponent channels where each of them has 8 bits. The same exponent value is used for all of the three mantissa parts of R, G, and B channels, and the largest mantissa has its leftmost bit set to 1 (i.e. the largest mantissa is between 128 and 255). The 32-bit pixels are encoded with run-length encoding. This file format can also be used with CIE XYZ color space to store color values without the limitation of RGB (BT.709 [ITU15a] or BT.2020 [ITU15b]) color gamut. In that case, the file format can be also called XYZE.

Tagged image file format (TIFF) is another file format used for storing HDR images. In the LogLuv TIFF HDR file format [Lar98b, Lar98a], the colors are converted to CIE (u',v') color coordinates, and luminance is found by taking the logarithm of the Y channel of CIE XYZ converted image. It uses either 24-bit (10 bits of luminance and 14 bits for combined chrominance channel) or 32-bit (16 bits of luminance and 8 bits for each chrominance channel) structure. The resulting image is stored in a TIFF image.

OpenEXR [BKH03] is a commonly used open-source HDR image file format (which uses `.exr` file extension). Each pixel consists of three half-precision floating point numbers (16-bit per-channel and 48-bit per-pixel) for three color channels, including a sign bit, 5 bits of exponent, and 10 bits of mantissa. Although the required size for the same image is greater compared to RGBE and LogLuv TIFF, OpenEXR can store a wider range of luminance and color thanks to its ability to store negative numbers. In addition to its precision, it can support lossless and lossy compression.

1.3.2 HDR Image and Video Compression

The wider range of luminance and color of HDR comes at the cost of large amount of data which is difficult to store, transmit and reproduce. Therefore, efficient compression algorithms are necessary for storing and transmitting the HDR content. Although the file storage formats discussed above apply lossless compression to the HDR images, lossy compression is also necessary to meet the needs of transmission channels for the images and videos. In this subsection, we discuss the image and video compression systems for

HDR content.

HDR Image compression

In order to compress HDR images, several compression methods were proposed throughout the last decade [WS06, STZ+07, KE13, Bol14, AMR+15]. All of the HDR image compression algorithms discussed here are backward-compatible.

In order to compress HDR images in a backward-compatible fashion, Ward and Simmons [WS06] proposed JPEG-HDR. This method uses tone-mapping which is dynamic range conversion from HDR to SDR image or video. In this encoding scheme, the HDR image is tone-mapped to an 8-bit SDR image and the residual image is found. Both the tone-mapped image and the residual image are encoded using standard JPEG compression algorithm, and the compressed tone-mapped image is placed in base layer whereas the compressed residual image is placed in the extension layer. This encoding scheme can be extended to use other backward-compatible image compression algorithms such as JPEG 2000, as it is done in [VDSLD14].

JPEG XR [STZ+07] and JPEG 2000 [Bol14] are other compression methods that can accommodate high bit-depth images which may include HDR images. Korshunov and Ebrahimi [KE13] also proposed an JPEG backward-compatible compression method for HDR images. They apply three simple tone-mapping operators, i.e. linear, logarithm, and gamma, and they compare the proposed method to three other HDR image compression schemes discussed above, i.e. JPEG-HDR, JPEG 2000, and JPEG XR.

JPEG-XT [AMR+15, AMR+16] (ISO/IEC 18477) is the backward-compatible HDR image compression standard of Joint Photographic Experts Group (JPEG). In addition to the enhancement layer, it generates a base layer bitstream which is compatible with legacy SDR JPEG decoder. In order to generate the base layer, the HDR image is tone-mapped and compressed with standard JPEG encoder. The residual image is then processed and encoded according to one of the several profiles chosen. The quality levels for both base layer and enhancement layer can be selected separately. Detailed information on JPEG XT can be obtained from [Ric13, AMR+15, Ric16, AMR+16].

HDR Video compression

Most of the video compression methods, as well as the state-of-the-art compression algorithms [TLSS09, SOHW12], are developed for SDR systems using 24-bit SDR images. Furthermore, these compression methods assume that the relationship between the real world luminance and the electrical signals are perceptually proportional. Although this assumption holds for the case of SDR, the perceptual uniformity of HDR is found to be different from the SDR case [AMMS08].

The relationship between the optical and electrical signals are defined through opto-electronic transfer (OETF) and electro-optical transfer (EOTF) functions. For legacy SDR

systems, a power-law EOTF [ITU11] was designed for cathode ray tube (CRT) systems for gamma correction, and until recently, all of the displays were using the same EOTF. However, the wider luminance range of HDR requires a more sophisticated EOTF and OETF for HDR display systems. For this purpose, perceptual quantizer (PQ) [MND12, SMP14] EOTF and hybrid log-gamma (HLG) [Bor14] OETF are proposed, and they are recommended for use in HDR television systems [ITU17a, ITU17b]. Particularly, PQ EOTF finds widespread use within the multimedia community. In the call for evidence (CfE) of Moving Picture Experts Group (MPEG) for HDR/WCG video compression [LFH15], PQ EOTF is used in the compression of anchor HDR video bitstreams. Additionally, some researchers focus on the development of an adaptive PQ [YJK16, LSH+17] in order to improve its performance.

The compression of HDR videos can be done using either backward-compatible or non-backward-compatible [MDBR+16] algorithms. The former [WS04, MEMS06, LK08, FKZ+17] generally use a tone-mapping operator (TMO) in order to generate a base layer stream which can be viewed in SDR displays, and a residual stream which contains additional information for HDR video decoding. The latter [Lar98a, MKMS04, GT11] can encode videos with high bit-depth quantization and employs state-of-the-art video encoders [TLSS09, SOHW12].

In addition to generating an SDR bitstream, backward compatible HDR video compression algorithms also consider temporal information to reduce any temporal artifacts, such as flickering and color change, and generate an enhancement/residual layer for conversion to HDR. In [MDBR+16], the HDR image encoding algorithm proposed by Ward and Simmons [WS04] was extended for video compression by using a newer version of photographic TMO [KRTT12] which considers temporal changes. The original algorithm uses photographic TMO [RSSF02] to extend standard JPEG and stores the ratio between the HDR image and tone-mapped image. Lee et al. [LK05] proposed an HDR video compression algorithm which uses a temporally consistent TMO [LK07] in order to create SDR bitstream, and a ratio stream is found similar to the *hdrjpeg*. This ratio stream is then filtered with a bilateral filter to reduce the noise. Mantiuk et al. [MEMS06] also proposed a backward compatible video compression algorithm which encodes the tone-mapped [RSSF02] SDR frames using MPEG-4 encoder. The residual frames are then processed to remove noise and encoded with MPEG encoder as well.

Although layer-based coding approaches are suitable and popular methods for backward compatible video compression, the coding approach (or structure) and backward compatibility is not directly correlated. Despite using a TMO-based encoding scheme, the work of Ozcinar et al. [OLVD16] does not have any constraint on the quality or plausibility of the SDR stream generated by their algorithm, and focus on the quality of decoded HDR frames. The SDR images (or video frames) generated using a TMO or pixel transformation for the purpose of HDR compression may not be always visually pleasing. Nevertheless, the generation of visually pleasing (including the artistic intent of the original scene) SDR images or videos is very important for backward-compatibility [KD13, MMNW13, GRG+16].

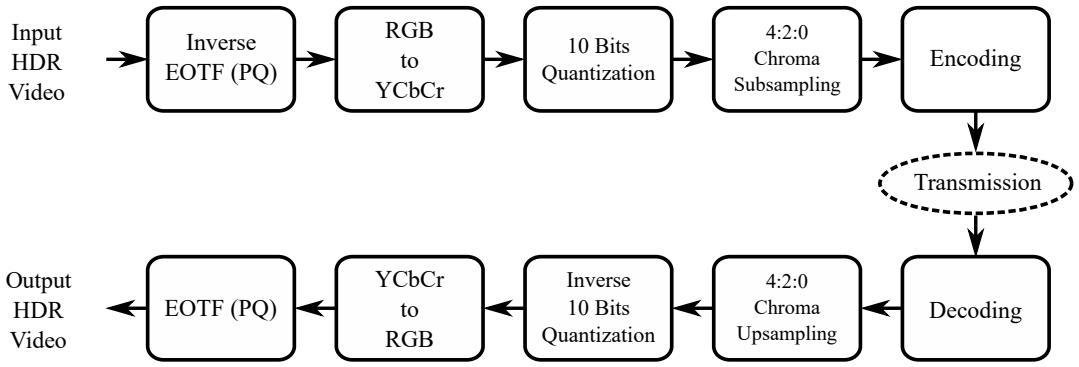


Figure 1.2 – HDR video compression pipeline as described in MPEG CfE [LFH15], and standardized in ITU-R BT.2100 [ITU17b]

In [GRG+16], Gommelet et al. proposed a TMO in order to preserve the SDR image quality and rate-distortion performance of a scalable coding approach for HDR content. They also found that even though it performs well in terms of the quality of reconstructed HDR frames, TMO proposed by Mai et al.[MMM+11] does not yield good objective quality compared to the reference SDR. Photographic TMO [RSSF02] is one of the TMO found to perform well for both visual appeal and visual quality [LCTS05, GRG+16], and it is used by many of the backward-compatible video compression algorithms discussed above [WS04, MEMS06, GRG+16, MDBR+16].

Non-backward compatible algorithms commonly map much wider range of luminance to a more compact range using a transform function such as PQ, logarithm, or a custom perceptual quantization function [MKMS04]. They also use a higher bit depth compared to backward compatible methods. Mantiuk et al. [MKMS04] proposed an HDR video coding method which converts the color space to Lpu'v' and extends MPEG-4 encoder to encode HDR content. Computed similar to LogLuv [Lar98a], Lpu'v' includes 11-bit luma channel and 8-bit u' and v' chroma channels. Garbas and Thoma [GT11] also proposed another method for encoding HDR videos. The Adaptive LogLuv algorithm [MT10] was modified and, in order to reduce flickering, temporal consideration was included. The colors of HDR video frame were converted to Lu'v' color space similar to LogLuv [Lar98a] before 12-bit luma and 8-bit u' and v' chroma channels are encoded.

In addition to these studies, the encoding scheme described in the MPEG CfE [LFH15] can also be considered as a non-backward compatible compression method. The compressed bitstreams need to be converted back to HDR pixel values by an EOTF, and they cannot be displayed after decoding. Due to its use of High Efficiency Video Coding (HEVC) Main 10 profile, this method is also called as *"HDR10"*, and its core components are standardized in ITU-R Recommentation BT.2100 [ITU17b], which include 10-bit or 12-bit bit depth; ITU BT.2020 [ITU15b] wide color gamut (WCG) color space; use of RGB, Y'CbCr, ICtCp (or ITP) [LPY+16] signal formats; and PQ [MND12, SMP14] or HLG [Bor14] optical transfer

functions. An example of content delivery chain for HDR10 can be seen in Figure 1.2. Dolby also recently published [Dol17] their own HDR content delivery format using 12-bit depth and dynamic metadata in addition to PQ EOTF and BT.2020 color space.

1.3.3 Reproduction and Display

Due to the brightness and contrast requirements, HDR reproduction and display is one of the most challenging aspects of HDR technology, and it was not possible until recently. Because the contrast of LCD is limited, locally dimmable backlights are used for HDR content reproduction [SHS+04]. This separation of brightness generation and color is called *dual modulation* [NDSLC16a], and the most widespread dual modulation method is 2D dimming via LED backlights and LCD panels.

Reproduction of HDR content may need processing, depending on the capabilities of the display device. In this subsection, we will briefly discuss the tone-mapping techniques and HDR display rendering methods.

Tone-Mapping Techniques

In order to reproduce captured and transmitted HDR scenes on a display with limited dynamic range, tone-mapping techniques are employed. The target media can be an electronic display which presents images and videos, or printed material. Regardless of the target media, tone-mapping techniques can be used to re-target the dynamic range of the HDR content. The tone mapping methods can be classified into two categories: global and local.

Global tone mapping algorithms affect the whole image, and the same operation is applied to all of the pixels. In general, these algorithms can be easy to implement, and they can work in close-to real-time. These algorithms can be very basic simple operations such as linear, logarithmic, or exponential functions [BADC11]. The global methods can have different objectives such as brightness reproduction [TR93], modeling visual adaptation [FPSG96], histogram adjustment [LRP97], adaptive log mapping [DMAC03], minimization of reproduction error [MMM+11], or display adaptive presentation [MDK08].

Local methods, on the other hand, affect parts of image as the tone-mapping operator is applied on a neighborhood of pixels at a time. These algorithms can preserve local contrast and can look more appealing. However, the selection of kernel size is crucial as it can create halos around the edges. Some examples include the work of Chiu [CHS+93], Pattanaik [PFFG98], Reinhard [RSSF02], and Ashikmin [Ash02].

For the video tone-mapping, a special care needs to be taken due to the temporal variation among video frames. In order to avoid strong temporal artifacts, temporally coherent TMOs should be developed [BCTB13]. For an exhaustive analysis of tone-mapping algorithms, interested readers can refer to [BADC11, BCTB13, EMU17].

HDR Display Rendering

Dual modulation algorithms aim to modulate the backlight and the color components separately. For reproduction of HDR content, the most commonly used scheme is LED-LCD dual modulation. LED values can be modified by rendering algorithm to ensure that the backlight is dimmed in the dark parts of the display. A similar approach is used in standard LCD displays of notebook computers and smartphones by global backlight dimming [LT08] for power consumption.

Local backlight dimming is a widely researched topic for SDR LCD displays [BNK+13, MBK+15, NDSLC16a]. Some of the local backlight dimming methods use simple measures to determine the backlight value such as maximum or average of the pixels [FKM00]. However, taking maximum pixel value may lead to LCD leakage [BNK+13] and may not yield any power saving whereas taking average pixel value can yield darker backlight where it should actually be brighter. Some other methods use additional information of the content. Cho and Kwon [CK09] use the average value with a correction term which is computed by finding the difference between the maximum and average brightness locally. Nam [Nam11] computes the backlight for all the image and segment-based parts of the image in order to improve local contrast and keep image appealing by avoiding discontinuities within the image. Lin et al. [LHL+08] use a histogram based approach. They divide the backlight into zones corresponding to LEDs and find weights for each zone by finding the inverse function of the cumulative distribution function (CDF) of the global histogram. Another method proposed by Atkins [Atk12] computes LCD values firstly, contrary to how it has been done in the local backlight dimming literature. Some other studies attempt to solve the dual modulation problem by formulating it as an optimization problem and solving it either globally [BNK+12, BNK+13] or using blocks [BMN+14, CCLS15]. Some of these studies are compared in [BNK+13] and [MBK+15].

Most of these existing works mainly focus on SDR LCD local backlight dimming displays, and they consider a much smaller number of LEDs or edge-lit displays. Due to the very high number of LEDs (~ 2000) and very large point spread function (PSF) of the light diffuser layer in HDR display systems, determination of LED and LCD values becomes an optimization problem. However, even with graphical processing units (GPU) and field-programmable gate arrays (FPGA), solving an optimization problem with this scale is not possible in real-time [THW+07]. Therefore, sub-optimal solutions are necessary.

The first attempt for HDR display rendering was done by Seetzen et al. [SHS+04]. In this method, the backlight is calculated by taking the square root of the target backlight intensity of the image. The LED values are found by carrying out a single iteration of Gauss-Seidel method, and the LCD values are found by dividing the image pixel values to the simulated backlight of the LED values found. This method is explained in more detail in the work of Trentacoste et al. [THW+07].

Another HDR display rendering algorithm was proposed by Narwaria et al. [NDSLC16a].

This method has several steps as following. First, the HDR image is converted from scene-referred values to display-referred pixel values, and, for each pixel, the maximum of each color channel is taken to find the target backlight image. The backlight image is filtered with a max-filter to ensure that it will have the minimum required luminance to show each pixel. Next, using a power law function, the backlight is split into two components, and LED values are normalized to keep the corresponding LCD values smaller than 255. The target backlight is downsampled, and with a gradient based optimization step, LED values are found. The backlight corresponding to the determined LED values is generated using the PSF of the LEDs. Lastly, LCD pixel values are found by dividing the image to the simulated backlight, and the LCD values are gamma corrected.

Albeit very similar to the method of Narwaria et al. [NDSLC16a], we propose another HDR display rendering algorithm in this thesis which differs in a few aspects. The main differences are the power law split and LED normalization for LCD values. In order to ensure the most accurate reproduction of HDR image pixels, the target backlight is not split with a power law function. Alternatively, it can be considered as we are using a power law split using the exponent power of 1 to generate a higher contrast backlight. Additionally, the LED values are not normalized for LCD pixel values. Because the backlight is not split, this normalization step is not needed. The proposed algorithm also considers the power limitations of the display in order to ensure that the displayed image luminance values are as close as possible to the intended luminance values. This consideration for power limitations of the display also allows for a much accurate estimation of the emitted luminance. Details of the proposed algorithm are discussed in Section 2.1.2.

1.4 Quality Assessment for HDR Content

The previous section briefly introduces the new conditions and limitations brought by HDR imaging and content delivery. In this section we discuss the subjective HDR quality assessment studies and the objective HDR quality assessment methods used.

1.4.1 Subjective Quality Assessment

The human visual perception of luminance is not proportional to the physical real world luminance. It follows the DeVries-Rose and Weber-Fechner laws for low luminance and higher luminance values, respectively [KP86]. The wider range of luminance may change the importance of distortion artifacts in the regions darker or brighter than SDR luminance range. Therefore, subjective quality assessment is essential to understand how human perception of quality changes in HDR luminance range. In addition to the assessment of HDR content quality, understanding human perception and preferences of the viewers for these new conditions of HDR are also important.

Perception of HDR Content

The effects of wider dynamic range are analyzed in many perceptual subjective studies using HDR displays. After a subjective experiment presenting distorted images on an SDR and HDR display, Aydın et al. [AMS08] found that the increased brightness increases the visibility of artifacts and distortions become more annoying. They also analyzed the human visual system for HDR luminance range and proposed perceptually uniform (PU) encoding in the same work. Akyüz et al. [AFR+07] answered two very important questions related to the perception of HDR and SDR contents in the presence of an HDR display: "Do viewers prefer HDR or SDR content on an HDR display?" and "What makes the HDR experience superior?". They found that viewers prefer HDR images to SDR images, and there is no significant difference between best exposure SDR image and tone-mapped HDR image. They also found that linearly augmented SDR images (i.e. the dynamic range of SDR image is increased using a linear mapping function) are perceived either better than or as good as HDR images.

Many studies were conducted to analyze the user preference regarding the peak luminance of the display and ambient illumination. Seetzen et al. [SLY+06] conducted a subjective experiment to find out the relationship between the peak luminance and the contrast. One of their findings was that the viewers prefer brighter displays. Yoshida et al. [YMMS06], Rempel et al. [RHLM09], and Hanhart et al. [HKE14b] also made similar observations. Yoshida et al. [YMMS06] conducted a series of subjective experiments in order to find good properties of a TMO and design one based on subjective experiment data. They found that the responses of the viewers depend on the question asked. If they were asked to find the best looking images, viewers enhanced contrast, whereas they avoided changing contrast when they were asked to find the image most similar to the original. In their study, Rempel et al. [RHLM09] found out that the viewers prefer lower display brightness for low ambient light. They also found that HDR displays do not cause visual fatigue and, regardless of ambient illumination, viewers prefer minimizing the black level of the display. In a recent study, a subjective study was conducted in order to analyze the brightness preference for SDR to HDR conversion by Bist et al. [BCMD17a]. The authors found that the viewers' preferred brightness is highly content dependent. In fact, subjects preferred to have lower peak brightness for images with higher rate of bright pixels.

Tone-Mapping

Several subjective studies were conducted for the evaluation and comparison of tone-mapping algorithms. Ledda et al. [LCTS05] conducted a subjective experiment to validate 6 different TMOs against the reference HDR image shown on an HDR display. They also proposed to use a specific pairwise comparisons test to evaluate TMOs, where two SDR and one HDR displays are prepared and users are asked to select the TMO they preferred. Subjective results were analyzed by analysis multiple comparison test, and iCAM [JF03]

and photographic tone mapping [RSSF02] are found to perform very well in different cases. Using a methodology similar to one suggested in [LCTS05], Narwaria et al. [NPDSLCP14] analyzed whether the viewers prefer single-exposure SDR image or tone-mapped image. The experiment results showed that there is no statistical evidence that single-exposure SDR images are better or worse than tone-mapped images, which is in agreement with the findings of Akyüz et al. [AFR+07].

In [NDSLCP14a], Narwaria et al. analyzed five different TMOs by conducting a subjective experiment. In this experiment, 210 compressed HDR images were used, which are compressed with JPEG 2000 using a backward-compatible encoding scheme [WS06]. The results were also statistically analyzed. In another study, Narwaria et al. [NDSLCP14b] conducted a subjective experiment in order to find how visual attention is affected by different TMOs, which can be important for backward-compatible compression methods. In a recent work, Krasula et al. [KNFLC17] conducted a subjective test for viewers' preference on TMOs and evaluated objective metrics using the subjective data collected. They found that the presence of reference affects the viewers' preferences significantly. They also proposed a methodology, similar to [LCTS05], to evaluate tone-mapped images with or without HDR reference. Furthermore, the created subjectively annotated data was made publicly available.

Image Compression

Subjective quality assessment methods are used to evaluate different HDR image compression methods and parameters [NDSLCP13, NDSLCP14a, VDSLD14, HBK+14, KHR+15]. They are also used to collect opinion scores for evaluation of objective quality metrics. Narwaria et al. [NDSLCP13] conducted a subjective experiment with 140 HDR images which are compressed with a tone-mapping based HDR image compression method using JPEG. The database was also used to evaluate four different objective quality metrics. As also mentioned in 'Tone-Mapping' part above, in another study, Narwaria et al. [NDSLCP14a] created an HDR image quality database by subjective annotation of 210 compressed HDR images. The compressed images were created using a backward-compatible compression scheme using JPEG 2000. Valenzise et al. [VDSLD14] conducted a subjective test using 50 compressed HDR images. The HDR images were compressed using JPEG XT and the backward-compatible compression scheme using JPEG and JPEG 2000. Subjective experiment results were then used to compare the performance of objective quality metrics such as HDR-VDP, PSNR, and SSIM. PSNR and SSIM were calculated using either perceptually uniform or logarithmically encoded pixel values. The results showed that the SDR quality metrics can successfully estimate objective quality when they are computed using PU encoded pixel values.

Hanhart et al. [HBK+14] evaluated 13 objective quality metrics by conducting a pairwise comparisons subjective experiment with 20 HDR images compressed using JPEG

XT Profile A. Subjective results were converted to MOS using a Thurstone Case V [Thu27] scaling method before evaluation. The results indicate that commonly used metrics such as PSNR and SSIM perform poorly. In another study, Hanhart et al. [HKE14a] evaluated 11 different TMOs for their use in the JPEG XT Profile A compression. For this purpose, two different crowdsourcing experiments were conducted. One of the experiments measured the quality of the compressed images whereas the selected TMOs were evaluated in the other experiment using a set of attributes, and the most suitable TMOs were found. Using DSIS methodology, Korshunov et al. [KHR+15] subjectively evaluated the quality of 240 HDR images compressed using JPEG XT encoder. For this HDR image quality database, 20 HDR images were selected and tone-mapped using either [RSSF02] or [MMS06]. The images were compressed with JPEG XT Profiles using 4 bitrates and 3 profiles. The subjective scores are used to evaluate the objective quality metrics for JPEG XT standard [AMR+15, HBP+15].

Video Compression

In order to compare several state-of-the-art HDR video compression algorithms, Mukherjee et al. [MDBR+16] conducted a subjective test. In order to keep the experiment short, a ranking based evaluation method was used. The videos were compressed to have either high or low quality level. The subjective experiment results were analyzed for statistical significance, and the video compression methods were ranked accordingly. Non-backward compatible compression algorithms were found to be performing better than the backward compatible algorithms.

Subjective experiments can also be conducted for exploratory intents such as understanding the requirements of a good objective quality metric. In order to investigate the performance of existing objective quality metrics, Řeřábek et al. [ŘHKE15] evaluated several objective quality metrics by conducting a subjective quality assessment using compressed HDR video sequences. Five video sequences were compressed at four bitrates using HEVC with either YCbCr color space, 10-bits, and `4:2:0` chroma subsampling or YCbCr color space, 12-bits, and `4:4:4` chroma subsampling.

Additionally, standardization activities of MPEG for HDR/WCG video compression has driven some other subjective quality assessment studies for video compression. In [HŘE15], Hanhart et al. conducted a subjective test in order to evaluate the responses (nine Category 1 and four Category 3a submissions) to the MPEG CfE for HDR/WCG video coding. Pairwise comparisons methodology with three alternatives was used as subjective test method. The subjective results were then used to evaluate objective quality metrics, and it is found that the quality differences among submissions and the anchor video can be detected using PSNR-DE1000, HDR-VDP-2 and PSNR-Lx metrics. In another study, Hanhart et al. [HRE16] evaluated the responses (five Category 1 and four Category 3a submissions) to the MPEG CfE for HDR/WCG video coding both subjectively and objectively. DSIS was used as the subjective test methodolody. They found that the HEVC video coding

efficiency for HDR video compression can be increased compared to the anchor of MPEG CfE.

Subjective assessment methods are versatile and can be used for many different applications as we discussed in this subsection. For quality assessment, objective quality metrics are much more efficient compared to subjective qualit assessment methodes, with an acceptable accuracy. In the following subsection, we discuss objective HDR quality assessment.

1.4.2 Objective Quality Assessment

As we mentioned previously, SDR multimedia quality assessment methods assume that the image or video is perceptually proportional to the human perception, which is not true for HDR. Still, objective SDR quality metrics can be used for HDR content provided that the pixel values are converted to perceptually uniform scale before computation [VDSLD14]. For this purpose, perceptually uniform (PU) encoding [AMS08], logarithmic function, or perceptual quantizer (PQ) EOTF [MND12, SMP14] are used [HŘE15, HRE16]. This usage of SDR quality metrics for HDR content is also analyzed in Chapter 4.

At the time of writing this thesis, only three objective HDR quality assessment metrics exist: DRIM [AMMS08], HDR-VDP [MDMS05, NMDSLC15], and HDR-VQM [NDSLC15].

Being the first quality metric designed for HDR content, HDR-VDP was developed by extending Visible Differences Predictor (VDP) [Dal93]. In this extension, HDR-VDP simulated the human eye for light scattering and modified the amplitude nonlinearity and contrast sensitivity function (CSF) in order to accommodate the luminance range of HDR. The HDR-VDP metric was further extended by Mantiuk et al. [MKRH11] to include luminance masking, multi-scale decomposition, and quality score estimation. Moreover, the pooling weights for quality prediction step was recalculated by Narwaria et al. [NMDSLC15] using both SDR and HDR images with subjective scores, making the metric more accurate for HDR image quality estimation.

Dynamic range independent image quality assessment metric (DRIM) [AMMS08] detects three types of changes in the structure of the image: loss of visible contrast, amplification of invisible contrast, and reversal of visible contrast. In order to find these changes, DRIM first determines whether the structure (or the contrast) is visible. This is done using the HDR-VDP's contrast detection model which yields a perceptually normalized map. This map is then split into several bands of different orientation and spatial bandwidth. Then a distortion map is generated after pooling of results from several subbands. Although this metric is able to detect the structural changes and is useful for qualitative analysis of the methods, it does not have a pooling mechanism to create a single quality score for the test image. Therefore, its use in quality prediction is rather limited.

HDR-VQM [NDSLC15] quality metric was designed specifically for HDR video. It estimates HDR video quality in a number of steps. First, the emitted luminance values

are found or simulated, and perceived luminance values are found after PU encoding. The frames are then filtered using log-Gabor filters, and subband errors are calculated. The quality score is predicted after the error-pooling step which includes pooling of subband error for each short-term spatio-temporal tubes followed by a spatial and long-term temporal pooling.

These metrics were evaluated in several studies for compressed HDR images [AMR$^+$15, HBP$^+$15] and videos [ŘHKE15, HRE16]. However, these evaluation studies use data sets which are limited either in size or types of distortions. In addition to these studies, results of an extensive evaluation of objective quality methods are presented in Chapter 4.

Chapter 2

Effects of Display Rendering on HDR Image Quality Assessment

Contents

As discussed in detail in Section 1.3.3, HDR display technology has considerably evolved in the last decade. HDR displays are now capable of reproducing much greater and much smaller luminance values, and thus, they have a higher peak luminance and high contrast compared to SDR displays. These conditions may introduce their own distortions compared to the case of SDR where the electro-optical transfer function (EOTF) is standard and the backlight is simple and uniform. Therefore, these augmented viewing conditions are needed to be considered for HDR image and video quality assessment. In our studies, we use a SIM2 HDR47E S 4K display. It is necessary to understand how the SIM2 display works and which parameters affect display rendering in order to analyze the viewing conditions and their effects on the quality assessment.

For objective quality assessment of HDR content, numerous studies try to take into account these new conditions [MDMS05, NMDSLC15, NDSLC15, AMMS08] some of which are evaluated in Chapter 4. Some other works [AMS08, VDSLD14] try to adapt the HDR image pixel values by encoding them as perceptually uniform in order to use existing

SDR objective quality metrics for HDR content. These new conditions are also taken into consideration for the subjective quality assessment of HDR images. Even though some studies address the effect of rendering algorithms [KMBF13, MBK+13], most of the subjective studies focus on specific parameters, mostly the increased brightness and contrast. In their work [HKE+15], Hanhart et al. showed that human observers prefer images displayed at high brightness levels to images visualized at low brightness levels, a result that was previously observed also by Akyuz et al. [AFR+07]. A similar observation was made by Mantel et al. [MKF+15] and Rempel et al. [RHLM09] in their study where they found that the preferred peak luminance increases with the increase in ambient light. This brings forth the thought that the quality experienced by humans viewing images on an HDR display may differ due to rendering differences [AMS08, MBK+13]. In this chapter, we try to find out how the HDR display rendering affects both the objective quality estimation and subjective HDR quality perception.

The accurate estimation of the luminance values emitted from the display is also crucial to understand and analyze objective quality assessment algorithms. HDR image and video quality algorithms HDR-VDP and HDR-VQM [MDMS05, NMDSLC15, NDSLC15] need luminance values for calculation. These luminance values can be acquired in two ways: a measurement of the display luminance can be made, or a simulation can be made to estimate the luminance values of the display. Simulation (or estimation) of the emitted luminance values is also necessary to design new objective quality metrics or to use existing SDR quality metrics for HDR content. In order to simulate luminance values, it is necessary to know the relationship between the pixel values of the input HDR file and the emitted luminance values for the display's rendering.

In this chapter, we assess the effects of the display rendering on both subjective and objective quality assessment. For this study, we used a SIM2 HDR47 display which uses the *dual-modulation* paradigm to generate higher brightness and contrast values. This assessment of the effects on quality was made by comparing two different display rendering methods: the built-in SIM2 rendering method, and a display rendering method which we propose in this chapter. In the following sub-sections, we describe the image and video reproduction of the proposed display rendering method, present the results of the experimental validation, and discuss the effects of using different display rendering methods on subjective and objective HDR image quality.

2.1 Accurate Reproduction of High Dynamic Range Frames

The most popular method for the production of HDR displays is using different layers for brightness and color adjustment. This is done by coupling a locally dimmed light source, such as a panel of LEDs, with a front LCD screen. This process allows both the generation of high peak brightness values and keeping black levels very low with the help of local dimming [SHS+04]. But, both LED and LCD pixel values have to be computed

in order to reproduce an HDR frame within this framework. Given an HDR picture, the problem of estimating the corresponding LED/LCD panel values is known as *dual modulation* [SHS+04, NDSLC16a].

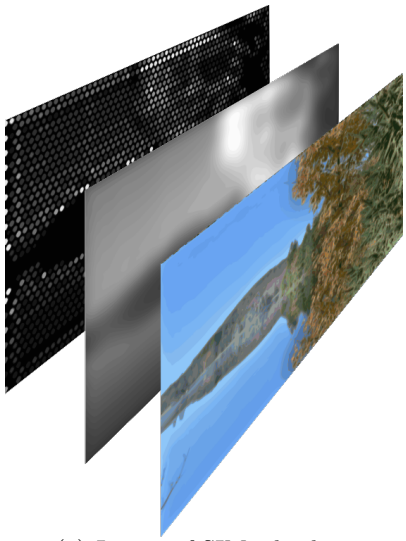
Due to the current technical limitations of HDR displays, dual modulation requires a global optimization approach, since the overall rendering of an HDR picture may change due to perturbations in the value of a single LED and can be influenced by the rendering of the previous video frames [ZVD16]. In practice, many dual modulation algorithms relax this globality constraint and act locally, trading reproduction accuracy for computational complexity. For example, the rendering algorithms built into HDR displays generally give up peak brightness and dynamic range in order to render HDR video in real time, at high frame rates (e.g. built-in display rendering of SIM2 HDR47 display [ZVD16]). Indeed, built-in rendering is often a common choice in many applications, e.g., it has been used in the subjective evaluation of HDR compression performance [KHR+15, LSH+17], tone-mapping studies [NDSLCP13, NDSLCP14a], and color studies [MSL+16, ZHV+17]. Nevertheless, in some psycho-visual experiments, it could be desirable to reproduce the luminance levels stored in the HDR content as accurately as possible, or at least to know the actual per pixel luminance emitted by the display with a sufficient precision. The accurate reproduction and estimation of luminance values, both for HDR images and videos, constitute the goal and the contribution of the work presented in this section.

In this section, we propose a dual modulation algorithm for HDR image and video content, which has the following three characteristics: *i)* it can accurately reproduce the intended HDR luminance; *ii)* it enables us to estimate precisely and with pixel granularity the luminance emitted by an image/video displayed using the proposed method; *iii)* it takes into account temporal dependencies in HDR video, reducing the impact of reproduction artifacts such as flickering. We tested the proposed algorithm on a SIM2 HDR47E S 4K display [SIM14], comparing it with the built-in rendering provided by the manufacturer. We show that our method is systematically more precise in reproducing HDR content and that we can accurately estimate the emitted luminance, which is unfeasible with the built-in rendering. At the same time, our results on HDR video are encouraging, showing that temporal fluctuations can be reduced substantially by smoothing LED values across time.

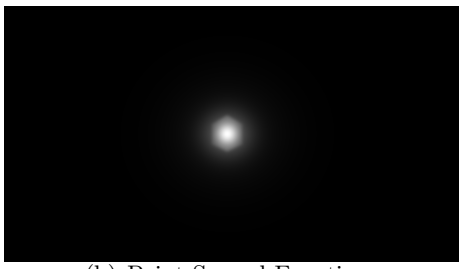
In the following sub-sections, we describe the characteristics of the SIM2 display used, the proposed dual modulation algorithm for rendering HDR images and video, and the experimental validation results.

2.1.1 Display Characteristics

The rendering algorithm we propose is designed to work on SIM2 HDR47E S 4K displays [SIM14]. The peak luminance of the display is measured as 4250 cd/m^2, and its contrast ratio is higher than $4 \cdot 10^6 : 1$. The screen is a dual-modulated display which



(a) Layers of SIM2 display



(b) Point Spread Function

Figure 2.1 – SIM2 HDR display has (a) three layers: LED array constituting the backlight layer, light diffuser layer, and the LCD panel. Light diffuser layer is necessary to avoid discontinuities on the final image, and it introduces (b) a point spread function (PSF).

includes an LED array for backlight, a light diffuser layer, and an LCD panel in this order, as shown in Figure 2.1.(a). There are 2202 independently controllable LED lights and a 1920×1080 pixel LCD panel which can be controlled separately. The SIM2 display can be controlled using the automatic built-in rendering (also known as HDR Mode), or through a custom dual modulation input provided by users (also known as DVI Plus Mode). In the built-in rendering mode, the user supplies the HDR image to a software that converts it to Log Luv color space. The Log Luv image is then processed internally by the display, which determines the values of LEDs and of LCD pixels *transparently* to the user, i.e., one cannot know the values of single LEDs and of LCD pixels obtained in this process. In DVI Plus mode, the user can supply the screen with customized LED and LCD values by indicating them in the first two lines of the image. Hence, it enables the rendering of HDR image and video frames using different rendering algorithms. In this work, the DVI Plus (also denoted as DVI+) mode was used to drive the display.

In order to develop the rendering algorithm, it is necessary to model the characteristics of the display properly. For this purpose, several measurements were made using the Konica Minolta LS-100 light meter. Afterwards, a set of parameters were found to model the display: the maximum power consumption of the display, average power consumption of each LED, the point spread function (PSF) induced by the light diffuser layer. The point spread function can be see in Figure 2.1.(b). Detailed explanations for these measurements and the findings can be found in Annex A.

We would like to note that the proposed approach is still valid with other display models, provided that some parameters of the device are known or previously measured.

Notably, these include the maximum power consumption of the display, as well as an estimation of the PSF induced by the light diffuser layer.

2.1.2 A Dual Modulation Algorithm for Image Reproduction

As discussed in Section 1.3.3, several dual-modulation algorithms for LED/LCD displays are proposed in the literature [FKM00, LT08, CK09, BNK+12, BNK+13, KMBF13, BMN+14, CCLS15]. Most of these works solve the dual modulation problem using some kind of approximation, e.g., they find LED backlight illumination by taking the maximum, average, or weighted average of pixel values [FKM00, BNK+12], or use local block-based approaches [LT08, CK09]. LCD values are generally obtained after the computation of the LED values, by dividing the HDR image luminance by the backlight. Dual modulation can be formulated more rigorously as an optimization problem [BNK+13, KMBF13, CCLS15]. However, these approaches have been mainly targeting low-resolution backlight panels, with tens of LEDs, in the context of low-consumption, locally dimmed LCD displays. A larger LED setup has been considered by Seetzen et al. [SHS+04], who use a local approximation of the gradient with a single Gauss-Seidel iteration to solve for 760 LEDs and 1280×1024 LCD pixels in their prototype HDR display. Current HDR technology, instead, requires dealing with thousands of LEDs, as well as with their large point spread function.

The rendering process on an HDR display with LED-LCD system is essentially a deconvolution problem, i.e., finding the values of the LEDs and of the LCD pixels in such a way to minimize the distance from a target input image. In this process, a critical factor is the asymmetry in the resolution of the LED and LCD panels – the number of pixels in the LCD panel is much greater than the number of LEDs, and the point spread function (PSF) of the LED diffuser has a size of approximately 1000×1000 pixels, which is necessary to avoid discontinuities in the LCD illumination. Additionally, there are other aspects such as the LCD *leakage* [BNK+13] and power constraint. Due to their non-ideal response, LCD cells allow a small percentage of incoming light to pass through them even when they are completely closed (black), and this phenomenon is called *leakage*. Power constraint, on the other hand, requires that the overall brightness should be modulated to account for the maximal power consumption of the display. In practice, this causes some very bright regions of the image to be *clipped*, causing detail loss [BNK+13].

Since SIM2 has 2202 LEDs, any direct optimization approach will be computationally very complex and infeasible to use. So, an iterative scaling algorithm was proposed [ZVDS+15, ZVD16]. The algorithm consists of the following parts:

- Preprocessing
- Computation of target backlight
- Iterative scaling

- Computation of LCD pixel values

The details of these parts are described part by part below.

Preprocessing

First, we find the target *display-referred* luminance values from the input HDR image. HDR images are generally *scene-referred*, i.e., they store values proportional to the physical luminance of the scene. However, the luminance range that can be reproduced by an HDR display is clearly inferior to that of the scene. Therefore, the images should be "graded" to the display capabilities manually or by an automatic process, e.g., by using the display-adaptive tonemapping of Mantiuk et al. [MDK08].

Here, we assume that the input images have been previously graded to the display, and we just saturate luminance values in excess of the maximum display brightness, i.e. 4250 cd/m^2. We denote the preprocessed image as I.

Computation of Target Backlight

Next, we find the target optimal backlight, BL_{target}, that minimizes the required backlight luminance to meet the power constraint and maximizes the fidelity to the target pixel values. In order to find BL_{target}, we define two other backlight images: BL_{min} and BL_{max}.

As liquid crystal cells can only block the light and cannot generate light, at least BL_{min} is required to make sure that the backlight is sufficient for all the pixels of LCD panel. If the backlight is sufficient enough, the intended luminance values can be reached by changing only LCD values. In order to find BL_{min}, we compute the local maxima of the target luminance over 30-pixel radius windows corresponding to the area of a single LED. This ensures that even a very small bright point will have enough backlight. BL_{min} can also be defined as:

$$BL_{min} = \max_{p \in A_p}(I(p)) \tag{2.1}$$

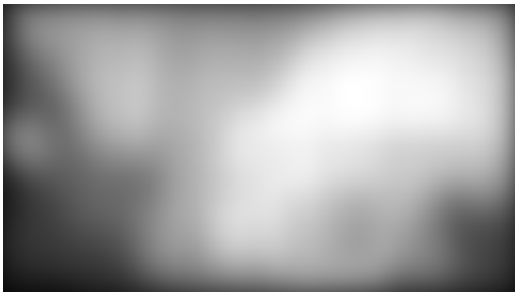
where I is the image and p is the pixel within the 30 pixel radius area A_p.

Liquid crystal cells are known to be non-ideal and leak some light even if they are completely closed. In order to control the effects of LCD leakage, the maximum luminance for each pixel, BL_{max}, is found by dividing the image luminance values of that pixel by the estimated LCD leakage factor $\epsilon = 0.005$. The LCD leakage factor ϵ is found empirically by measuring LCD leakage in different test patterns, using a Minolta LS-100 luminance meter. BL_{max} can also be defined as:

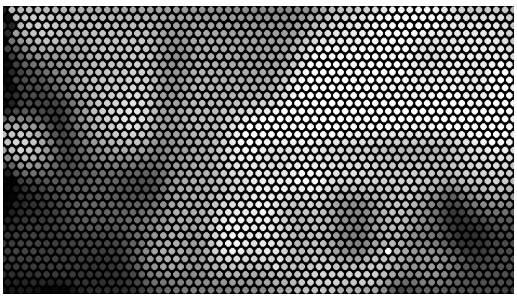
$$BL_{max} = \frac{I}{\epsilon} \tag{2.2}$$

The resulting backlight images are compared pixelwise and the minimum values of BL_{min} and BL_{max} for each pixel are collected within an image $BL_{allowed}$. The $BL_{allowed}$

(a) HDR Image (Tonemapped using [MDK08] for representation)



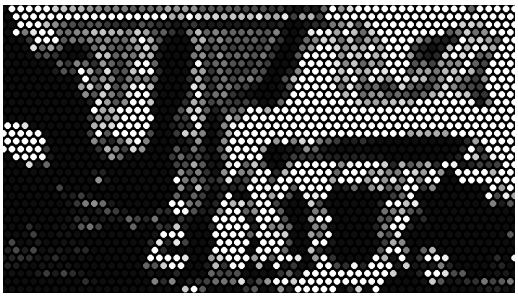
(b) BL_{target} – Target backlight



(c) LED_0 – LED Array initialized by sampling target backlight

(d) S_0 – Scale for the first iteration

(e) BL_2 – Backlight found in second iteration



(f) LED_{final} – LED array found

(g) BL_{final} – Backlight found

(h) LCD pixel values

Figure 2.2 – Steps of the HDR image rendering algorithm for the HDR image *Market3*

values are smoothed and median filtered, and BL_{target} is found. This filtering is done in order to avoid any spurious peaks resulting from single bright spots or noise. This also enables us to meet the energetic constraint of the display. Examples of an HDR image and its target backlight can be seen in Figure 2.2.(a) and (b), respectively.

After the computation of the target backlight, the LEDs and backlight are initialized by sampling BL_{target} on LED locations and taking the convolution with the Point Spread Function (PSF) (found in an earlier study [ZVDS$^+$15]), respectively. That is:

$$BL_t = LED_t * PSF \tag{2.3}$$

where t is the iteration number and $t = 0$ for initialization. LED_0 corresponds to the initial LED values found by sampling BL_{target}, and BL_0 is the backlight of LED_0. An example of LED_0 can be seen in Figure 2.2.(c).

Iterative Scaling

A scale map is generated in order to update the LED values using the following equation:

$$S_t = \frac{BL_{target}}{BL_t} \tag{2.4}$$

where t is the iteration number. The LED values are multiplied with the scale map found as follows:

$$LED_t = LED_{t-1} \times S_{t-1} = LED_{t-1} \times \left(\frac{BL_{target}}{BL_{t-1}}\right) \tag{2.5}$$

LED_t is then clipped to take values in $[0, 1]$, i.e., it is projected onto the set of feasible LED values at each iteration. After the LED values are found, backlight values are also found using the Equation 2.3.

The operations in Equations 2.3 and 2.5 are carried out consecutively by increasing the iteration number until $\sum ||PU(BL_t) - PU(BL_{t-1})||^2$ falls below a threshold. Taking perceptually uniform (PU) [AMS08] encoded backlight makes the computation of the cost function perceptually meaningful and speeds up convergence. When the iterative scaling converges, the resulting LED_{final} values are possibly further scaled linearly to meet the power constraints of the display. Examples of a scale map, backlight of the second iteration, final LED array, and final backlight –BL_{final}– can be seen in Figure 2.2.(d)-(g).

Computation of LCD Pixel Values

LCD pixel values are found by dividing (pixel-wise) each color channel of the original image by the final backlight estimate, and by applying gamma correction, i.e.:

$$LCD_k = \left(\frac{I_k}{BL_{final}}\right)^{1/\gamma_{k,p}} = \left(\frac{I_k}{LED_{final} * PSF}\right)^{1/\gamma_{k,p}} \tag{2.6}$$

where I is the HDR image, $k \in \{R, G, B\}$ is the RGB channel indicator, $p \in \{0, 1, 2, ..., 255\}$ is the LCD pixel value, and $\gamma_{k,p}$ is the gamma correction factor, determined experimentally (see Annex A.2).

As explained in the Annex A.2, the gamma values are different for each channel which is in agreement with the results of Nam [Nam10]. However, the measured gamma values are not constant through the color channel. Therefore, the gamma correction is carried out using a look-up-table. For this purpose, the gamma values were found by measuring each color channel for LCD pixel values $p \in \{0, 1, 2, ..., 255\}$. The resulting γ values were found by dividing the measured luminance of that color channel to the input value of the LCD panel. Hence, the used $\gamma_{k,p}$ is a function of both k and p where k is the RGB channel indicator and $p \in \{0, 1, 2, ..., 255\}$ is the LCD pixel value. An example of LCD pixel values can be seen in Figure 2.2.(h).

Our Matlab implementation of this algorithm takes an average of 19 seconds (about 24 iterations) on an Intel i7-3630QM 2.40 GHz 8 GB RAM PC for rendering a 1920×1080 pixels image. Examples of tonemapped HDR images, LED values, backlights, and LCD values of HDR images "AirBellowsGap", "DevilsBathtub", "MasonLake(1)", "LasVegasStore" contents from Fairchild's HDR dataset [Fai07] are presented in Figure 2.3.

Estimation of Emitted Luminance

Knowing the values of LEDs and LCD pixels, we can estimate the emitted luminance. The HDR image pixels produced by the display are the product of backlight and LCD values. That is, for each color channel k, the rendered image I'_k is:

$$I'_k = (LED_{final} * PSF) \times LCD_k. \tag{2.7}$$

Assuming ITU-R BT.709 primaries [ITU15a], we can compute the emitted luminance as:

$$L = 0.2126 \times I'_R + 0.7152 \times I'_G + 0.0722 \times I'_B, \tag{2.8}$$

2.1.3 A Dual Modulation Algorithm for Video Reproduction

Rendering HDR video requires additional care compared to HDR image, as frame-by-frame rendering might lead to temporal flickering due to high-frequency changes in the backlight. Even though it is possible to reduce the flickering using post-processing [NMBF13], it is preferable to directly handle it during rendering. Burini et al. [BMN+14] considered temporal variation in video sequences and, in order to reduce the flickering effect, implemented an infinite impulse response (IIR) filter integrated into their dual modulation method. They proposed a block-based gradient descent algorithm, and they minimize both the reproduction error and the power consumption required by the LEDs at the same time. However, their work was effective for an LCD display with only 16 LEDs, and it is

(m) HDR Image (Tonemapped) (n) LED values (o) Backlight (p) LCD values

Figure 2.3 – Examples of tonemapped HDR images, LED values, backlights, and LCD values of HDR images (top to bottom) “AirBellowsGap”, “DevilsBathtub”, “MasonLake(1)”, “LasVegasStore”

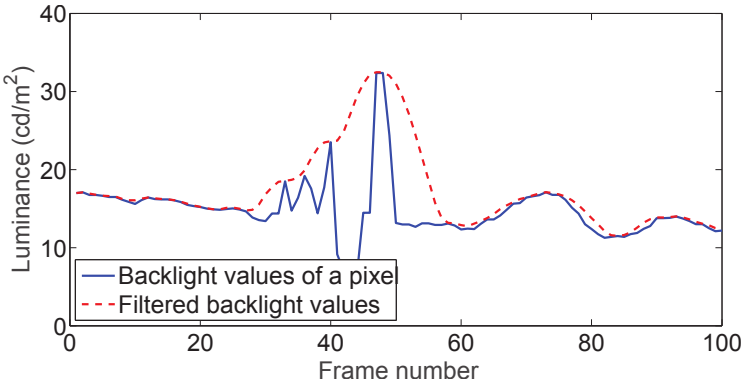


Figure 2.4 – Example of temporal smoothing of backlight pixel trajectories for the "ChristmassTree" video sequence.

computationally too demanding to be extended to configurations with thousands of LEDs.

The HDR dual modulation algorithm described in the previous section provides accurate HDR reproduction. However, small perturbations in the original HDR pixel values may lead to overall changes in the produced backlight. In order to reduce the impact of flickering, we consider two solutions. First, we initialize the LED values for the current frame f using those of the previous frame, i.e., $LED_0^f = LED_{final}^{f-1}$. Second, we smooth the target backlights across time, over consecutive overlapping windows, as described in the following.

Given a video frame f, its initial target backlight BL_{target}^f is computed as explained in Section 2.1.2's *Computation of target backlight* step. Then, for each frame, we consider a look-ahead window of N frames, and we arrange their corresponding backlights in a stack A^f, i.e.:

$$A^f = \left[BL_{target}^f BL_{target}^{f+1} ... BL_{target}^{f+N-1} \right]. \tag{2.9}$$

where the dimensions of A^f are $1080 \times 1920 \times N$. Afterwards, we aim at smoothing the trajectory of backlight pixel values over the window, by computing their upper envelope. To this end, we extract the backlight pixel signal across time for each pixel location (i, j), i.e., we obtain the N-dimensional column vector $A_{i,j}^f$. In order to compute the envelope of this signal, one cannot simply employ a low-pass filter, as averaging may produce lower target backlight than necessary, thus reducing peak brightness and reproduction fidelity. Instead, we adopt a simple approach that consists of convolving each sample independently by a Gaussian window and taking the maximum at each time instant.

More precisely, let S_l be a $N \times N$ matrix such that $S_l(a, b) = 1$ if $a = b = l$ and 0 otherwise. Multiplying $A_{i,j}^f$ by S_l yields:

$$T_{i,j,l}^f = [0 \dots A_{i,j}^f(l) \dots 0]^T, \tag{2.10}$$

i.e., a vector with all zeros but the l^{th} element, which is the l^{th} entry of $A_{i,j}^f$. Now, let $W_{i,j,l}^f = T_{i,j,l}^f * w_\sigma$ be a low-pass version of $T_{i,j,l}^f$ obtained by convolution with a Gaussian

smoothing filter w_σ of variance σ^2. The envelope signal $M^f_{i,j}$ is then obtained by stacking the vectors $\{W^f_{i,j,l}\}$ for $l = 1, \dots N$ into an $N \times N$ matrix

$$B^f_{i,j} = [W^f_{i,j,1} \dots W^f_{i,j,N}] \tag{2.11}$$

where $B^f_{i,j}$ is an $N \times N$ array. Then, $M^f_{i,j}$ can be found by taking the maximum value across the columns of $B^f_{i,j}$.

The procedure described through Equations 2.9 - 2.11 is repeated using a sliding window approach, i.e., the backlight target is updated as $BL^f_{target} = M^f$, and the frame index f is increased by one. An example of filtered target backlight for a given pixel position of the "ChristmassTree" [ABDD+14, BDAPN14] video sequence is shown in Figure 2.4. Once the smoothed target backlight has been computed, the rest of the rendering part follows the algorithm described in Section 2.1.2.

2.1.4 Experimental Validation

It is important to understand how the developed rendering method performs before any further use of the algorithm. In order to analyze the performance of a display rendering algorithm, test patterns and light meters are used generally. These light meters measure the luminance of not directly a point but an area (i.e. a small solid angle). Therefore, they cannot measure the pixel-wise luminance of the display.

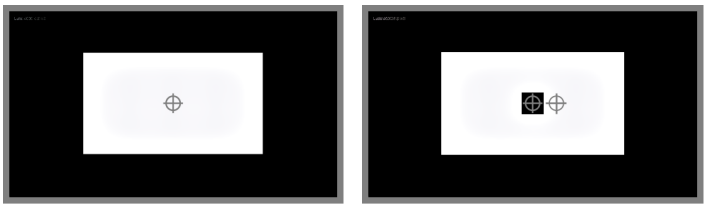
In order to measure the emitted luminance on a pixel-wise granularity, we use a DSLR camera and a light meter to capture an HDR image, register the HDR image pixels and the pixels of the captured HDR image, apply morphological transformation to align the images, and compare the resulting pixel images. The details of this process are given below in the *Measurement of pixel-wise luminance* part.

The proposed display rendering algorithm is able to reproduce HDR images and video frames, and it can also estimate the emitted luminance values. In this part, we report the results of the experimental validation of the proposed HDR display rendering algorithm.

We measured and compared the peak brightness, the local contrast, and the fidelity of reproduction of both the built-in rendering algorithm and the proposed rendering algorithm. Furthermore, we also measured the accuracy of the estimated luminance values and the temporal variation of the backlight for the case of video.

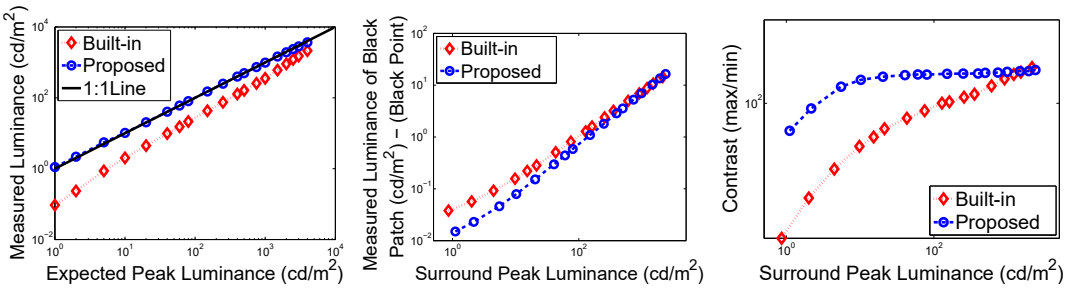
Brightness and Local Contrast

We characterize the performance of the rendering algorithm described in Section 2.1.2 with respect to the built-in mode in terms of accuracy of brightness rendering and local contrast. Since an evaluation of these two measures on complex content (such as natural images) is itself a challenging and content-dependent task, we considered here simple stimuli, which also enable a more accurate measurement of displayed luminance using the Minolta LS-100



(a) Measurement of linear brightness response and peak brightness.

(b) Measurement of black level and local contrast.



(c) Peak luminance

(d) Black point

(e) Contrast

Figure 2.5 – The test patterns for (a) brightness response and (b) local contrast, and the resulting (c) Comparison of peak brightness, (d) black level luminance, and (d) local contrast of the built-in rendering mode and the proposed rendering mode.

luminance meter. Specifically, we considered the following test patterns:

- *Linear brightness response and peak luminance*

We used the pattern of Figure 2.5(a) to measure the accuracy of produced luminance with respect to the target one. The pattern consists of a white box covering 30% of the display surface, surrounded by a black background. The 30% area is selected in order to be within the limits of maximum power consumption of the display. A sequence of test patterns as in Figure 2.5(a) was generated, for values of luminance levels of the white box ranging from 1 to 4000 cd/m^2. Figure 2.5(c) shows the value of luminance in cd/m^2 measured in correspondence of the cross in Figure 2.5(a), as a function of the target input luminance at the same spot. The black solid line indicates the ideal case of a perfectly linear response, i.e., measured luminance matches exactly the required one. This plot shows that: *i)* the proposed rendering algorithm matches more precisely target luminance; *ii)* it also achieves a higher peak brightness than the built-in rendering mode.

- *Local contrast*

Local contrast was tested with the pattern in Figure 2.5(b). This stimulus contains again a white box of 30% of the screen area, but in the middle of the white area, there is a 64×64 pixels square black patch. The small black square width was chosen in order to gauge how LCD leakage affects local contrast in different renderings. We considered several versions

of this pattern with different luminance levels of the white region. Figure 2.5(d) shows the measured luminance of the center black surface versus the measured luminance of the white box. Both measurements were made at the spots shown in Figure 2.5(b). The plot shows how the black level of the center black square is darker for the proposed rendering compared to the built-in rendering, i.e., the proposed rendering is better at handling LCD leakage. The effect of this on local contrast, measured as the ratio between the luminance of the white and black patches, is shown in Figure 2.5(e), which highlights the better local contrast achievable with the proposed rendering.

Measurement of Pixelwise Luminance

In addition to the measurements made with test patterns, another set of measurements were made using a DSLR camera. The use of DSLR camera enables us to measure the pixel-wise luminance and compare the rendering methods using complex natural images.

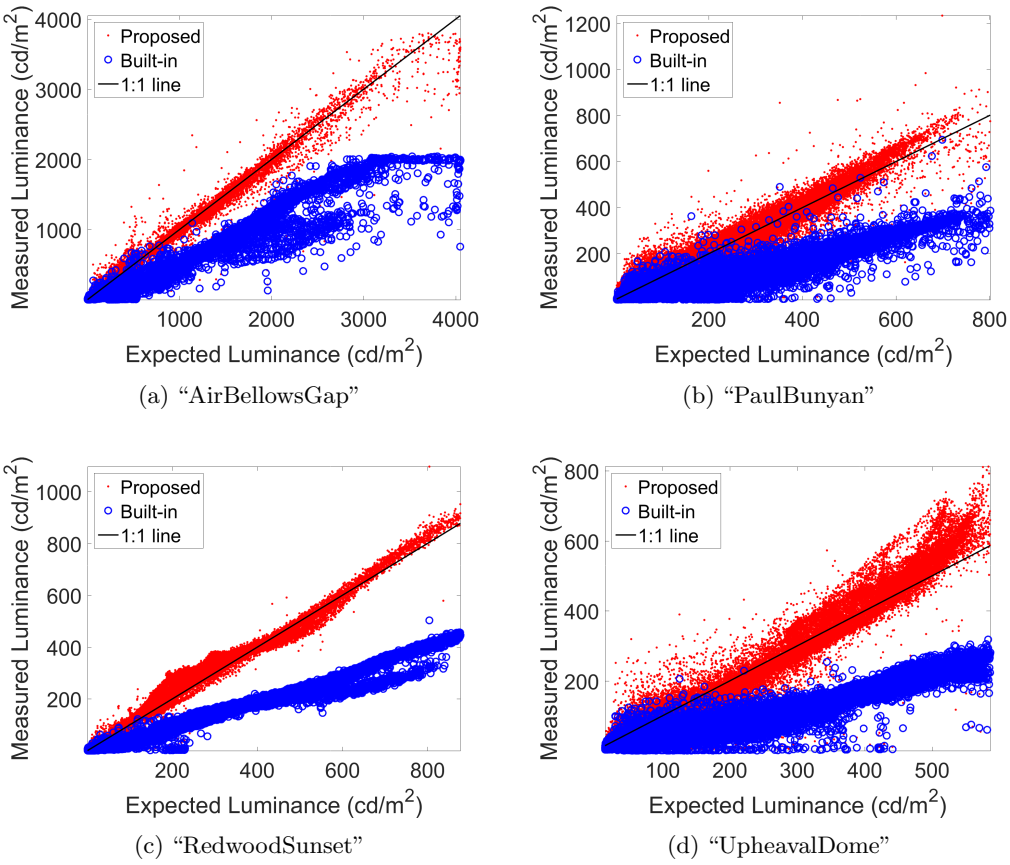
First, 7 raw[1] images with different exposures with exposure compensation values $\in \{-2\frac{1}{3}, -1.5, \frac{-2}{3}, 0, \frac{2}{3}, 1.5, 2\frac{1}{3}\}$ were captured using a Canon EOS700D DSLR camera with 18-55 mm lens. The camera was fixed using a tripod 1.7 meters away from the display to capture the photographs of the display presenting HDR content, and the focal length of the lens was fixed at 35mm. The distance and the lens zoom were selected to avoid or minimize the moiré patterns. The raw images were captured by controlling the DSLR camera using a third party software, and they were built into an HDR image using HDR-Toolbox [BADC11] with the help of the following code:

```
saturation_value = 2^13 - 1;
stackExp = ReadRAWStackInfo(<folder_name>, 'CR2');
stack    = ReadRAWStack(<folder_name>, 'CR2', saturation_value);
[img, ~] = BuildHDR(stack, stackExp, 'linear', [], 'Deb97');
```

Using the captured raw images and their exposure times, `BuildHDR` function generates an HDR image. Captured with different exposure times, the images (which are already `'linear'`) are merged into a single HDR image file by taking a weighted sum of the images using `'Deb97'` [DM97] weight function. These newly created HDR images were registered and aligned to the 1920×1080 resolution. This operation was done for both the built-in rendering method and the proposed algorithm. At the same time, the luminance values on the 4 different spots of the image were measured using a Konica Minolta LS-100. Created HDR images were then adjusted using these luminance measurements.

The resulting HDR images had the pixel-wise luminance values, and these images were used to measure the luminance values of the two rendering methods compared. With the help of this measurement, we were also able to understand the relationship between the

[1]The word 'raw' is used here to denote the uncompressed version of the captured image with linear luminance values. Raw images are not tonemapped and/or gamma corrected.



(a) "AirBellowsGap" (b) "PaulBunyan"

(c) "RedwoodSunset" (d) "UpheavalDome"

Figure 2.6 – Plots of measured luminance vs. expected luminance

HDR image pixel values and the emitted luminance values for the built-in rendering of the SIM2 display.

In the following we refer to these three quantities, which are all expressed in cd/m^2:

- *Expected luminance*, corresponding to the display-referred luminance stored in the HDR file. Values higher than display peak luminance are clipped.
- *Estimated luminance*, i.e., pixel-wise luminance values estimated as in (2.8).
- *Measured luminance*, obtained following the procedure described above. The generated images store the real emitted luminance values.

Fidelity of Reproduction

To gauge the fidelity of HDR reproduction, the measured luminance values were plotted against expected luminance values for 7 different HDR contents from Fairchild database [Fai07] selected in [VDSLD14]; namely, *AirBellowsGap*, *DevilsBathtub*, *HancockKitchenOutside*, *MasonLake(1)*, *PaulBunyan*, *RedwoodSunset*, and *UpheavalDome*. Example plots can be seen in Figure 2.6. Notice that the built-in rendering method provides lower

luminance and higher scatter compared to the proposed rendering algorithm. The fact that the luminance emitted with the proposed algorithm matches the expected one (points clustered on the 45° line) demonstrates the higher accuracy of our rendering. It is also evident, especially from Figure 2.6.(a), that the built-in rendering has a linear response and saturates after a point that is the practical maximum brightness of the display for that case.

In order to have a more quantitative evaluation, Pearson Correlation Coefficient (PCC) and Root Mean Squared Error (RMSE) indices were computed, measuring the linear dependence between two variables and the variance of estimates, respectively. The resulting PCC and RMSE values are reported in Table 2.1. As can be seen, the results of the proposed algorithm are more precise than built-in rendering for each content.

Table 2.1 – Correlation results for luminance measurement for expected luminance and measured luminance.

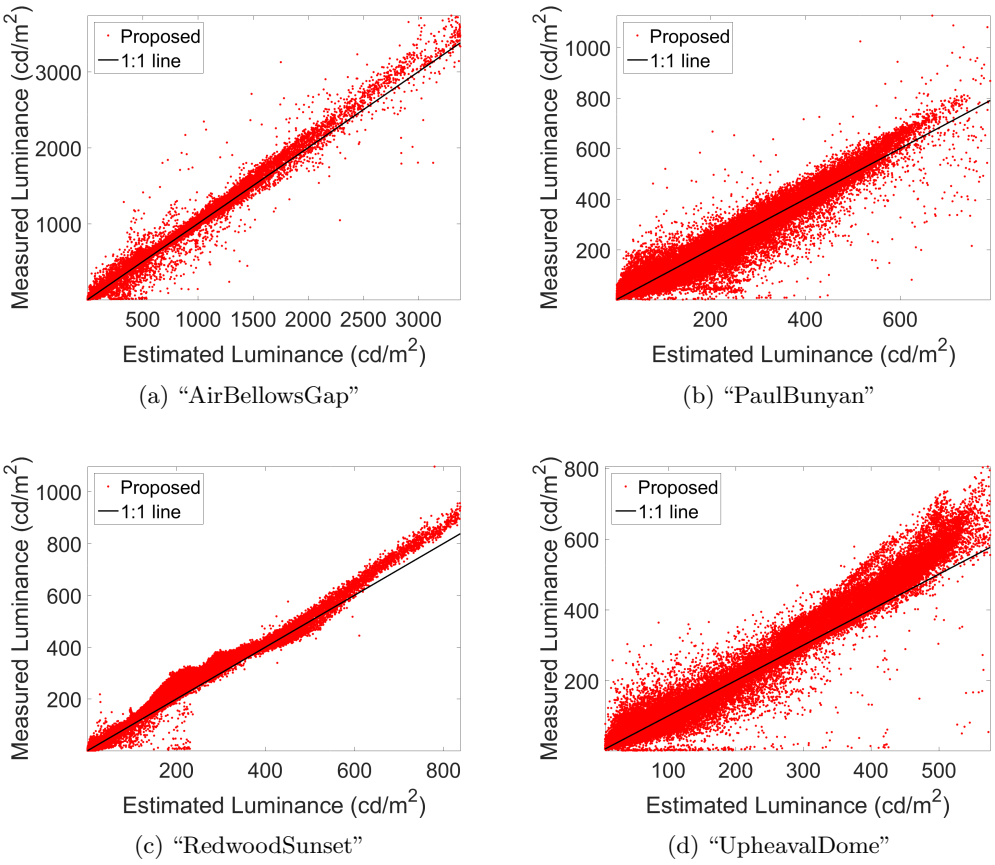
Image	Built-in		Proposed	
	PCC	RMSE	PCC	RMSE
AirBellowsGap	0.9807	311.16	0.9924	65.61
DevilsBathtub	0.8692	96.17	0.9089	59.61
HancockKitchenOutside	0.9400	103.88	0.9572	36.89
MasonLake(1)	0.9159	188.41	0.9312	76.10
PaulBunyan	0.9633	143.73	0.9703	37.83
RedwoodSunset	0.9933	107.69	0.9936	20.83
UpheavalDome	0.9782	142.11	0.9798	37.17

Fidelity of Estimation

In order to assess the emitted luminance estimation accuracy, the measured luminance values were plotted against the estimated luminance values. We report as an example, the scatter plots for four HDR images, namely “AirBellowsGap”, “PaulBunyan”, “RedwoodSunset”, and “UpheavalDome”, in Figure 2.7. The plots show that the estimated luminance values are in a linear relationship with the measured luminance, and the estimated luminance values are very close to 45° line.

Pearson Correlation Coefficient (PCC) and Root Mean Squared Error (RMSE) indices were also computed to have the quantitative results as in the *Fidelity of reproduction* part above. The results are presented in Table 2.2. The PCC scores are above 0.90 and are very close to 1 for the cases of “AirBellowsGap” and “RedwoodSunset”.

Regarding these results, we can say that the accuracy of the luminance estimation of the proposed rendering algorithm has been validated for complex stimuli rather than test patterns.



(a) "AirBellowsGap" (b) "PaulBunyan"

(c) "RedwoodSunset" (d) "UpheavalDome"

Figure 2.7 – Plots of measured luminance vs. estimated luminance. These results are presented only for the proposed rendering.

Temporal Variation

As discussed above, small high-frequency variations in the backlight across time may produce flickering in the displayed HDR video. Although the proposed video rendering method was not validated subjectively, in this part, we try to measure the temporal variation using an objective calculation. In order to measure temporal variation in the backlight produced by our rendering method, we compute the temporal perceptual information index (TI) defined in ITU-T Recommendation P.910 [ITU08]. Since the main source of the temporal variation on video is the backlight, we computed the frame differences over backlight instead of image. TI is computed as:

$$TI = \max_{f} \left(\text{std}(BL^{f}_{final} - BL^{f-1}_{final}) \right), \tag{2.12}$$

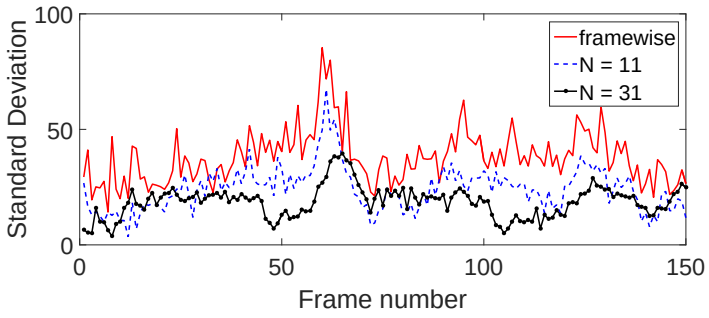
where `std` denotes standard deviation computed over space, and f denotes frame number.

We computed TI for 5 different video sequences, namely *Balloon*, *FireEater2*, *Market3*, *Tibul2* [LLF13], and *ChristmassTree* [ABDD+14, BDAPN14], rendered using both the frame-by-frame algorithm described in Section 2.1.2 and the video algorithm described in

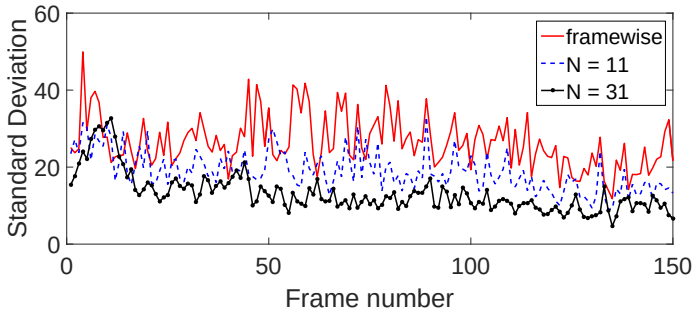
Table 2.2 – Correlation results for luminance measurement for estimated luminance and measured luminance, using the proposed rendering method.

Image	PCC	RMSE
AirBellowsGap	0.9937	62.83
DevilsBathtub	0.9089	59.61
HancockKitchenOutside	0.9575	36.40
MasonLake(1)	0.9312	76.10
PaulBunyan	0.9695	38.29
RedwoodSunset	0.9947	23.86
UpheavalDome	0.9815	40.00

Section 2.1.3. For the latter, we considered different window sizes $N = 1, 11, 21, 31$, where $N = 1$ corresponds to frame-by-frame processing with the only difference that LEDs are initialized using previous LED values. The frame-wise standard deviations for "FireEater2" and "Market3" sequences are illustrated in Figure 2.8. As expected, increasing the window length N reduces the standard deviation of frame difference.



(a) "FireEater2"



(b) "Market3"

Figure 2.8 – Standard deviation change with respect to frame number for (a) "FireEater2" and (b) "Market3" sequences

To compare the results, the TI values of each video sequence are reported in Table 2.3. Again, TI values are dropping with increasing window length. These results show that the proposed HDR video rendering algorithm can effectively reduce the effects caused by temporal variation.

Table 2.3 – TI for different video contents. (*BL*: Balloon, *CT*: ChristmassTree, *FE*: FireEater2, *MK*: Market3, *TB*: Tibul2)

Rendering	BL	CT	FE	MK	TB
framewise	30.68	184.79	85.37	49.92	142.61
$N = 1$	27.32	184.28	63.38	49.18	127.43
$N = 11$	27.00	115.13	67.41	33.05	76.29
$N = 21$	18.44	74.20	48.65	35.72	54.07
$N = 31$	18.92	44.51	39.66	32.63	55.53

2.2 Effects of Display Rendering

The quality metrics developed for the assessment of HDR image and video quality [MDMS05, AMMS08, NMDSLC15, NDSLC15] require as input the per pixel luminance values (expressed in cd/m^2) that an observer in front of the display would see. Moreover, the estimation of pixel-wise luminance values is found to be important for the computation of other objective quality metrics as well [AMS08, VDSLD14]. As a result, different renderings could also have a potential impact on the calculation of objective quality. A different display rendering can also impact the viewers' experience and perception of quality due to the visual changes such as brightness and contrast. In spite of this close connection between quality evaluation and HDR visualization, the effect of different rendering on HDR subjective and objective quality assessment has not been sufficiently investigated so far.

In the previous section, we introduced a simple, yet effective, HDR display rendering algorithm for the SIM2 HDR47 display [SIM14]. We compared the proposed method with the proprietary built-in visualization offered by the display. The proposed rendering algorithm has clear differences from the built-in one, e.g., it yields brighter images, with higher local contrast at low luminance levels.

Equipped with this new rendering, we conducted a subjective study to judge the quality of compressed HDR images, using the same settings as in the previous work of Valenzise et al. [VDSLD14], except that we displayed images with the proposed rendering algorithm. The collected subjective quality scores were compared using multiple comparison analysis in addition to the qualitative analysis of the resulting HDR images.

In order to understand the effect of rendering on the HDR objective quality assessment, we estimated per pixel luminance produced by the display with our rendering algorithm and used this as input to quality metrics for both pristine and compressed content. Since a precise estimation of pixel-wise luminance using SIM2 HDR47 display is not available with the built-in rendering mode, we simulated the luminance values using a simple linear model which scales HDR pixels into the physical bounds of display luminance and clips values that exceed the peak luminance of the device [VDSLD14].

In the following sub-sections, we compare two different rendering methods and assess the impact of HDR image rendering on both subjective and objective scores.

2.2.1 Impact on Subjective Evaluation

In this sub-section, we analyze how a different display rendering can affect subjective quality, for the scenario of HDR image compression. In order to understand the impact of different display renderings on the subjective quality perception, a subjective quality experiment is needed. Due to its design, switching from built-in rendering to DVI+ mode on the SIM2 HDR47 display takes several seconds. Additionally, it cannot be automated and requires a manual intervention of the experimenter. Thus, designing a test presenting the results of both renderings at the same time is not feasible. Therefore, in order to simplify the experiment design, we only used the DVI Plus mode of the display with the display rendering algorithm proposed in Section 2.1.2.

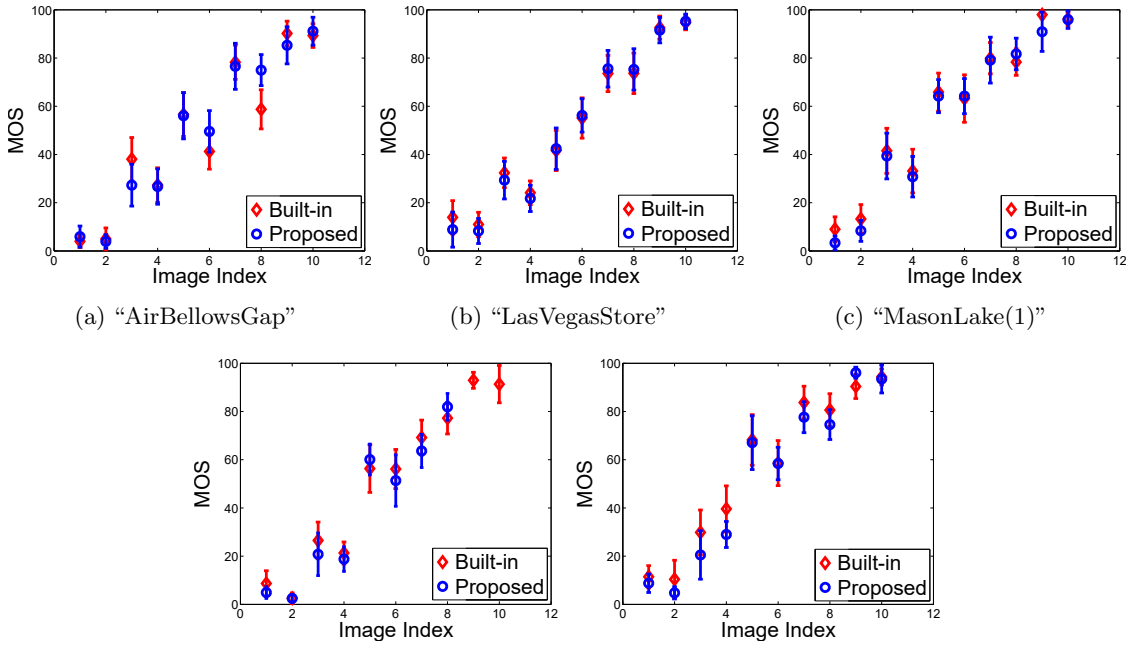
A subjectively annotated HDR image database was created and made publicly available in the work of Valenzise et al. [VDSLD14]. They used the built-in rendering algorithm of the SIM2 display while collecting the MOS values. This subjective test was designed to use the same test material and same experimental conditions as in [VDSLD14] in order to ensure that the only variable that was changed is the rendering mode. In the following parts, we first summarize the test environment and methodology, and we analyze the differences among the results through analysis of variance and multiple comparisons.

Test Environment and Methodology

In [VDSLD14], HDR images were compressed using three different encoders: JPEG, JPEG 2000, and JPEG XT. These compressed HDR images were displayed using the built-in rendering of the SIM2 display, and a subjective experiment was conducted using the Double Stimulus Impairment Scale (DSIS) methodology. In order to rule out all of the possible independent variables except the different rendering, we kept the same test environment and material. The experiment was conducted in a gray surfaced test space which was isolated from all external light sources, conforming to the ITU recommendations BT.500-13 and BT.2022 standards [ITU12b, ITU12a]. The amount of ambient light, not directed to the observer, was 20 cd/m^2. The viewers were seated at about 1 meter distance from the display.

As done in [VDSLD14], we also used DSIS [ITU12b] as the experiment methodology. Two images, reference image A and distorted image B were shown to subjects in a sequential manner. Before the experiment, a training session was conducted to familiarize the subjects with the levels of distortion to be expected during the experiment. The subjects were also told that the image A will always be the reference and image B will be the distorted image. The subjects were asked to rate the distortion appearing in the distorted image B using 5 distinct adjectives (“Very annoying”, “Annoying”, “Slightly annoying”, “Perceptible but not annoying”, “Imperceptible”), on a continuous scale between 0 and 100, 0 being “Very annoying” and 100 “Imperceptible”.

During the pilot test, it was noticed that the magnitude of distortion in the images was



(a) "AirBellowsGap" (b) "LasVegasStore" (c) "MasonLake(1)"

(d) "RedwoodSunset" (e) "UpheavalDome"

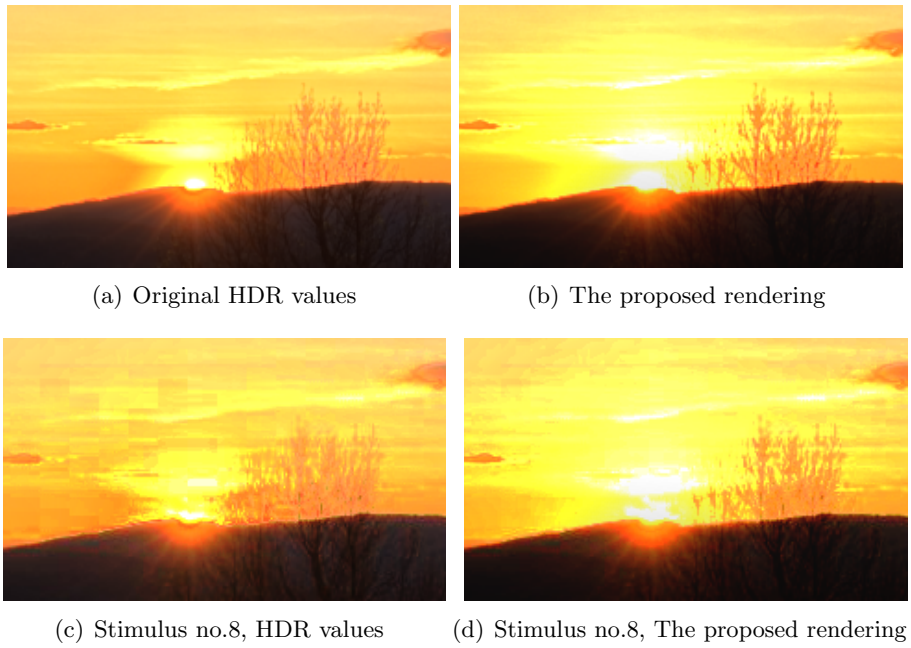
Figure 2.9 – Mean Opinion Scores by different renderings for the tested contents. Points indicate MOS values and bars indicate confidence intervals.

larger than that of [VDSLD14]. Therefore, it was more difficult to judge the quality of the images compared to the case of [VDSLD14]. So, differently from the previous experiment, compressed images were displayed for a duration of 8 seconds instead of the 6 seconds used for the dataset of [VDSLD14]. The dataset consisted of 50 images in total, spanning several contents and coding conditions as explained in the original paper [VDSLD14]. The experiment was paused during the interactive voting, giving the subjects as much time as they wish to complete the task. During the pilot test, it was noted that the average voting time is between 4 and 8 seconds. Hence, one session of the experiment took approximately 20 minutes on average.

Sixteen people (fourteen men and two women) participated in the subjective experiment with the developed rendering. The subjects were aged between 23 and 39, and the average age was 27.75. All the subjects reported normal or corrected-to-normal vision. Two of the subjects were found to be outliers with the standard detection procedure [ITU12b]. The mean opinion score (MOS) and confidence interval (CI) for each of the 50 tested images are calculated after outlier removal, assuming that scores follow a *Student's-t* distribution [ITU12c].

Experiment Results

The resulting MOS values for each content are shown in Figure 2.9. After concluding the tests, we noticed that two samples of "Perceptible" level of the "RedwoodSunset" content



(a) Original HDR values (b) The proposed rendering

(c) Stimulus no.8, HDR values (d) Stimulus no.8, The proposed rendering

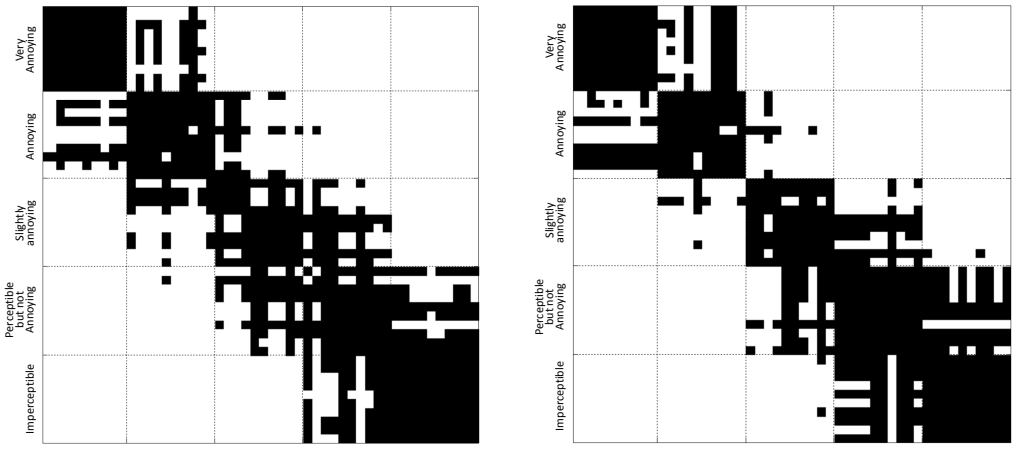
Figure 2.10 – A detail of the "AirBellowsGap" content showing clipping effects on small and very bright regions. Stimulus number 8 corresponds to JPEG compression with a quality factor of 90. The subfigures show (a) the original and (c) the compressed HDR values as stored in the HDR file, as well as (b) the proposed rendering of the original and (d) the proposed rendering of the compressed HDR image. Here the clipping artifacts overcome compression artifacts, i.e., the latter become invisible and thus the MOS of this stimulus is significantly higher with the proposed rendering. Images are tone-mapped for visualization purposes.

were erroneously repeated twice in place of the corresponding "Imperceptible" level. Hence, we excluded them from this comparison. The results of the proposed rendering were compared with MOS values collected using the built-in rendering, which are published with the associated dataset [VDSLD14]. These plots show a substantial level of agreement between the scores obtained with the two renderings, with some differences in some specific contents such as "AirBellowsGap" and "UpheavalDome".

Overall, the collected MOS values using the built-in rendering and the rendering proposed in Section 2.1.2 have a linear correlation of 0.99. A qualitative analysis shows that the distortion in "UpheavalDome" becomes more visible, due to an increased brightness of the rendering, while for "AirBellowsGap" the opposite happens, i.e., details and blocking artifacts become invisible around the sun region, which is clipped in our proposed DVI+ rendering since its brightness is much higher than that of the built-in rendering mode. Examples of the latter phenomenon are illustrated in Figure 2.10.

More details about the differences produced by the two renderings were obtained by performing a one-way analysis of variance, followed by multiple comparison analysis on the MOS values of the built-in rendering and the proposed rendering separately.

Multiple comparison enables us to group stimuli in each dataset according to their



(a) The built-in rendering (b) The proposed rendering

Figure 2.11 – Multiple comparison results for MOS of subjective experiments with different renderings. Each of the 50 rows/columns in each matrix corresponds to a pair of MOS values. For convenience, stimuli are grouped according to their adjectives, as found in the test material selection procedure [VDSLD14].

equivalence in terms of observed mean opinion scores. Changes in the results of multiple comparison may reveal significant differences in the relative perceived quality levels of the stimuli with the two renderings. The results of multiple comparison analysis are reported in Figure 2.11, where the two binary matrices were obtained by comparing all the pairs of MOS values in each dataset and applying the Tukey's honestly significant difference criterion. A black entry in the matrix indicates that no statistical evidence has been found to show that the corresponding pair of MOS values is significantly different. In both Figure 2.11.(a) and (b), we can observe that stimuli are grouped around five clusters, which correspond approximately to the five adjectives of the quality scale which are reported for convenience in the figure.

A qualitative evaluation of Figure 2.11 suggests that the clustering of stimuli MOS values does not change significantly with the two rendering modes. In the highest quality levels, i.e., "Perceptible" and "Imperceptible", though, the results are more intertwined. Considering only these two adjectives (i.e., 190 pairs), there are only 26 pairs of stimuli whose quality appears to be significantly different with the proposed DVI+ rendering algorithm. For the built-in rendering, this number grows to 41. Overall, the proportion of significantly different pairs of stimuli is the same in the built-in rendering and the proposed rendering cases. This suggests that with the proposed rendering subtle details become less visible at higher quality levels, i.e., display artifacts overcome compression artifacts. Conversely, the higher brightness and local contrast offered by the proposed rendering make distortion differences more visible at lower quality levels, with respect to the built-in rendering mode.

2.2.2 Impact on Objective Evaluation

The techniques for measuring HDR image quality can be broadly divided into two classes. On one hand, metrics such as the HDR-VDP and HDR-VQM [MDMS05, NMDSLC15, NDSLC15] accurately model visual perception in such a way to predict and quantify significant visual differences between images. On the other hand, many quality metrics commonly used in the case of SDR imaging directly assume that input values are *perceptually linear* in order to compute meaningful operations on pixels. The perceptual linearization is implicitly done for the case of SDR images by the gamma encoding of sRGB [ITU11].

In the case of HDR signals, a typical mapping function is the perceptually uniform (PU) encoding [AMS08]. Both HDR-VDP and PU-metrics (metrics computed on PU-encoded values) require as input photometric values of the displayed images. Generally speaking, these values can be estimated by the display rendering algorithm. In practice, using the built-in rendering mode of SIM2 HDR47 display, the displayed luminance values are not known and cannot be estimated accurately. Therefore, in the work of Valenzise et al. [VDSLD14], displayed luminance values were simulated (or estimated) assuming a simple linear response of the display, with saturation at the maximum display luminance, i.e., $L_{out} = \min(L_{in}, L_{max})$, where L_{in} is the "display-referred" luminance values to display, and $L_{max} = 4250\,cd/m^2$. We call this simulation the *"linear display model"* or *"linear model"* throughout the thesis. In fact, the results presented in Section 2.1.4 suggest that this linear model is a scaled version of the true response of the built-in rendering of SIM2. An advantage of the DVI+ rendering algorithm described in Section 2.1.2 –and validated in Section 2.1.4– is that it can accurately estimate per pixel displayed luminance, and these estimated pixelwise luminance values can be used as input to HDR quality metrics.

In this sub-section, we compare the performance of several objective metrics when their input is provided by either a simple linear model of the display or by a sophisticated estimate obtained through the knowledge of rendering algorithm. Specifically, we computed the predictions of six quality metrics computed on either L_{out} values (i.e. linear model) or on the estimated luminance values of our proposed DVI+ algorithm, and we correlated them with the MOS values obtained from the subjective experiment discussed in the *Experiment results* part of the Section 2.2.1. Considered full-reference metrics include the peak signal to noise ratio (PSNR), the structural similarity index (SSIM) [WBSS04] and its multi-scale version [WSB03], the information fidelity criterion (IFC) [SBDV05], the visual information fidelity (VIF) [SB06], and the HDR-VDP 2.2 [NMDSLC15]. The source code for the objective quality metrics is taken from `http://sourceforge.net/projects/hdrvdp/files/hdrvdp/` for HDR-VDP-2.2, and from `http://ollie-imac.cs.northwestern.edu/~ollie/GMM/code/metrix_mux/` for other objective metrics. All the metrics except HDR-VDP are computed on PU encoded values [AMS08].

Due to the limited size of the dataset, we evaluated the performance of metric predictions using a non-parametric index such as the Spearman rank-order correlation coefficient

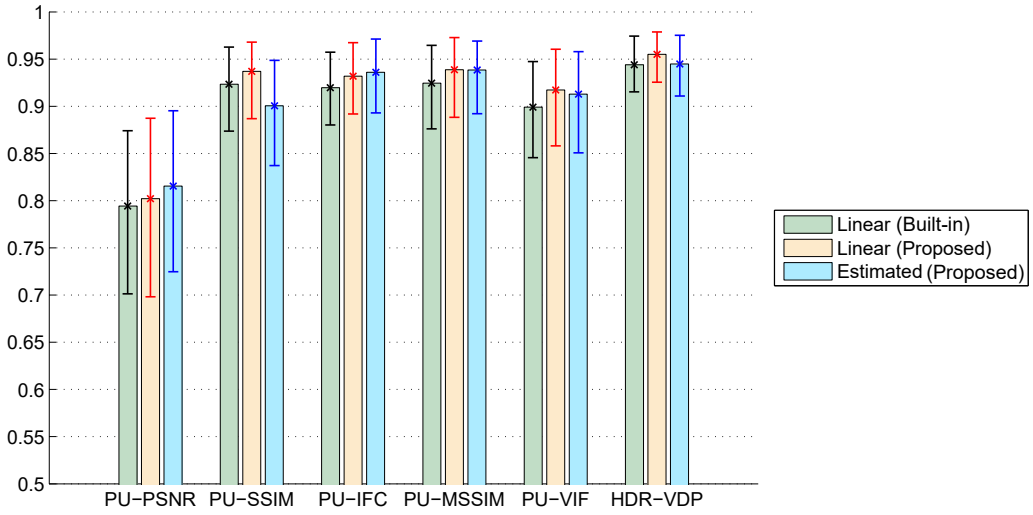


Figure 2.12 – SROCC with 95% confidence intervals for three scenarios: *i)* displayed luminance computed with the linear model and MOS values collected with the built-in rendering mode [VDSLD14]; *ii)* displayed luminance computed with the linear model and MOS values collected with the proposed display rendering algorithm; *iii)* displayed luminance estimated by the proposed display rendering and MOS values collected with the proposed display rendering algorithm.

(SROCC), which measures the degree of monotonicity of MOS estimates. In addition to SROCC values, we also computed confidence intervals of the correlation coefficients using bootstrap (`bootci` Matlab function with bias-corrected accelerated percentile method, 2000 bootstrap repetitions). Figure 2.12 reports the SROCC values with their 95% confidence intervals for the linear model –denoted as Linear (Proposed)– and the estimation of the proposed rendering –denoted as Estimated (Proposed)–. For comparison, we also report the results of the previous experiment of Valenzise et al. [VDSLD14], i.e., the SROCC between metrics computed using the linear model and MOS values obtained with the built-in rendering –denoted as Linear (Built-in)–.

As Figure 2.12 illustrates, the three sets of correlations are very close to each other, and there is no clear gain in using more accurate luminance as input to HDR metrics. To confirm this observation, we tested the significance of the difference of SROCC values for each metric, using the method for comparing dependent [2] correlation coefficients proposed by Zou [Zou07]. This method constructs a confidence interval for the difference of the correlation coefficients. If zero is within this interval, the null hypothesis that the two correlations are equal must be retained. Based on this test, we found the two following results: *i)* the linear model to compute displayed luminance gives statistically indistinguishable performance for two different renderings (the built-in and the proposed rendering, respectively); *ii)* an accurate knowledge of displayed luminance (with the proposed rendering) does not

[2] The dependency of the two correlations is apparent, due to the fact that they are computed on the same dataset of images. In addition, the correlations of Linear (Proposed) and Estimated (Proposed) are overlapping, since they are computed against the same MOS values.

significantly increase the performance of objective metrics with respect to the linear model. In fact, for the case of PU-SSIM, there is a visible decrease of the SROCC coefficients. However, PU-SSIM results are generally very close to one, which makes it difficult to understand the discriminability of this metric in practice for the case of HDR.

This result is quite surprising, as it contradicts the assumptions of many HDR quality metrics, which compute fidelity using the displayed physical luminance as input. A possible explanation for this phenomenon is that, despite the differences between the built-in rendering and the proposed rendering, the reproduced outputs are highly correlated, as the MOS analysis of Section 2.2.1 shows. Furthermore, the saturation in the linear model reduces the effect of outliers in the scene-referred HDR and improves its performance significantly. On the other hand, the proposed DVI+ rendering has a globally linear behavior for the majority of rendered pixels – clipped regions are limited to highlights such as the sun in Figure 2.10, but the saturation in the linear model actually produces a very similar result. This justifies the effectiveness of the linear model for the case of the proposed rendering (i.e. the case denoted as 'Linear (Proposed)'). Finally, there is one important caveat to take into account. The results presented here are valid for a very specific, although popular, processing task, i.e., HDR image compression, and it is known that for simple additive distortion even simple arithmetic metrics such as the PSNR perform quite well [HTG08].

2.3 Discussion

In this chapter, we analyzed the impact of a different display rendering on both subjective and objective quality assessment of compressed HDR images. For this purpose, the characteristics of the SIM2 display were modeled, and a simple iterative HDR image and video frame reproduction algorithm was developed for the widely used SIM2 HDR47 display, which yields higher brightness and contrast than the built-in rendering method. The algorithm employs a sliding window-based filter to avoid flickering for the case of video. It is also validated that the proposed dual modulation method can reproduce HDR content and estimate the emitted luminance accurately. The frame reproduction algorithm presented in this chapter was published in [ZVD16], and it is proved to be of use for psychophysical experimental studies where the knowledge of the emitted luminance is crucial to understand human perception of light and contrast. It is used in a number of studies [HVP$^+$16, HDVD17, KHV$^+$17] in order to find the perceived dynamic range for HDR content and the preference of the viewers on the display gamma.

Using this rendering, we conducted a subjective study to analyze the impact of a different rendering on the subjective quality, and the findings of this comparative study were published in [ZVDS$^+$15]. To understand the effects of display rendering, a subjectively annotated HDR image quality database [VDSLD14] was used where the MOS values were collected using the built-in rendering of the SIM2 display. Test conditions were kept as similar as possible in order to single out the differences in mean opinion scores due

to the only varying factor, i.e., visualization. The MOS values acquired were compared through a multiple comparison analysis. The results show that, overall, MOS values are not dramatically impacted by the employed rendering, although in some cases small and localized compression artifacts might become invisible due to rendering artifacts. At the same time, distortion may become more visible in darker or uniform regions, due to increased brightness.

From the point of view of objective quality metrics, our experiments do not bring enough evidence to support the hypothesis that giving accurate estimates of displayed luminance in input to HDR image quality metrics does bring significant advantages or changes over using a simple linear model of the display response. This simple linear model requires only the peak brightness of the display. Nevertheless, the results of the objective quality analysis show that a simple linear model, which is almost independent from the display, can provide reliable results as if a detailed knowledge of the reproduction display were available.

Another hypothesis for the lack of meaningful difference can be that the simple linear model was accurate enough in the first place. The measured luminance values of the built-in rendering of the SIM2 display show that the actual response of the SIM2 display is, in fact, a scaled version of the linear model used. It can be clearly seen, especially from Figure 2.6.(a), that the built-in rendering has the same characteristics as the linear model. The luminance values saturate after the practical maximum brightness of the display.

This result has important practical implications, since it suggests that HDR quality estimation can be performed with only a rough knowledge of the characteristics of the reproduction device.

Chapter 3

Effects of Color Space on HDR Video Compression and Quality

Contents

The effect of color on the perceived quality is normally discarded in the case of SDR, especially for image and video compression scenarios. However, color may influence the perceptual quality in HDR conditions, as a result of its augmented brightness and contrast levels, due to some aspects of color appearance phenomena, e.g. the Hunt effect, the Bezold-brucke hue shift, etc. [Fai13]. In this chapter, we try to understand the effect of color on the perceived quality. For this purpose, we selected a practical and realistic application scenario, HDR video compression, and we compared the effects of three different color spaces on HDR video compression performance.

For SDR videos, it is common to transform the RGB signal to Y'CbCr color space prior to compression [ITU15a], as it is done in state-of-the-art video compression standards, i.e. H.264/AVC [TLSS09] or H.265/HEVC [SOHW12]. Similarly, in the standardization efforts of MPEG [LFH15] for HDR, this color space transformation is utilized, while the Y channel is coded with Perceptual Quantization (PQ) [MND12, SMP14] instead of the gamma correction function [ITU11]. In addition to Y'CbCr color space transformation, Lu et al. [LPY+16] recently proposed the ITP (ICtCp) color space transformation which

shows better baseband properties than Y'CbCr for HDR/WCG compression with 10-bit quantization. LogLuv [Lar98a] color space transformation is another commonly used transformation among existing HDR video compression algorithms [MKMS04, GT11]. This color space transformation has been slightly modified in this thesis in order to use the same 10-bit encoding scheme and find the effects of it independently of the effects of bit depth. Therefore, we define Ypu'v' which converts pixel values from RGB to Lu'v', and encodes L channel with PQ EOTF [SMP14] in order to get Yp, hence takes the name Ypu'v'. In this chapter, we investigate the effect of these three color spaces on HDR video compression: Y'CbCr [ITU15a], and ITP (ICtCp) [LPY+16] and Ypu'v'.

To this end, we conducted a psychophysical study to compare video sequences coded at different bit rates with the three aforementioned color spaces. We employed a reduced-design pairwise comparison methodology to get the most precise results, comparing stimuli across different bit rates with the goal of converting the obtained preferences to *quality scores*. The choice of compression levels (bit rates) in this case is crucial, and requires selecting test stimuli carefully in such a way to avoid cases where viewers would unanimously prefer one stimulus over the other, or where they would not be able to observe any difference between pairs of video sequences coded at two consecutive bit rates. Therefore, prior to the main experiment, we conducted a preliminary subjective test to select the bit rates for the stimuli, and we selected four bit rate levels for each content. Specifically, in the preliminary experiment, we presented stimuli coded at different bit rates, with the goal to select compression levels spaced apart by one *just noticeable difference* (JND), i.e., such that 50% of participants could observe a quality difference in a pair of stimuli. Using these four bit rate levels, the videos are compressed and used in the main experiment.

The results of the main experiment were analyzed by scaling the preference probabilities for each pair of stimuli into global *just objectionable differences* (JOD) scores, as described in Section 3.2.2. One JOD difference between two stimuli corresponds to selecting one video as higher quality than the other in 75% of the trials. We employ the term JOD instead of JND in this case to emphasize that, in the main experiment, participants were asked to give a *quality* judgment (i.e., select the video which has better overall quality), rather than assess whether a difference between the stimuli exists (as in the preliminary experiment). JOD can then be interpreted similarly to the DMOS concept, and this makes it possible to compare different methods using quality-rate curves. We completed the analysis by testing the statistical significance of JOD differences among different color spaces, and found that, overall, there is no substantial gain of ITP over Y'CbCr, while Ypu'v' has slightly lower performance for some sequences.

3.1 Selection of the Test Stimuli

The pairwise comparisons methodology was selected to compare the effects of different color spaces because the differences between videos compressed with different color spaces

are subtle. For some pairs, viewers can unanimously decide that one stimulus is better than the other, or vice versa. In order to acquire meaningful data without such decisions, the test stimuli should be selected carefully. For this purpose, we conducted a preliminary experiment to select the stimuli for the subjective experiment for color space comparison.

This preliminary experiment was designed to find perceptually uniform distances between compressed HDR video sequences, rendered at different levels of compression. These distances are measured in just noticeable difference (JND) units. For each content, four JND steps, starting from the uncompressed sequence, were found. During the experiment, only the sequences encoded using Y'CbCr color space were examined, and their corresponding bit rates were used as a reference for compression of the sequences in other two color spaces for the main study.

3.1.1 Details of the Subjective Experiment for Stimuli Selection

Experiment Design

This experiment was conducted in four sessions using two alternative forced-choice (2AFC) pairwise comparisons (PC) evaluation, where the question was: *"Can you observe any quality difference between the two displayed videos?"*, and the subjects were able to respond either 'Yes' or 'No'. In the study, the perceptual responses of the participants were evaluated in a randomized design. The sequence and the compression rate were the independent variables. The dependent variable was the user preference.

The dataset contained 7 video sequences, with a significant variance in image statistics, as described in the *Selected Materials* part below. In each session, for each scene, five to seven sequences with different levels of compression (with different quantization parameters (QP)) were generated, so that $QP_{k,i} = QP_k^{ref} + j_i$, where $j = \{1, 2, 4, 6, 8, 10, 14\}$, i is the index, and k is the quality level. Each of these sequences, compressed using $QP_{k,i}$, were compared to the reference sequence with QP_k^{ref}. Then, the QP corresponding to 1 JND difference from the reference, $\widehat{QP}_k$, is selected. In the first session, the uncompressed sequence was the reference and the lowest compression level was selected in the pilot study made with expert viewers. In subsequent sessions, the reference, QP_k^{ref}, was the previously found sequence, $\widehat{QP}_{k-1}$, with one JND from its own reference, QP_{k-1}^{ref}.

In each trial, two videos of the same content but different compression levels were displayed in a side-by-side fashion. Videos were 5 seconds long, and they were repeated once. Upon the video presentation, the voting sign was displayed allowing the participants to make their choice. They were asked if they can perceive any difference in quality with respect to compression artifact, previously demonstrated during the training session. The voting time was not restricted. The next set of stimuli was presented one second after the user voted. The test design is visualized in Figure 3.1.

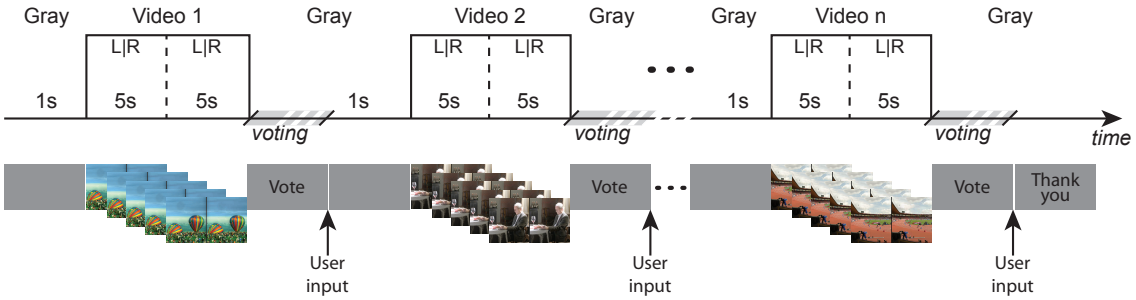


Figure 3.1 – Visualization for the subjective experiment for stimuli selection

Selected Materials

7 HDR video sequences were used in the study: the "BalloonFestival" (Balloon), "Market3Clip4000r2" (Market) and "Tibul2Clip4000r1" (Tibul) sequences proposed in MPEG by Technicolor and CableLabs [TF15], the "Bistro-01-pad16" (Bistro) and "Showgirl-02-pad16" (Showgirl) sequences from the Stuttgart HDR Video Database [FGE+14], "EBU-04-Hurdles" (Hurdles) and "EBU-06-Starting" (Starting) sequences from EBU Zurich Athletics 2014 (https://tech.ebu.ch/testsequences/zurich). Names in parentheses will be used to refer to the scenes for the rest of the chapter. See Figure 3.2 for the screenshots of each video sequence. The frame rates and the horizontal crop locations are reported in Table 3.1.

Table 3.1 – The corresponding frame rates and horizontal crop windows (in pixels) of the test sequences used in the preliminary experiment

Sequence	fps	H-crop
Balloon	24	921–1872
Bistro	30	855–1806
Hurdles	50	1–952
Market	50	471–1422
Showgirl	30	429–1380
Starting	50	541–1492
Tibul	30	481–1432

The sequences were selected based on the image statistics and the pilot study, so that the dynamic range (DR), image key (IK), spatial (SI) and temporal (TI) perceptual information measures and image content vary and are evenly distributed across the data set, see Figure 3.3. These features are briefly summarized below:

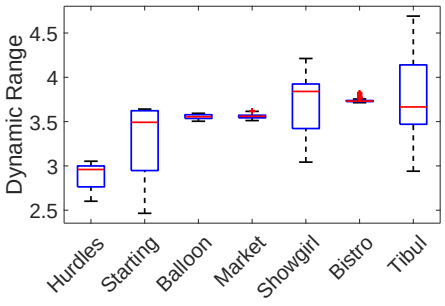
- **Dynamic Range:** Simply $DR = \log_{10}(Lum_{max}/Lum_{min})$, where Lum_{max} and Lum_{min} is the maximum and minimum luminance values of the HDR image.
- **Image Key:** Indicates the brightness of the image, $Key = \frac{\log Lum_{avg} - \log Lum_{min}}{\log Lum_{max} - \log Lum_{min}}$ where the $\log Lum_{avg}$ is calculated as $\log Lum_{avg} = mean(\log Lum)$, and Lum is the luminance of the HDR image.

(a) Balloon (b) Bistro (c) Hurdles

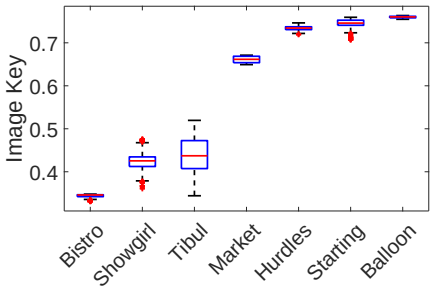
(d) Market (e) Showgirl (f) Starting

(g) Tibul

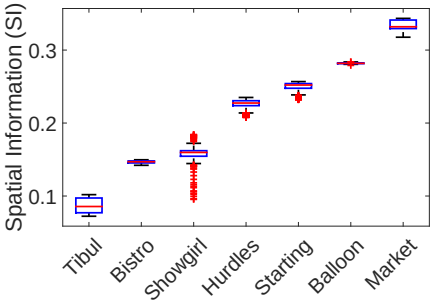
Figure 3.2 – Starting frames of the 7 HDR video sequences used in the preliminary experiment. The Balloon, Market and Tibul sequences were proposed in MPEG by Technicolor and CableLabs [TF15]; the Bistro and Showgirl sequences are from the Stuttgart HDR Video Database [FGE+14]; and Hurdle and Starting sequences are from EBU Zurich Athletics 2014 (`https://tech.ebu.ch/testsequences/zurich`). The images were tonemapped [MDK08] for representation. Showgirl and Tibul scenes were not used in the main study.



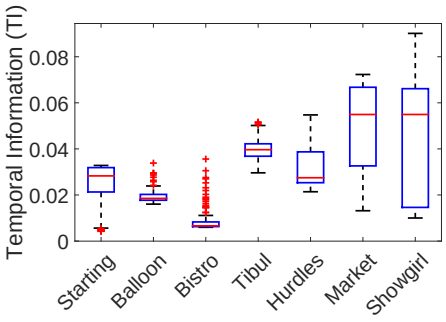
(a) Dynamic Range



(b) Image Key



(c) Spatial Perceptual Information (SI)



(d) Temporal Perceptual Information (TI)

Figure 3.3 – Image statistics for selected scenes. The values were sorted for better representation.

- **Spatial Perceptual Information (SI):** As proposed in ITU-T Rec. P.910 [ITU08], SI is calculated as

$$SI = \max_n \left(\text{std}_{space} \left[\text{Sobel}(Lum_n) \right] \right)$$

where n is the frame number. It was designed and recommended for SDR images. In order to adapt it for HDR, we modified the physical luminance values (in cd/m^2) by PQ encoding [SMP14] and mapping them to the range of $[0, 1]$ before calculation.

- **Temporal Perceptual Information (TI):** TI is calculated as

$$TI = \max_n \left(\text{std}_{space} \left[Lum_n - Lum_{n-1} \right] \right)$$

where n is the frame number. It was designed and recommended for SDR videos [ITU08]. As in the case of SI, the physical luminance values (in cd/m^2) were PQ encoded and mapped to the range of $[0, 1]$ before calculation.

Since the SIM2 HDR47 display is only capable of displaying the color gamut described in ITU-R Recommendation BT.709, all of these sequences were processed in BT.709 color gamut. Several of these sequences (including *Balloon*, *Hurdles*, *Market*, *Starting*, and *Tibul*) were directly acquired from MPEG files, and their pixel values were already in the BT.709 color gamut. Two other sequences, *Bistro* and *Showgirl*, were acquired from Stuttgart

HDR Video Database [FGE+14] in ALEXA-Wide-Gamut OpenEXR format. Their pixel values were clipped to fall within BT.709 color gamut. Normally, such a clipping operation is expected to create strong artifacts such as banding or saturation. However, no visible artifacts were observed both on the undistorted reference and on the compressed videos since most of these clipped pixels are in the dark regions and mostly camera acquisition noise.

Test sequences were generated using the following chain of operations: First, the RGB HDR frames were encoded using PQ EOTF and then transformed to Y'CbCr color space. After `4:2:0` chroma subsampling, Y'CbCr frames were encoded using HEVC Main-10 profile with HM 16.5 [SOHW12, BFSS17]. The encoded bit streams were then decoded and both the color transformation and EOTF encoding were inverted. The resulting frames were stored in an AVI file as uncompressed video frames. After JND was found for that level, the set of videos for the next level with different QPs were generated, as described in the *Experiment Design* part above.

The experiments were conducted in a dark, quiet room, with the luminance of the screen when turned off at 0.03 cd/m^2. The stimuli were presented on a calibrated HDR SIM2 HDR47E S 4K 47" display with 1920×1080 pixel resolution, peak brightness of 4250 cd/m^2, used in its native built-in rendering mode. The distance from the screen was fixed to three heights of the display (i.e. approximately 180 cms), with the observers' eyes positioned zero degrees horizontally and vertically from the center of the display [ITU98]. The framework was developed in MATLAB R2014b and run on a Dell T5500 computer with Intel Xeon X5680 processor at 3.33GHz, 24GB RAM and NVIDIA Quadro FX 580 graphics card. Due to the immense content size, we stored all the test materials on an SSD hard drive for faster content loading and seamless display.

Participants and Procedure

33 people (20 men and 13 women) with an average age of 33.6, volunteered for the experiment. In each of the four sessions there were 13, 17, 11 and 12 participants respectively, among whom most took part in two nonconsecutive sessions. All of them reported normal or corrected-to-normal visual acuity.

Prior to the experiment, the participants were briefed about the purpose of the experiment. This was followed by a verbal explanation of the experimental procedure and a short training session with 8 sample trials. At this time, two sequences that were not used in the study, rendered at several levels of compression, were utilized and the nature of the artifacts was explained. Towards the end of the training, the participants were asked to evaluate a few pairs of stimuli. Doing so, the experimenter was able to understand whether the participants understood the task. This further helped to stabilize their opinion, to adjust to the magnitude of the quality degradation, and to further familiarize themselves with the experimental framework. Following the training, the experiment commenced and no

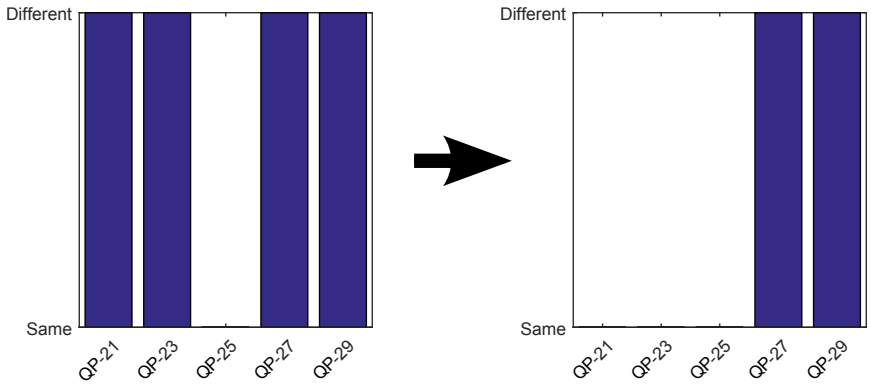


Figure 3.4 – Removing the inconsistency in user voting. In this case, QP=27 was selected as the threshold.

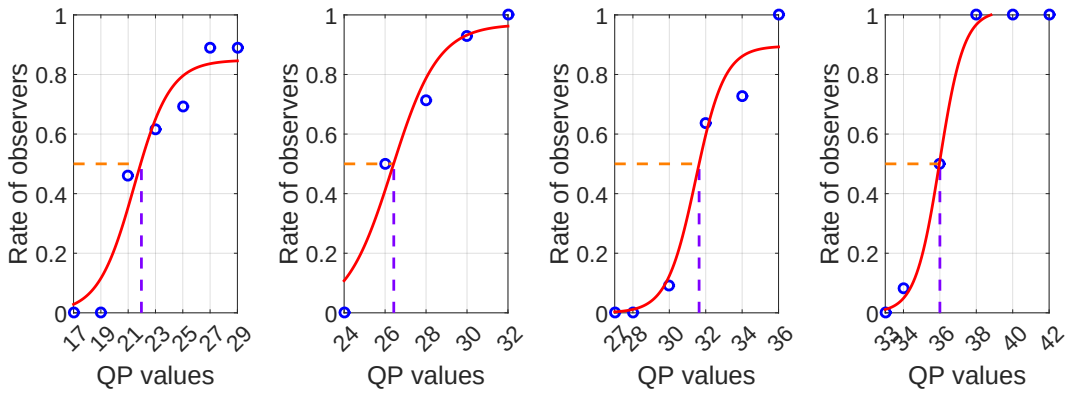
further interaction between the participant and the experimenter occurred until debriefing once all trials were conducted.

3.1.2 Stimuli Selection for the Color Space Experiment

In order to find the stimuli 1 JND apart from each other, the experiment was conducted in an iterative fashion as described in the *Experiment Design* part above. To find the stimulus which was 1 JND apart from the anchor video, the following operations were carried out in this order for each level k. For the sake of simplicity and continuity throughout the chapter, we call these levels JOD_k where $k \in \{1, 2, 3, 4\}$ is the quality level.

After each session, the resulting data was gathered and screened for consistency. The results of each video sequence for each participant were grouped together and analyzed. Some of the results were interesting as all the responses of a participant for a particular scene were "Same" (or "Different"). This means either that there was no perceivable difference between any of the pairs (for the case of all are "Same") or that difference was perceived in all compared pairs (for the case of all "Different"). The former is highly unlikely to happen as the distortion is eminently noticeable between the sequences generated with the $\Delta QP \approx 10$. The latter is even less likely since even the expert viewers could not perceive any quality difference at $\Delta QP = 1$ for any of the tested scenes. Therefore, the cases where all the responses were "Same" or "Different" were considered as outliers, and the results of that particular participant for that particular scene were discarded. During the whole analysis, 37 out of 273 comparisons (per participant and per scene) were removed in total.

The results were further analyzed for their consistency. The gathered results for each participant and each video sequence were expected to follow a simple pattern: users would not see any difference in the videos with $QP_{k,l,m,i} < \theta_{k,l,m}$ until a certain threshold point and would see the difference in all of the stimuli after that point, i.e. $QP_{k,l,m,i} \geq \theta_{k,l,m}$, where $\theta_{k,l,m}$ is the QP threshold point for k^{th} quality level for l^{th} observer and m^{th} scene. The results which did not follow this expected behavior were considered inconsistent, and



(a) $\widehat{QP}_1 = 22$ (b) $\widehat{QP}_2 = 26$ (c) $\widehat{QP}_3 = 32$ (d) $\widehat{QP}_4 = 36$

Figure 3.5 – The process of finding 1 JND distance videos for Balloon video sequence. The corresponding QP values are found where only 50% of the participants could see the difference between the videos. The values zero and one on the y-axis indicates that the difference can be observed by none or all of the observers, respectively.

they were modified to be consistent. That is, the $\theta_{k,l,m}$ was found as the minimum QP value which ensured that all of the following data points were all different. Assume that $R_{k,l,m,i}$ is the pairwise comparisons results obtained for $QP_{k,l,m,i}$:

$$\theta_{k,l,m} = min\{QP_{k,l,m,i} \mid R_{k,l,m,(i+n)} = \text{“Different”}, \quad \forall n \geq 0\} \tag{3.1}$$

Afterwards, all of the data points $R_{k,l,m,i}$ with $QP_{k,l,m,i} < \theta_{k,l,m}$ was set to “Same”. Please notice that $\theta_{k,l,m}$ is different for each observer l and scene m. This operation is visualized in Figure 3.4, where $\theta_{k,l,m} = 27$.

The modified results $\widehat{R}_{k,l,m,i}$ were summed across the participants, and the result was plotted. The result of this sum resembles –as expected– a cumulative distribution function (CDF) of the probability of seeing a difference. For the videos with lower QP (and higher bitrate), there is little or no difference to the video of the previous level, $\widehat{QP}_{k-1}$. After one point, the CDF becomes 1. This means that every video with that QP was and every other video with higher QP will be noticed by all of the observers. The underlying CDF was estimated by using a logistic fitting. The video QP which yields 1 JND with the anchor video, i.e. $\widehat{QP}_k$, was determined by finding the closest QP value corresponding to the 50% of observers seeing difference. Examples of this operation are shown in Figure 3.5, and the process of determining $\widehat{QP}_k$ is indicated with dashed lines.

The resulting QP and bitrate values are reported in Table 3.2 on the rows indicated as Y’CbCr. As can be seen from the table, there is not any simple relationship between the QP values of the quality levels and their bitrates. Both the QP value and the bitrate of a video seem to be related to the characterstics of the video. Bistro video sequence appears to have some special characteristics considering its very low bitrate. These very low bitrate

values may be due to the low key, SI, and TI values of the video.

The QP values for videos compressed with other color spaces, namely ITP and Ypu'v', were found by finding the QP values minimizing the bitrate difference between the Y'CbCr video. For this purpose, the raw videos were compressed using a set of QP values $QP_{k,i} = \widehat{QP}_k^{\text{Y'CbCr}} + j_i$ where $j \in \{-2, -1, 0, 1, 2\}$ for k^{th} quality level and i is the index. Afterwards, the bitrates of these videos were compared, and the QP value of the video yielding minimal bitrate difference was chosen. The resulting QP values and bitrates are reported in Table 3.2.

Table 3.2 – Compression levels: All of the QP values and the corresponding bit rates (in kbps) across scenes and JOD levels.

Sequence	Color Space	JOD_1		JOD_2		JOD_3		JOD_4	
		$\widehat{QP}_1$	BR_1	$\widehat{QP}_2$	BR_2	$\widehat{QP}_3$	BR_3	$\widehat{QP}_4$	BR_4
Balloon	Y'CbCr	22	4945.69	26	2653.65	32	1151.99	36	678.08
	ITP	22	5128.45	26	2742.05	32	1185.61	36	705.98
	Ypu'v'	23	4531.33	27	2466.26	33	1072.66	37	650.01
Bistro	Y'CbCr	23	520.48	27	278.21	32	143.54	34	111.04
	ITP	23	525.09	27	287.10	32	148.58	34	114.80
	Ypu'v'	25	569.49	28	278.77	33	141.20	35	111.42
Hurdles	Y'CbCr	22	7077.85	25	4147.72	28	2557.50	31	1674.57
	ITP	22	6610.88	25	3923.48	28	2465.01	31	1644.14
	Ypu'v'	22	7526.22	25	4364.39	29	2390.68	32	1567.09
Market	Y'CbCr	25	6252.36	29	3108.43	34	1339.03	36	969.68
	ITP	25	5797.64	29	2934.46	34	1283.06	36	936.23
	Ypu'v'	25	6090.28	29	3089.60	34	1348.33	36	988.16
Starting	Y'CbCr	19	7210.73	25	2170.69	28	1337.33	30	1010.40
	ITP	20	6879.85	25	2250.69	28	1373.15	30	1052.30
	Ypu'v'	22	7575.62	27	2260.09	30	1342.80	32	927.63

3.2 Color Space Effect on Compression

In this main experiment, the compression performance when coding HDR video sequences using different color spaces was investigated. The bit rates were selected based on the results of the preliminary experiment as explained in the Section 3.1.2 above.

3.2.1 Details of the Subjective Experiment

Experiment Design

For the main experimental task, we chose paired comparisons methodology which provides higher sensitivity and easier experimental task than direct rating. However, this method may require comparing an excessive number of pairs when a large number of conditions is involved [MTM12], as in our case. For the complete design in our experiment, it would be necessary to make 390 unique comparisons, which would require multiple long sessions.

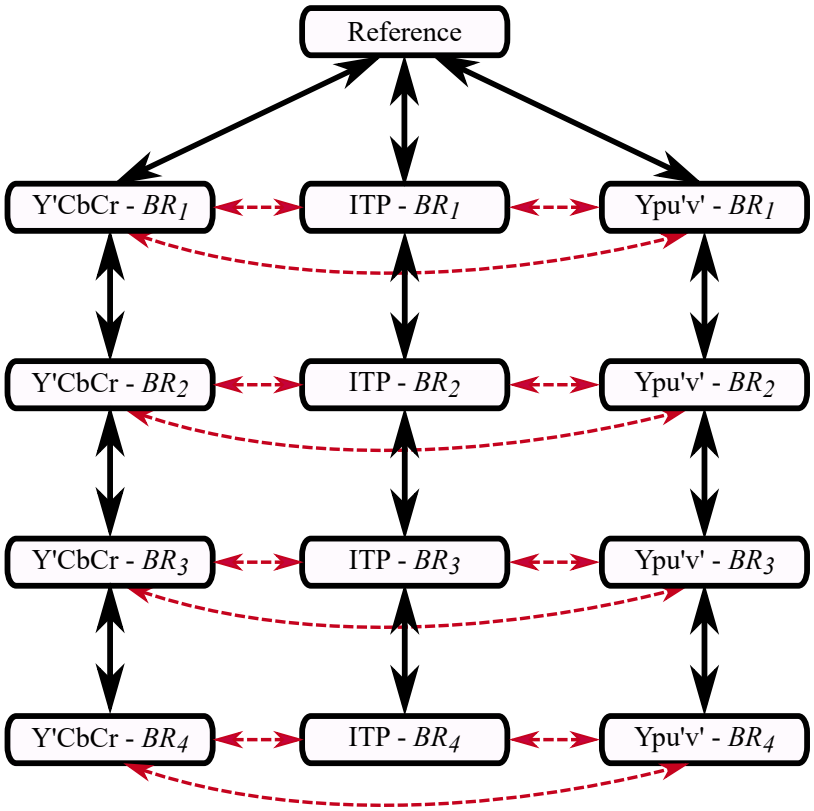


Figure 3.6 – Incomplete pairwise comparisons experiment design used for the main color space subjective experiment. Solid black lines indicate the comparisons made within the same color space, and dashed red lines indicate the comparison accross color spaces for the same bitrate.

At the same time, comparing stimuli with significantly high perceptual difference leads to obvious and unnecessary results. Therefore, an incomplete design in which only the relevant pairs were compared was employed.

HDR video sequences were compared across bit rates for the same color space, and across color spaces using the same bit rate, as shown in Figure 3.6. For the former, only the sequences compressed at the neighboring bit rates were compared, e.g. BR_1 vs BR_2, or BR_3 vs BR_4. The uncompressed sequence was compared only with the three videos compressed at the highest bit rates. In each trial, the participants had to select the sequence with higher quality, i.e. with lower magnitude and amount of perceivable artifacts.

Selected Materials

Due to the high inconsistencies of the preliminary experiment results for Tibul and Showgirl scenes, these sequences were discarded in this experiment. Showgirl scene has the face of a showgirl which is the only salient part, and because of that, other parts of the video become not important for the users. Moreover, the face starts on the mirror and changes

location during the sequence. This rapid change in the speed of the face makes it harder for viewers to understand the quality. Additionally, in this scene, the salient region is close to the display boundary. As mentioned in the Annex A.3, the boundary effect present in the built-in rendering heavily affects the user votes due to the brighter boundary regions. Tibul scene also suffers from similar problems. The salient region change quickly due to the movement of the ship. It also has the same boundary problem as Showgirl, and Tibul looks unnatural due to its lighting and color scheme.

The test sequence generation was done similar to the description made in *Selected Materials* part in Section 3.1.1. RGB videos were either transformed to Y'CbCr after PQ EOTF encoding, to ITP, or to Ypu'v'. After `4:2:0` chroma subsampling, converted frames were encoded using HEVC Main-10 profile with HM 16.5 [SOHW12, BFSS17]. The encoded bit streams were then decoded, and color transformation and EOTF encoding were inverted. The resulting frames were stored in an AVI file as uncompressed video frames. As described in the previous section, Section 3.1.1, JND levels were found using only Y'CbCr color space transformation. The QP values for ITP and Ypu'v' were found by finding similar bit rates to selected Y'CbCr videos corresponding to different JND levels.

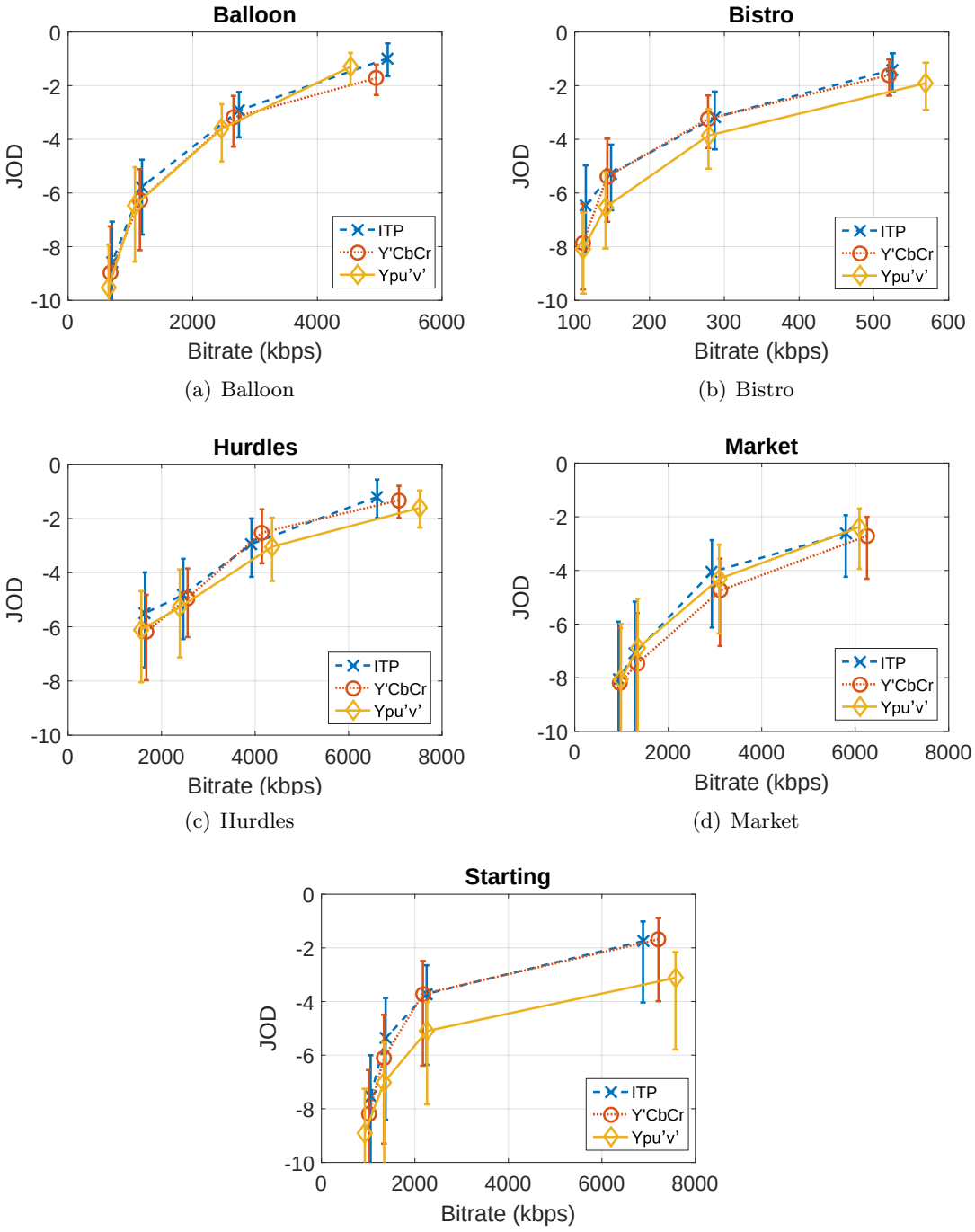
Participants and Procedure

18 people (14 men and 4 women), with an average age of 29.44, volunteered for the main experiment. All of them reported normal or corrected-to-normal visual acuity and were tested for color acuity using Ishihara test. This time, the participants were asked to select the sequence with the higher quality and thus fewer compression artifacts, or otherwise make the best guess. 14 participants took part in two sessions, composed of the same pairs but displayed in different order, i.e. A vs B, and B vs A. The total number of user responses per pair was 32.

3.2.2 Analysis of the Subjective Results

The subjective results were collected using the subjective test the details which were explained in the previous sub-section. These results were analyzed after a scaling was done.

The results of a pairwise comparisons test are generally gathered in a preference or comparison matrix. This matrix includes the preference ratios of the stimuli, and these preference ratios can be converted into quality scores by a procedure called *"scaling"*. There are several methods to carry out scaling [BT52, Thu27, LDSE11, TG11], and these methods use two models: Bradley-Terry model [BT52], and Thurstone's model [Thu27]. Scaling pairwise comparisons data is discussed in more detail in Section 5.1.



(a) Balloon

(b) Bistro

(c) Hurdles

(d) Market

(e) Starting

Figure 3.7 – Image scores obtained by scaling preferences to relative quality distances (in JOD units) for the three tested color spaces.

Scaling of Pairwise Comparisons Results

The obtained pairwise comparison results were scaled using publicly available *pwcmp* software[1]. The software uses a Bayesian method, which employs a maximum-likelihood-estimator to maximize the probability that the collected data explains the scaled quality scores under the Thurstone Case V assumptions. The optimization procedure finds a quality value for each pair of stimuli that maximizes the likelihood, which is modeled by the binomial distribution. Unlike standard scaling procedures, the Bayesian approach can robustly scale pairs of conditions for which there is unanimous agreement. Such pairs are common when a large number of conditions are compared. It can also scale the result of an incomplete and unbalanced pair-wise design, when not all the pairs are compared and some pairs are compared more often than the others.

The distribution parameters of the software are adjusted so that the difference of one quality value corresponds to the 75% preference rate. 75% rate is the mid-point between the same quality (i.e. 50% or random guess) and different quality (i.e. 100%), and it implies that only half of the observers were able to see a difference. Although this is very close to the *just noticeable difference (JND)* as a concept, we use the term *just objectionable difference (JOD)* to indicate that these quality scores are distances from the original perfect quality, as JOD indicates overall quality. That is, two stimuli may have several JNDs between them but they may both have 1 JOD difference from the original. In this sense, JOD values can be viewed as quality scores similar to MOS (or DMOS) values, and they can be used to understand the overall quality of the stimulus.

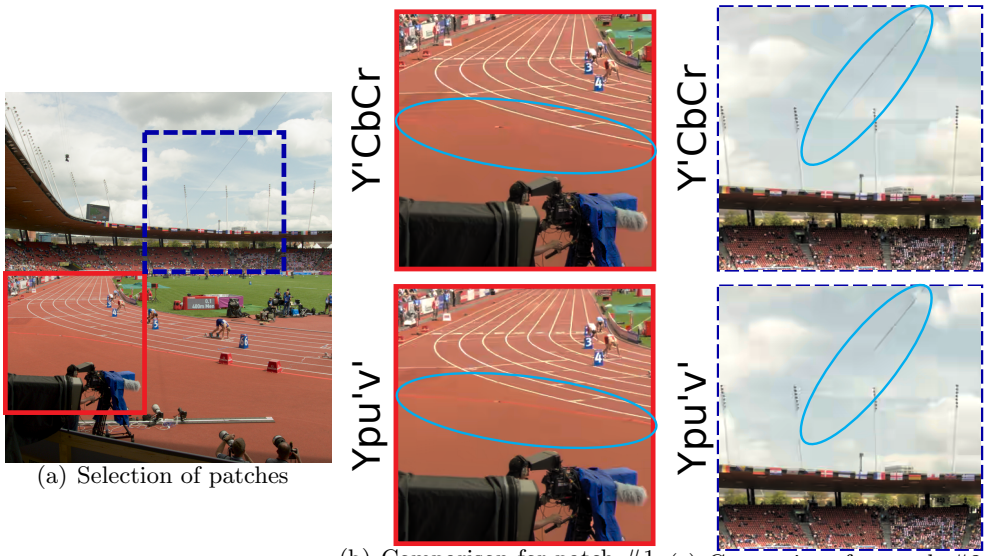
As the pairwise comparisons can provide only relative information about the quality, the JOD values are also relative. To maintain consistency across the video sequences, we always fix the starting point of the JOD scale at 0 for different distortions and thus the quality degradation results in negative JOD values.

Comparison of Pairwise Scaling Results

The comparison matrix for each video sequence was formed separately since each stimulus was compared to another stimulus with the same content. For each video sequence, the original uncompressed video was fixed to have zero JOD value in order to fix the relativity to the original video. Afterwards, the JOD values were found for the stimuli using the *pwcmp* scaling software. The confidence intervals were found using bootstrapping.

The resulting JOD values are reported in Figure 3.7 for each video sequence. The videos compressed with three color spaces have very similar JOD values. Looking at the scaled data, we can say that, overall, there is no significant difference between the video compression performances using tested color spaces despite the numerical differences. However, there are a few cases where a preference of using one color space over the other is evident, e.g. *Starting* scene at higher bit rates. In this sequence, there were two predominant regions of

[1] *pwcmp* toolbox for scaling pairwise comparison data `https://github.com/mantiuk/pwcmp`



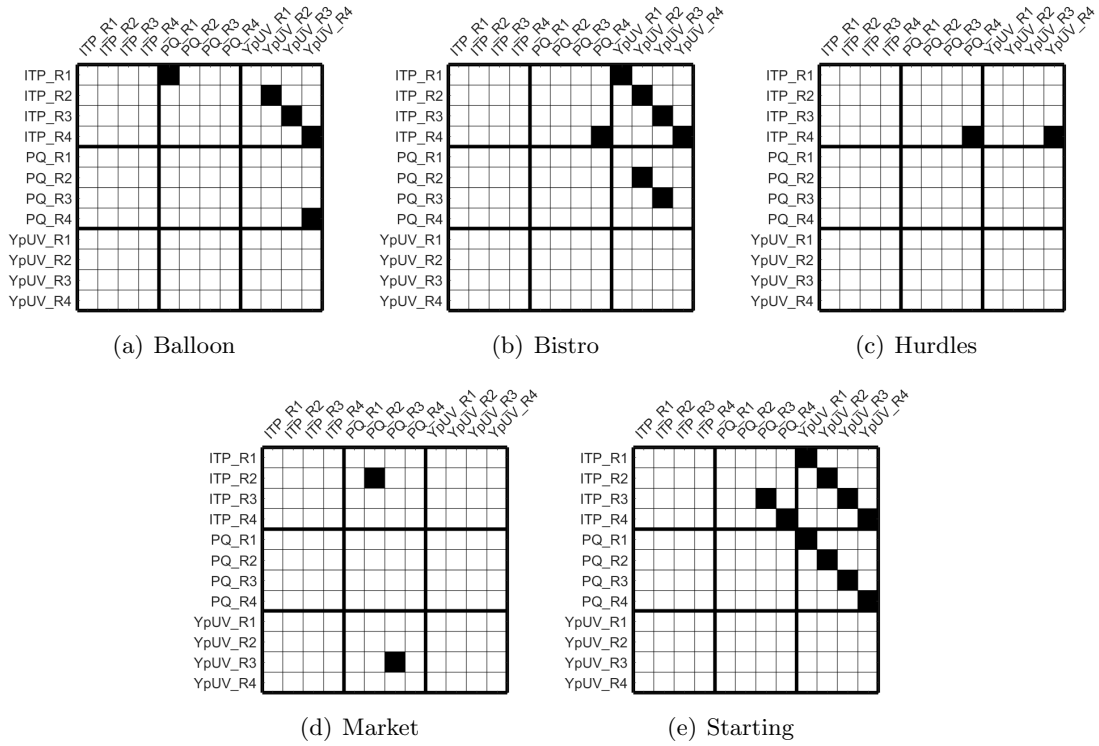
(a) Selection of patches (b) Comparison for patch #1 (c) Comparison for patch #2

Figure 3.8 – An example of the difference in compression performance between the Y'CbCr and Ypu'v' color spaces, both compressed at BR_2 level. The color spaces affect the artifacts differently in the (b) bottom-left (patch #1) corner of the scene and in the (c) top (patch #2) part of the scene.

interest, where the compression artifacts appeared to be the most obvious. These regions are, specifically, the red tape on the ground (patch #1, indicated with solid red box) and the wire which is visible in the sky (patch #2, indicated with dashed blue box), as shown in Figure 3.8. The red tape on the ground has better quality for the case of Ypu'v' and worse quality for Y'CbCr. On the other hand, the wire in the sky retains a bigger portion of it for the case of Y'CbCr and a smaller portion for Ypu'v'.

At high bit rates (BR_1 and BR_2), user ratings were highly dependent on which of these two regions they were focusing while making the comparison. Due to the conflicting appearance of the artifacts, the confidence intervals for *Starting* sequence for the cases of BR_1 and BR_2 quality levels are larger then those of other video sequences. At lower bit rates (BR_3 and BR_4), the details in the bottom (patch #1) part were corrupted in all the methods, resulting in more uniform responses and more similar JOD values.

In order to test the statistical significance between the color spaces, two methods were used. The first method we used was the statistical significance test of the *pwcmp* software. In order to test the significance of the compared pairs, the `pw_plot_ranking_triangles` and `pw_signifiance_matrix` functions were used. These functions use the covariance matrix, C, found as a result of the JOD calculation and calculate the probability of two conditions being different based on the variance of the difference between the said two conditions. The variance between the conditions i and j is found by $v = C_{i,i} + C_{j,j} - 2 \times C_{i,j}$. Assuming that the difference follows normal distribution, the significance of the pair is found with $\alpha = 5\%$.



(a) Balloon (b) Bistro (c) Hurdles

(d) Market (e) Starting

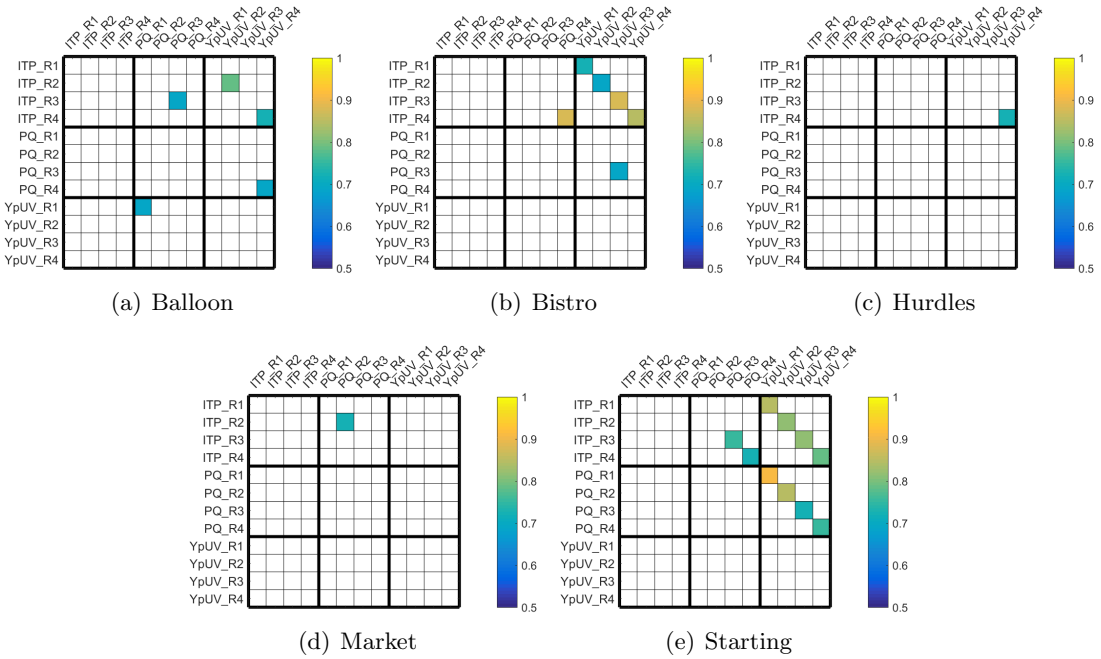
Figure 3.9 – Difference between test conditions after significance test on JODs. Only the conditions at the same bit rate are reported. Entries named as PQ refer to the color transformation with PQ and Y'CbCr. Black entries at position (i, j) indicate that stimulus i has been found to be significantly better than stimulus j, at 95% confidence. Similar results are obtained by performing a pairwise binomial test on raw (unscaled) data.

As a results of this significance test, several cases were found as significantly better than their counterpart as shown in Figure 3.9. This test was in accordance with the results from Figure 3.7, showing that Ypu'v' color space mainly has the worst effect on compression performance, while ITP is not significantly better than Y'CbCr except for a few cases.

Second, we conducted a binomial test between the different color spaces using the unscaled experimental data, only at the same bitrate as the differences are obvious for different bitrates. The results are shown in Figure 3.10. The colored cells show that the p-value of the test is lower than $\alpha = 5\%$. The associated intensity at the position (i, j) is not the p-value of the comparison, but it is the probability that stimulus i was selected over stimulus j as found in the test. The results of both significance tests are in agreement with each other.

Comparison of Objective Quality Scores

In addition to the subjective results, the video quality was predicted using two objective quality metrics: an objective quality metric for HDR video, i.e. HDR-VQM [NMDSLC15], and a color difference metric, i.e. ΔE_{2000} [LCR01]. HDR-VQM was computed using only the



(a) Balloon (b) Bistro (c) Hurdles

(d) Market (e) Starting

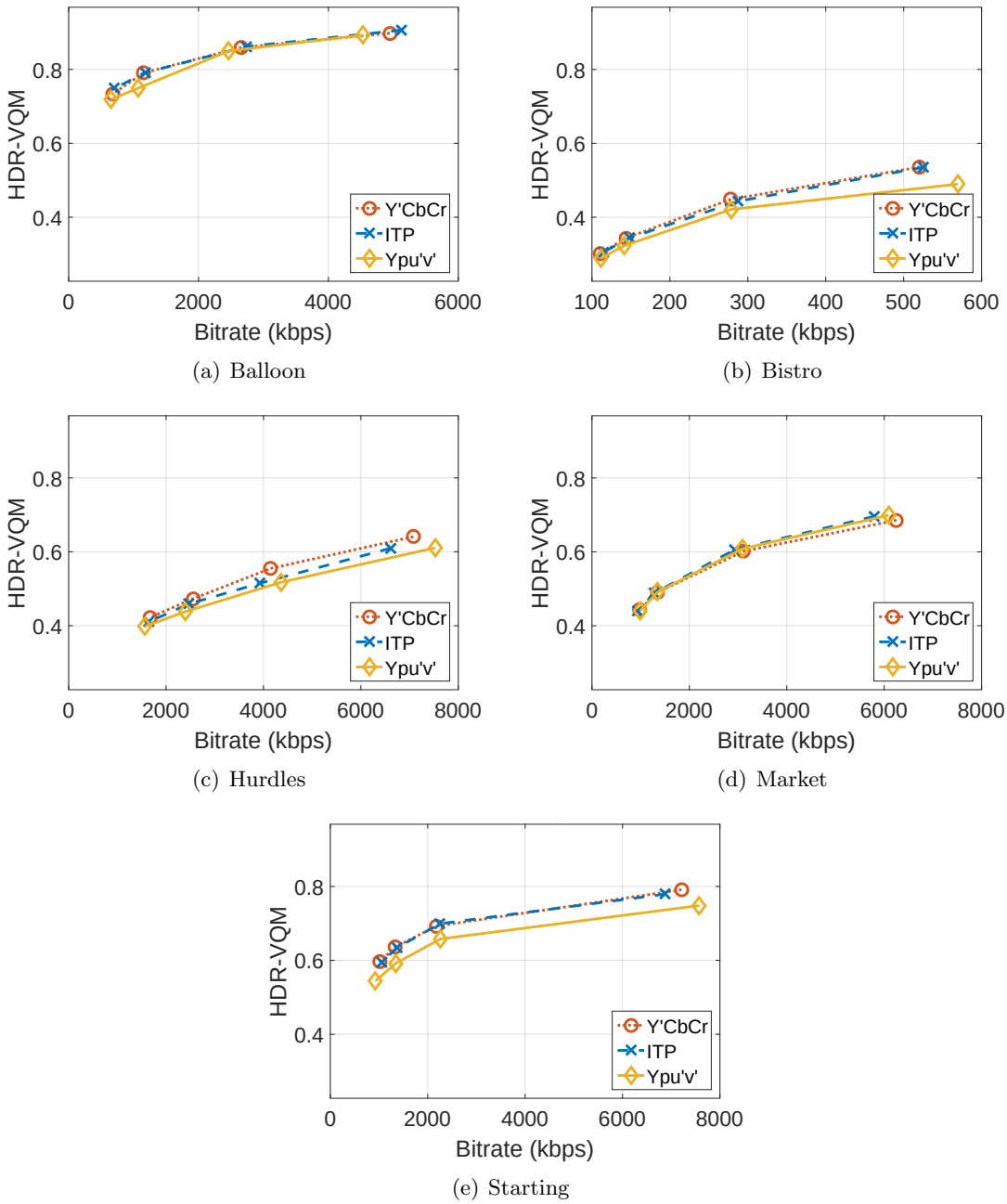
Figure 3.10 – Difference between test conditions after the binomial test on the raw (unscaled) experimental data. Only the conditions at the same bit rate are reported. Entries named as PQ refer to the color transformation with PQ and Y'CbCr. Colored entries at position (i, j) indicate that the p-value of the test is lower than 5%. Intensity values indicate the probability of stimulus i being significantly better than stimulus j.

luminance channel. The results of the HDR-VQM quality metric are shown in Figure 3.11, and the results of ΔE_{2000} are shown in Figure 3.12.

Comparing the same stimuli using the HDR-VQM objective metric, we found almost identical results to the subjective experiment. In most of the cases, compression with Ypu'v' color space results with the lowest quality, except for the Market scene where there is almost no difference in scores. We observed a similar situation in the Balloon scene, where only at low bit rates, a minor difference between the three methods is found. Notice that we selected the HDR-VQM metric for objective evaluation since this is the only HDR full-reference metric specific to *video*.

The compressed videos were also evaluated using ΔE_{2000} color difference metric. ΔE_{2000} was calculated for each frame, and the calculation results were placed in a vector. The markers in Figure 3.12 indicate the average ΔE_{2000} value for each case of compressed video, and the whiskers indicate the span of ΔE_{2000} values from minimum to the maximum. Looking at the results reported, we can say that there is not any overall conclusion. In terms of color difference, Y'CbCr appears to yield less color difference for *Starting* sequence whereas ITP yields more color difference for *Market* and *Bistro* sequences. For all other cases, the differences are not significant.

Considering the ΔE_{2000} results, we would expect that Y'CbCr should have higher



(a) Balloon

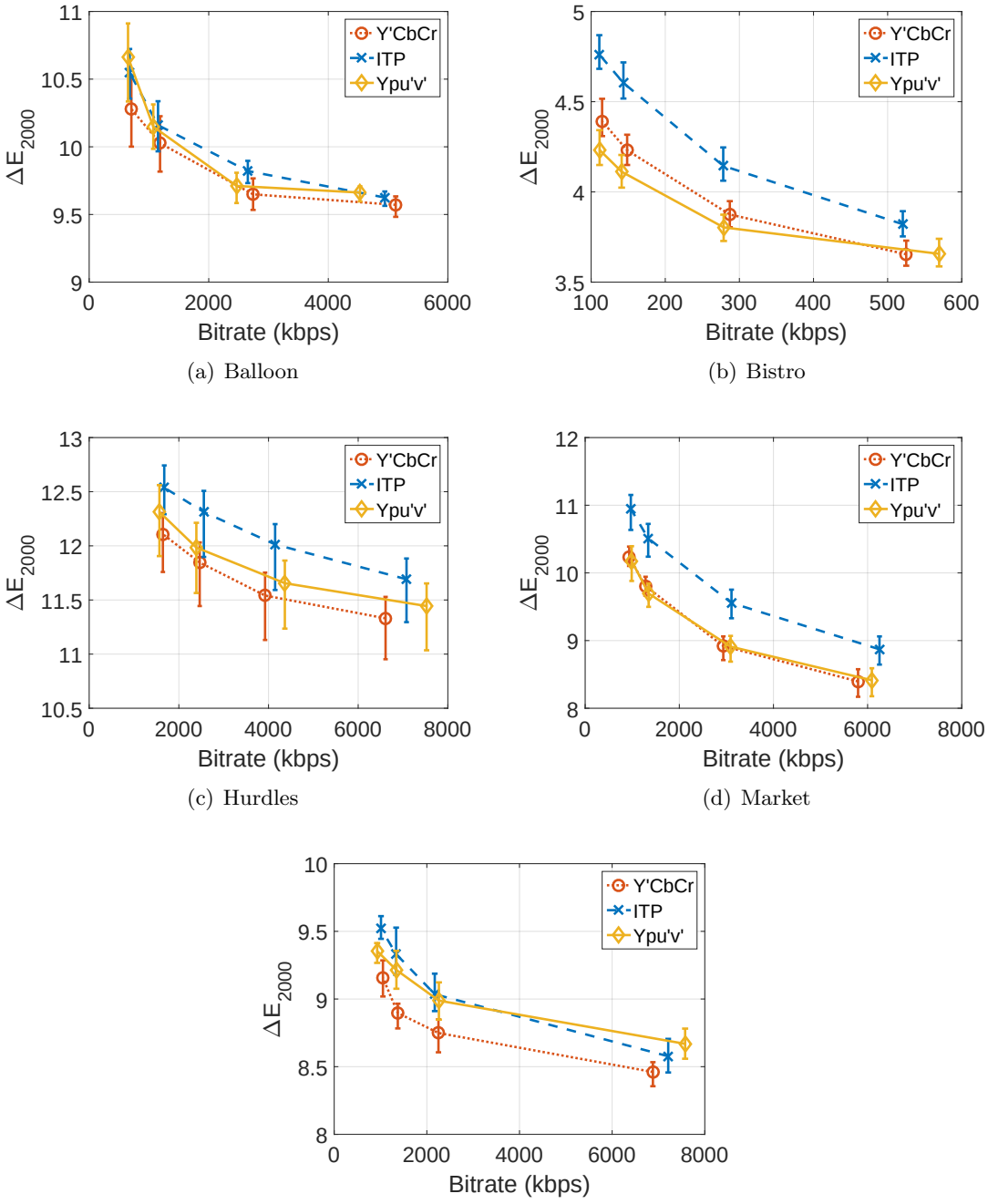
(b) Bistro

(c) Hurdles

(d) Market

(e) Starting

Figure 3.11 – The results obtained by comparing all the scenes for the three color spaces using the HDR-VQM metric. All scores are normalized, where 1 means perfect quality and lower scores represent a decrease in quality.



(a) Balloon

(b) Bistro

(c) Hurdles

(d) Market

(e) Starting

Figure 3.12 – The results obtained by comparing all the scenes for the three color spaces using the ΔE_{2000} metric. Higher ΔE_{2000} scores represent an increase in the color difference, and stimuli more similar to the original video yield lower scores.

preference rates (or lower JOD values) for *Starting* sequence. Similarly, ITP should have lower preference rates for *Market* and *Bistro* sequences. However, the results of ΔE_{2000} are different from what has been observed in both subjective experiment results and HDR-VQM objective metric results. Even though it is a color-blind metric, HDR-VQM predicts the quality of the compressed videos better than ΔE_{2000}. Nevertheless, we observe that the prediction accuracy of HDR-VQM is not sufficient to predict the quality scores found in the subjective experiment precisely, e.g. *Hurdles* and *Bistro* have significantly lower predicted quality than the actual one. The color difference metric ΔE_{2000} has been found to predict the visual quality poorly. These findings are in agreement with the previous works [SW03, OCHZ09, OJKP16, ZVD17].

3.3 Discussion

In this chapter, the effect of color space on compression performance of HDR videos was investigated. RGB videos were transformed to either Y'CbCr after PQ EOTF encoding; ITP; or Ypu'v' color space. The videos were compressed using HEVC Main-10 profile with HM 16.5 after `4:2:0` chroma subsampling. Two subjective tests were conducted: a preliminary experiment to find the bit rates at approximately one JOD level from each other, and the main color space experiment where five HDR video sequences, rendered at four previously found bit rates, were compared. To verify the significance of the user preference, the significance test on JODs was performed and validated using the pairwise binomial test. The resulting data was finally compared with the results generated by two objective quality metrics: HDR-VQM and ΔE_{2000}. The obtained results after both subjective and objective analysis of the effects of color space on compression and quality were published in [ZHV+17].

The results of our test on the effect of color space on compression performance for HDR video reveal that the influence of color space on coding performance is, in general, small. With the exception of a few specific, content-dependent cases, we did not find evidence of the ITP color space being significantly better than Y'CbCr. Instead, we observed that Ypu'v' has in general a lower performance for coding, although the differences in quality are generally small. Even in those cases where a difference can be observed, we found that this is strongly content dependent, and is highly influenced by the visual attention patterns of each observer. This produces larger confidence intervals in the estimated quality scores, indicating that the problem of assessing visual quality for small differences in the magnitude of the distortion across stimuli (such as those produced by changing the color space) can be strongly subject-dependent and requires both a careful choice of test material and appropriate analysis tools. We did so in this chapter by selecting test stimuli through a preliminary subjective study, aimed at well conditioning the scaling procedure carried out after the main study to find JOD quality scores.

Our results also confirm that the HDR video quality metric, HDR-VQM, can predict the

general trend and ranking between stimuli, but it is not sufficiently precise to distinguish very tiny perceptual differences and predict absolute quality levels. This motivates further studies in that direction. The color difference results of ΔE_{2000} cannot be generalized as the results are highly content dependent. Moreover, ΔE_{2000} results are not in agreement with the subjective quality results or the objective HDR-VQM results. Both this disagreement and the HDR-VQM's ability to predict the general trend of the subjective results indicate that the perceived quality for HDR video compression is dominated by the structural distortion caused by the changes in the luminance channel. These findings are in agreement with the previous studies on color in the case of SDR content [SW03, OCHZ09, OJKP16]. Therefore, we expect that the color-blind (or luminance-only) metrics will perform at least as efficiently as the color metrics for the HDR compression scenario.

Chapter 4

Performance Evaluation of Full-Reference HDR Image Quality Metrics

Contents

As a concept and technology, high dynamic range imaging augments current imaging technologies. It enables the acquisition and reproduction of everyday scenes with a larger brightness range as well as a wider range of color. Very bright and very dark objects can be simultaneously captured and displayed together [DLCMM16]. These properties of HDR make it a great tool to improve the human experience of visual media, compared to standard dynamic range technologies. HDR image and video cameras and displays have become available for commercial market, and parts of HDR storage and compression are in the process of standardization within MPEG [LFH15, HRE16] and JPEG [Ric13] communities. Therefore, it is necessary to understand the capabilities and shortcomings of the objective HDR quality assessment algorithms.

Compared to SDR quality assessment, new challenges emerge for the evaluation of HDR visual quality [NdSLC$^+$16b]. The visibility of the artifacts are increased with the increased luminance and widened color range of HDR. On the other hand, this increase in

brightness also alters where viewers focus and their attention patterns compared to the case of SDR [NDSLCP14b]. Additionally, increased brightness augments the effect of color distortions within the overall perception of quality [Fai13]. HDR quality is affected by all these factors. Although they are time-consuming, expensive to design, and need expertise, the most accurate methods of assessing HDR quality are subjective quality assessment experiments. In addition, special equipment such as HDR displays and light meters are required in the case of HDR. All of these limitations and the growing interest in HDR led to increasing research for the design and fine-tune of *full-reference* HDR quality metrics in the past few years [MKRH11, NDSLC15, NMDSLC15, AMS08, NLCV+16].

As discussed in detail in Section 1.4.2, two of the full-reference quality metrics developed exclusively for HDR images and videos, HDR-VDP [MKRH11] and HDR-VQM [NDSLC15] respectively, model the human visual system and estimate the visual quality according to the HVS model. The early stages of the HVS such as intra-ocular scattering, luminance masking, and achromatic response of the photoreceptors are accurately modeled by HDR-VDP. HDR-VQM also models the HVS by considering the average fixation duration of the human eye during the calculation of spatio-temporal errors and humans assess the videos by making "continuous assessments of the impact of short term errors" during the pooling step of its own quality estimation.

On the other hand, the quality metrics developed for SDR content can be used also for HDR content. As discussed in Section 1.2.1, these metrics can be arithmetic (PSNR, MSE), structural (SSIM [WBSS04] and its multiscale version MSSIM [WSB03]) and information-theoretic (e.g., VIF [SB06]). The SDR quality metrics were developed for the gamma-corrected 8-bit (or 24-bit if colored) images, and the gamma-correction ensures that the pixel values of these SDR images are perceptually linear. However, pixel values of HDR images are captured and stored as proportional to the physical luminance of the scene, and they are not perceptually linear. Human perception has a complex behavior: it can be approximated by a square-root in low luminance values and is approximately proportional to luminance ratios in higher luminance values, as expressed by the DeVries-Rose and Weber-Fechner laws, respectively [KP86]. Thus, in order to employ these metrics, the HDR content needs to be perceptually linearized, e.g., using a logarithmic or perceptually uniform (PU) encoding [AMS08].

Both the metrics developed exclusively for HDR content and the SDR metrics with perceptual linearization are compared against the mean opinion scores (MOS) of the subjects in several subjective studies for compression scenarios [VDSLD14, HBP+15, NDSLCP13, NDSLCP12]. The purpose of these studies is to show the performance of the considered objective quality metrics; however, the results and the conclusions of these studies differ from each other. For instance, the correlation values of PU-SSIM, i.e., SSIM metric applied after the PU encoding of [AMS08], differ substantially between the study of Narwaria et al. [NMDSLC15] and that of Valenzise et al. [VDSLD14]. This difference can be explained by considering the two studies using different sets of stimuli. While 50

subjectively annotated HDR images compressed using JPEG, JPEG 2000 and JPEG-XT encoders are used in [VDSLD14], Narwaria et al. [NMDSLC15] used a larger set using a number of subjectively annotated databases with different experimental conditions which have different distortions. Apart from these, Hanhart et al. [HBP+15] evaluate objective quality metrics on HDR images with a single distortion: compression with JPEG-XT encoder. Even though all of these studies have their strengths and advantages, it is very hard to draw a simple and clear conclusion for the considered objective quality metrics' performance.

In this chapter, we aim to bring more clarity to this field, by providing an extensive, reliable, and consistent benchmark of the most popular HDR image fidelity metrics. To achieve this, a new database was created using different image encoders and pixel encoding functions. In addition, all the available public HDR image quality databases were collected, and the MOS values of all images were aligned using the iterated nested least square algorithm (INLSA) proposed in [PW03b], in order to obtain a common subjective scale. This aligned database consists of a total of 690 compressed HDR images, and it is the largest set on which HDR metrics have been tested so far to the best of our knowledge. Using this large set of data, we analyze the prediction accuracy and the discriminability (i.e., the ability to detect when two images have different perceived quality) of 25 fidelity metrics, including those tested in MPEG standardization.

The main contributions include:

- the most extensive evaluation (using 690 subjectively annotated HDR images) of HDR full-reference image quality metrics available so far;
- the proposal of a new subjective database with 50 distorted HDR images, combining 3 image codecs and 2 pixel encoding algorithm (SMPTE-2084 Perceptual Quantization [SMP14] and a global tone-mapping operator);
- an evaluation of metric discriminability, that complements the conventional statistical accuracy analysis, based on a novel classification approach.

HVS has different perception mechanisms for image and video because of the fixation duration of the eye and because the temporal characteristics of image and video are different. Therefore, the quality assessment of image and video are different. However, some commonly used image quality metrics –e.g. PSNR, MSE, or SSIM– are often applied to the cases of video on a frame-by-frame basis. Therefore, the result of this work could be indicative of frame-by-frame objective metrics performance in the case of video as well.

4.1 Considered Subjective Databases

There is a large number of publicly available repositories of high-quality HDR pictures [DM04, Fai07, DM08, EMP13, pfs15]. They include high-resolution and high-quality

Table 4.1 – Number of observers, subjective methodology, number of stimuli, compression type and tone mappings employed in the HDR image quality databases used in this paper. TMOs legend: *AS*: Ashikmin, *RG*: Reinhard Global, *RL*: Reinhard Local, *DR*: Durand, *Log*: Logarithmic, *MT*: Mantiuk.

Database No	Observers	Methodology	Stimuli	Compression	TMO
#1 [NDSLCP13]	27	ACR-HR	140	JPEG†	iCAM-06 [KJF07]
#2 [NDSLCP14a]	29	ACR-HR	210	JPEG 2000†	AS [Ash02] RG [RSSF02] RL [RSSF02] DR [DD02] Log
#3 [KHR+15]	24	DSIS	240	JPEG-XT	RG [RSSF02] MT [MMS06]
#4 [VDSLD14]	15	DSIS	50	JPEG† JPEG 2000† JPEG-XT	Mai [MMM+11]
#5	15	DSIS	50	JPEG† JPEG 2000†	Mai [MMM+11] PQ [MND12, SMP14]

† The distorted images are generated through a scalable coding scheme [WS06]: the HDR image is converted to SDR using a TMO; then, the SDR picture is encoded & decoded by a legacy codec; finally, the image is converted back to HDR range.

undistorted images, even with luminance measurements of each image for some of these repositories. Compared to this availability of undistorted HDR images, the number of publicly available subjectively annotated HDR image quality databases is very small.

We selected four publicly available HDR image quality assessment databases for this analysis. In addition, we propose a new database which is described in Section 4.1. Each of these databases contains compressed HDR images and MOS values for these HDR images. The compression algorithms, number of observers, the number of stimuli used, and the experiment methodologies are different, and these parameters are summarized in Table 4.1. The interested reader can refer to original publications for further details.

Database #1 – Narwaria et al. (2013) [NDSLCP13]

In their work, Narwaria et al. [NDSLCP13] proposed a tone mapping based HDR image compression scheme and conducted a subjective experiment for the subjective quality assessment. The subjective experiment was conducted in a controlled test room which had a 130 cd/m^2 room illumination. A SIM2 HDR47E S 4K display was used for the experiment, and the distance from the display was set as $3 \times H$ (approximately 178 cm). The participants were asked to rate overall image quality using the Absolute Category Rating with Hidden Reference (ACR-HR) methodology, employing a five-level discrete scale where 1 is bad and 5 is excellent quality. The test material was obtained from 10 pristine HDR pictures, including both indoor and outdoor, natural or computer-generated scenes. The distorted images are generated through a backward compatible scheme [WS06]:

the HDR image is the first converted to SDR by using a tone mapping operator (TMO); then, the SDR picture is coded using a legacy image codec; finally, the compressed image is expanded by inverse tone mapping to the original HDR range. The coding scheme in [NDSLCP13] employs iCAM06 [KJF07] as TMO, and JPEG compression at different qualities. In addition, the authors proposed two criteria to optimize the quality of the reconstructed HDR. As a result, a total of 10 contents $\times$ 7 bitrates $\times$ 2 optimization criteria = 140 test images were evaluated. 27 subjects participated the test. This database is publicly available at `http://ivc.univ-nantes.fr/en/databases/JPEG_HDR_Images/`.

The analysis in [NDSLCP13] shows that Mean Squared Error (MSE) and Structural Similarity Index Measure (SSIM) perform well in estimating human predictions and ordering distorted images when each content is assessed separately. However, these results do not apply when different contents are considered at the same time. HDR-VDP-2 was found to be the best performing (in terms of linear correlation with MOSs) metric, but not statistically different from the metric proposed in [NLM+12].

Database #2 – Narwaria et al. (2014) [NDSLCP14a]

In another work, Narwaria et al. [NDSLCP14a] subjectively assess the effect of using different TMOs on HDR image compression. The test material includes 6 original scenes, both indoor and outdoor, from which a total of 210 test images were created using JPEG 2000 image compression algorithm after the application of several TMOs, including Ashikmin [Ash02], both local and global versions of Reinhard [RSSF02], Durand [DD02], and logarithmic TMO. The experiment setup was the same as in Narwaria et al. (2013) Database #1 described above. The subjective test is conducted with 29 observers using ACR-HR methodology.

Results show that the choice of TMO greatly affects the quality scores. It is also found that local TMOs, with the exception of Durand's, generally yield better results than global TMOs as they tend to preserve more details. No evaluation of objective quality metrics is reported in the original paper [NDSLCP14a].

Database #3 – Korshunov et al. (2015) [KHR+15]

In the study of Korshunov et al. [KHR+15], an HDR image quality database, publicly available at `http://mmspg.epfl.ch/jpegxt-hdr`, was created using backward-compatible JPEG-XT standard [Ric13] with different profiles and quality levels. For this database, 240 test images were produced, using either Reinhard [RSSF02] or Mantiuk [MMS06] TMO for the base layer, 4 bit rates for each original image and 3 profiles of JPEG-XT. The test room was illuminated with a 20 lux lamp, and a SIM2 HDR display was used. At any time, 3 observers took the test simultaneously. The subjective scores were collected from 24 participants, using Double Stimulus Impairment Scale (DSIS) Variant I methodology, i.e., images were displayed side-by-side, one of the images was the reference and the other

the distorted one.

This subjective database was used in the work of Artusi et al. [AMR+15]. In this work, an objective evaluation of JPEG-XT compressed HDR images was carried out. The results show that SDR metrics such as PSNR, SSIM, and multi-scale SSIM (MSSIM) give high correlation scores when they are used with the PU encoding of [AMS08], while the overall best correlated quality metric is HDR-VDP-2.

Database #4 – Valenzise et al. (2014) [VDSLD14]

Valenzise et al. [VDSLD14] were the first to collect subjective data with the specific goal to analyze the performance of HDR image fidelity metrics. Their database is composed of 50 compressed HDR images, obtained from 5 original scenes in the Fairchild HDR image survey [Fai07]. Three different coding schemes were used to produce the test material, i.e., JPEG, JPEG 2000 and JPEG-XT. In the first two cases, the HDR image is first tone mapped to SDR using the minimum-MSE TMO proposed by Mai et al. [MMM+11]. The images were displayed on a SIM2 HDR47E S 4K display, with an ambient luminance of 20 cd/m^2. Subjective scores were collected using DSIS methodology, i.e., pairs of images (original and distorted) were presented to the viewers, who had to evaluate the level of annoyance of distortion in the second image on a continuous quality scale ranging from 0 to 100, where 0 corresponds to very annoying artifacts and 100 to imperceptible artifacts. Fifteen observers rated the images. The database is available at `http://webpages.l2s.centralesupelec.fr/perso/giuseppe.valenzise/download.htm`.

The results of this study showed that SDR fidelity metrics could accurately predict image quality, provided that the display response is somehow taken into account (in particular, its peak brightness), and that a perceptually uniform (PU) encoding [AMS08] is applied to HDR pixel values to make them linear with respect to perception.

Database #5 – New subjective database

In addition to the databases described above, we construct a new subjective HDR image database of 50 images, as an extension to the previous work of Valenzise et al. [VDSLD14]. The new database features 5 original images, selected in such a way to be representative of different image features, including the dynamic range, image key, and spatial information. The five images are shown in Figure 4.1. The images *"Balloon"*, *"FireEater2"*, and *"Market3"* are chosen among the frames of the MPEG HDR sequences proposed by Technicolor [LLF13]. *"Showgirl"* is taken from Stuttgart HDR Video Database [FGE+14]. *"Typewriter"* is from HDR photographic survey dataset [Fai07]. All images have either 1920×1080 pixels spatial resolution, or are zero-padded to have the same resolution.

Similarly to [VDSLD14], the test images are obtained by using a backward compatible HDR coding scheme [WS06], using JPEG and JPEG 2000 (with different bitrates) as SDR codecs. We did not include JPEG-XT in this experiment since some of the contents we

(a) "Balloon" (b) "FireEater2" (c) "Market3"

(d) "Showgirl" (e) "Typewriter"

Figure 4.1 – Original contents for the new proposed image database described in Section 4.1, rendered using the TMO in [MDK08].

selected (e.g., *"Showgirl"* and *"Typewriter"*) were already part of the Database #3. In order to convert HDR to SDR, we use two options: *i)* the TMO of Mai et al. [MMM+11]; and *ii)* the electro-optical transfer function SMPTE ST 2084 [MND12, SMP14], commonly known as Perceptual Quantization (PQ). The latter is a fixed, content-independent transfer function which has been designed in such a way that the increments between codewords have minimum visibility, according to Barten's contrast sensitivity function [Bar99]. We choose this transfer function as an alternative to tone mapping, as it has been proposed as the anchor scheme in current MPEG HDR standardization activities [LFH15]. Both PQ and Mai et al.'s TMO were applied per color channel.

The test environment and methodology were carefully controlled to be the same as in Database #4 (Valenzise et al. (2014)) [VDSLD14]. The DSIS methodology was employed, where the reference image was shown for 6 seconds, followed by 2 seconds of mid-gray screen and 8 seconds of degraded image. The asymmetry in timing between distorted and reference image was determined in a pilot test, taking into account the fact that the reference image is shown several times, while the degraded image is different at each round and requires a longer evaluation interval. After both the original and distorted images are displayed, the observer takes all the time she/he needs to rate the level of annoyance on the same continuous scale as in [VDSLD14]. The sequence of tested images is randomized to avoid context effects [DS12]. Moreover, too bright ("Market3") and too dark ("FireEater2") stimuli are not placed one after another in order to avoid any masking caused by sudden brightness change. In addition to randomization, stabilizing images (one from each content and featuring each quality level) are shown at the beginning of the experiment to stabilize viewers' votes (which are discarded for those images).

In addition to the contents reported in Figure 4.1, a small subset of the stimuli of

Database #4 was included in the test. This enabled aligning the two databases, #4 and #5, in order for the corresponding MOS values to be on the same scale [PEB+11]. Thus, in the following, we will refer to the union of these two databases as Database #4 & 5.

A panel of 15 people (3 women, 12 men; average age of 26.8 years), mainly Ph.D. students naive to HDR technology and image compression, participated in the test. Subjects reported normal or corrected-to-normal vision. The outlier detection and removal procedure described in BT.500-13 [ITU12b] resulted in no detected outlier. Then, mean opinion scores and their confidence interval (CI) were computed assuming that data follows a *t-Student* distribution. These scores, together with the test images, are available at `http://webpages.l2s.centralesupelec.fr/perso/giuseppe.valenzise/download.htm`.

4.2 Alignment of MOS Values

During the training phase of subjective experiments, the subjects are generally instructed to use the whole range of grades (or distortions) in the scale while evaluating. However, the quality of the test material for different experiments may not be the same when they are compared to each other. The viewers may not share the same understanding and expectations of image or video quality. Hence, the MOS values generally do not show the absolute quality of the stimuli. In Figure 4.2, we observe the MOS distribution for non-aligned databases as a function of the HDR-VQM metric. Due to the characteristics of the experiments and the test material of each database, a similar level of impairment in the subjective scale may correspond to very different values of the objective metrics. Therefore, in order to use the MOS values of different subjective databases in a consistent way, these need to be mapped onto a common quality scale.

In order to align the MOS values of all five HDR image databases, we use the *iterated nested least square algorithm* (INLSA) proposed in [PW03b][1]. INLSA aligns the subjective quality values collected in different subjective experiments using some common external variables. These external variables are chosen as the objective quality metrics' estimations for the case of multimedia quality.

The alignment is done by changing the weights of the objective quality metrics, w, and changing the weights of the subjective quality scores, (a_i, b_i), iteratively. Before any other operation, the subjective quality scores s_i from the i^{th} experiment are normalized between 0 and 1, according to Equation 4.1:

$$s_i = \frac{s_i^o - best_i}{worst_i - best_i} \tag{4.1}$$

where s_i^o are the original values of the subjective scores, $best_i$ and $worst_i$ are the best and worst subjective quality values repectively. After the normalization of the scores, INLSA

[1]INLSA implementation on Matlab was downloaded from `http://www.its.bldrdoc.gov/resources/video-quality-research/guides-and-tutorials/insla-code.aspx`

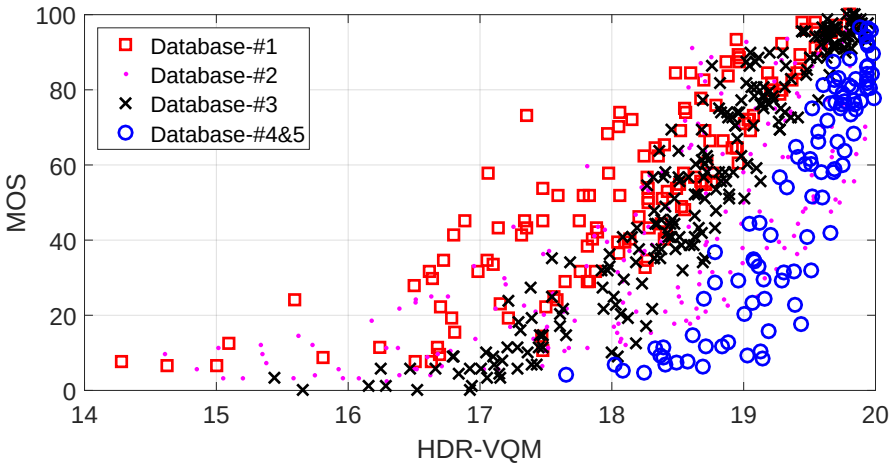


Figure 4.2 – MOS vs HDR-VQM scores before INLSA alignment.

brings all the subjective scores onto a common scale by changing the weights, a_i and b_i:

$$\tilde{s}_i = a_i s_i + b_i 1 \tag{4.2}$$

where s_i are the normalized subjective quality scores, $\tilde{s}_i$ are the resulting 'corrected' quality scores. After this first step, iteration is continued with the second step to find the best weights for the objective quality scores, q:

$$\tilde{s} \approx \hat{s} = Pw \tag{4.3}$$

P is the parameter matrix $P = [p_1 \;\; p_2 \;\; \dots \;\; p_r]$ which consists of parameter vectors p_i for r different objective quality metrics, and p_i consists of n quality scores q_j, $p_i = [q_1 \;\; q_2 \;\; \dots \;\; q_n]^\intercal$.

INLSA does not change the images themselves. It only changes the subjective quality values. This means that the objective metric results are not changed after INLSA alignment. Thus, INLSA only alters the subjective quality values in a linear manner, and it inherently assumes that the relationship between the objective quality estimates and the collected subjective quality values is linear. Considering that each HDR image has a distinct distortion, the objective quality metric score for that image should be unique to that particular image. However, it has been discussed that there are different perceptual effects [DS12] that influence human subjects' vote during the experiment. Hence, it is possible for these HDR images to have different absolute quality scores. This makes using the linear model feasible. Use of a non-linear model, on the other hand, may lead to biasing the data. Such kind of alignment may change the MOS values in a non-linear way, which is not intended in the original experiment in the first place. Additionally, we believe that non-linear correction should not be done during the alignment, but in the metric itself. These points show that a linear alignment is both necessary and sufficient for the task at hand.

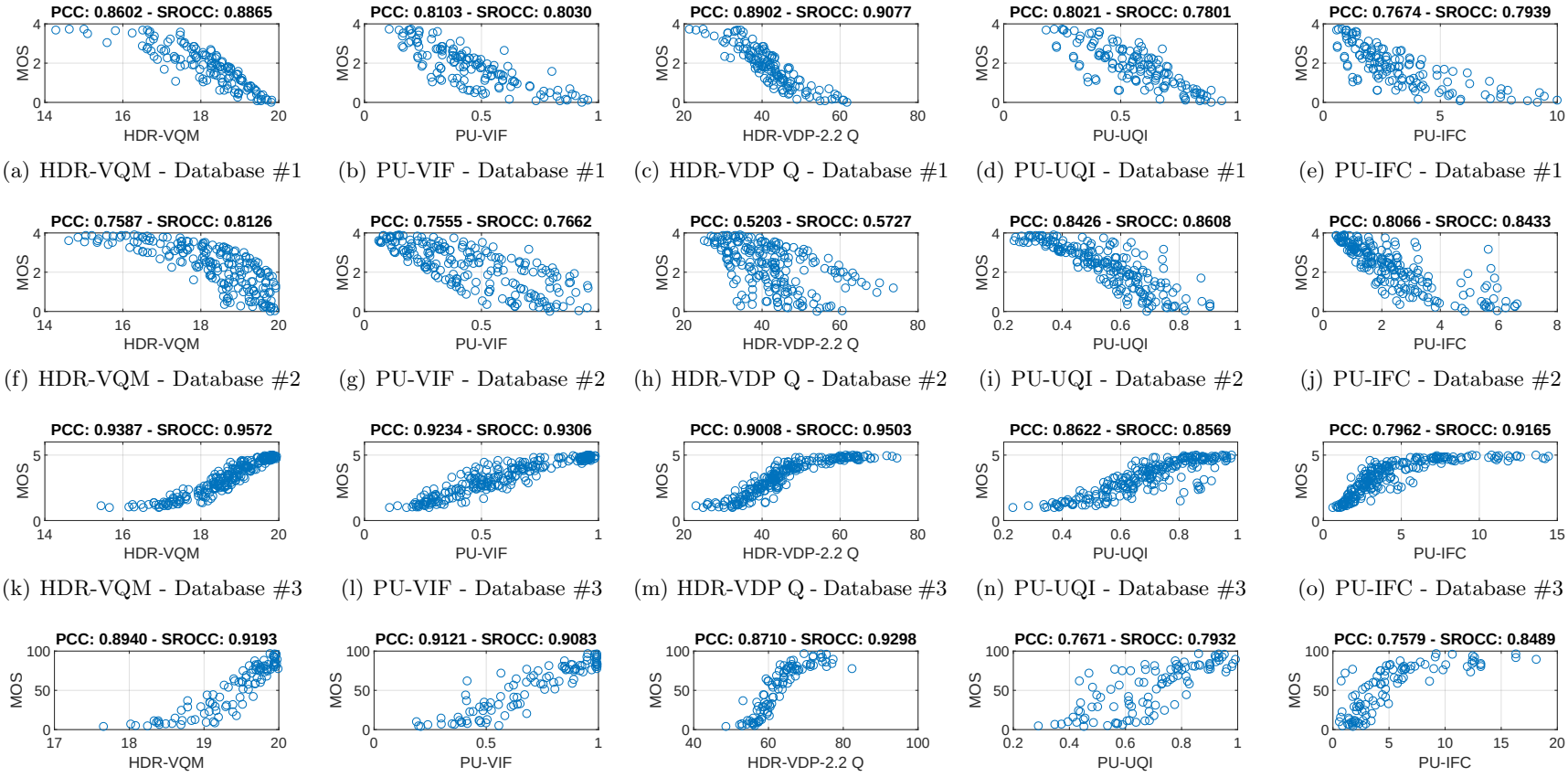
Selection of the Anchor Metrics

INLSA requires objective parameters (i.e. objective quality metric results) for the alignment, under the assumption that those are linear and sufficiently well correlated with respect to MOS. Therefore, we analyzed the considered metrics (see Section 4.3.1) in order to select the best candidates for this operation. We call these metrics 'anchor metrics'.

To select any metric as an anchor metric, we need to be sure that the quality estimation of this metric is accurate for the considered cases and robust to different conditions presented by the considered databases. Selecting very few metrics will dominate the results in their favor. On the other hand, using all the metrics will introduce noise and reduce the effectiveness of the INLSA alignment. Thus, the most correlated 5 metrics were chosen for alignment in order to reduce the dominance of any particular metric and avoid introducing noise. Since PCC is a correlation index showing the linearity of the data and SROCC is a correlation index showing the monotonicity of the data, we found the 5 metrics which have the highest value for the product of PCC and SROCC as shown in Table 4.2: HDR-VDP-2.2, HDR-VQM, PU-IFC, PU-UQI, and PU-VIF (the calculation of PU-metrics will be explained in detail in Section 4.3.1). The linear behavior of the metric results is clear from the plots of subjective quality vs objective quality from Figure 4.3.

Table 4.2 – Selection of Metrics for INLSA alignment - Correlation indices were calculated without applying non-linear fitting prior to calculation. Last column indicates the product of PCC and SROCC for each metric. Bold typeface indicates the selected metrics.

Metrics	PCC	SROCC	Product
HDR-VQM	0.859	0.894	0.768
PU-VIF	0.843	0.842	0.710
HDR-VDP-2.2 Q	0.793	0.836	0.663
PU-UQI	0.810	0.818	0.663
PU-IFC	0.781	0.847	0.661
Log-IFC	0.779	0.846	0.659
Photometric-UQI	0.805	0.812	0.654
PU-MSSIM	0.756	0.864	0.653
Log-UQI	0.801	0.810	0.649
Photometric-IFC	0.765	0.831	0.636
PU-SSIM	0.699	0.860	0.601
mPSNR	0.718	0.745	0.535
Log-PSNR	0.716	0.735	0.526
Log-SSIM	0.552	0.856	0.472
tPSNR-YUV	0.639	0.649	0.414
Photometric-VIF	0.614	0.650	0.399
Log-MSE	0.540	0.735	0.397
Log-VIF	0.596	0.646	0.385
PU-PSNR	0.613	0.611	0.374
$CIE\ \Delta E^{S}_{00}$	0.574	0.624	0.358
$CIE\ \Delta E_{00}$	0.552	0.555	0.307
PU-MSE	0.468	0.611	0.286
Photometric-SSIM	0.417	0.590	0.246
Photometric-PSNR	0.437	0.464	0.203
Photometric-MSE	0.282	0.451	0.127



(a) HDR-VQM - Database #1 (b) PU-VIF - Database #1 (c) HDR-VDP Q - Database #1 (d) PU-UQI - Database #1 (e) PU-IFC - Database #1

(f) HDR-VQM - Database #2 (g) PU-VIF - Database #2 (h) HDR-VDP Q - Database #2 (i) PU-UQI - Database #2 (j) PU-IFC - Database #2

(k) HDR-VQM - Database #3 (l) PU-VIF - Database #3 (m) HDR-VDP Q - Database #3 (n) PU-UQI - Database #3 (o) PU-IFC - Database #3

(p) HDR-VQM - Database #4 (q) PU-VIF - Database #4 (r) HDR-VDP Q - Database #4 (s) PU-UQI - Database #4 (t) PU-IFC - Database #4

Figure 4.3 – Plots of MOS vs objective quality scores for the selected objective metrics selected showing the linearity of the metric estimations

Alignment Results

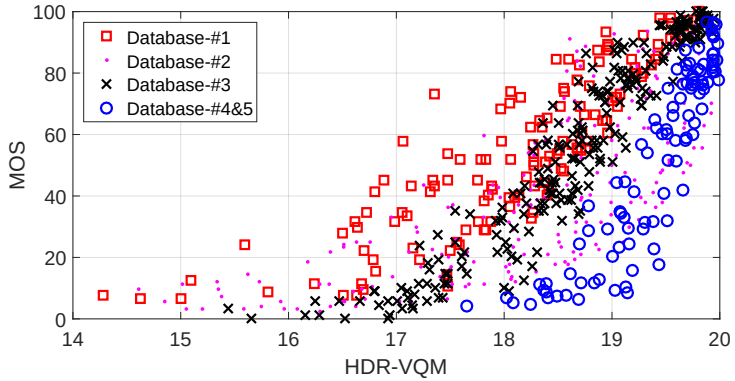
The MOS values of all of the 5 databases were brought together and aligned using the INLSA algorithm with the help of five anchor metrics selected. The scatter plots of MOS values vs. objective quality values estimated by HDR-VQM and PU-VIF metrics after alignment can be seen in Figure 4.4. It can be observed that data points having similar objective quality values have similar MOS values after INLSA alignment. After the alignment, all the MOS values were mapped onto a common subjective scale, and they can be used in the evaluation of the objective quality metrics.

From Figure 4.4.(b), 4.4.(d) and the initial observations of the test images, we notice that the images in Database #2 [NDSLCP14a] have very different characteristics compared to others, and MOS values are much more scattered than other databases after the alignment. This behavior is also evident in Figure 4.3. The metric results become more scattered for the case of Database #2. This is mainly due to the characteristics of this database, i.e., the stimuli were mainly obtained by changing the tone mapping algorithm used in the compression, including many TMOs which are definitely not adapted to be used in coding as they produce strong color artifacts in the reconstructed HDR image, and they are therefore not used in any practical coding scheme. Also, different kinds of distortion are present simultaneously, such as color banding, saturation etc. In some cases, it is noticed that false contours are generated, and some color channels are saturated. This makes the quality assessment problem much more difficult for any objective metric. It may be the case for Database #2, these artifacts decrease the subjective quality a lot whereas the saturation and false contours may limit the decrease in objective quality. Initial inspection of both test images and objective metric results indicate that the considered metrics do not capture the effect of color on quality as humans do.

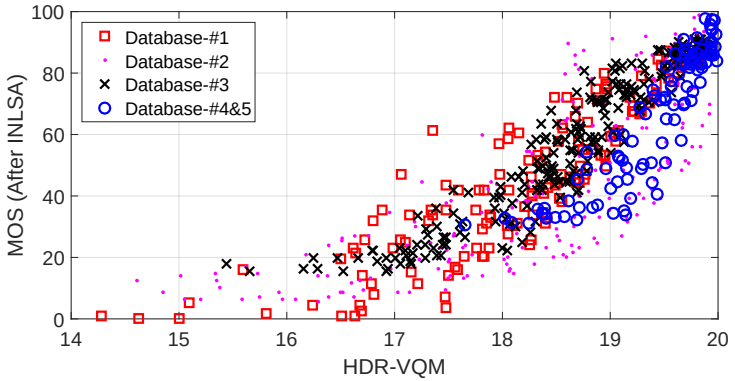
As viewers were rating very different distortions with respect to the other databases, which contain similar kinds of visual impairments, Database #2 is very challenging for all the quality metrics we considered in this work. Therefore, in order to provide a complete overview of the performance of HDR fidelity metrics, in the following, we report results both with and without Database #2 in the evaluations.

4.3 Analysis of Objective Quality Metrics

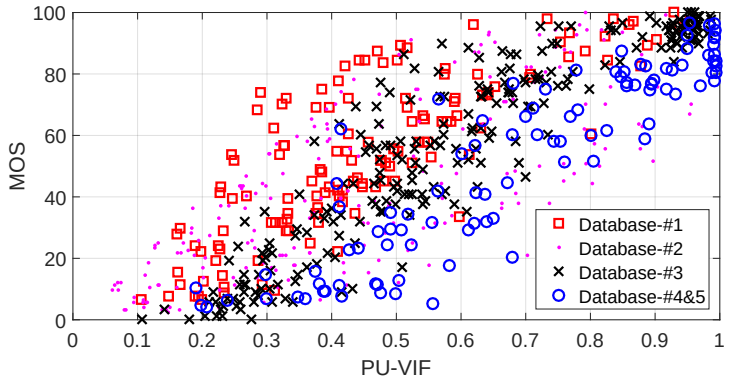
After the alignment of MOS values of the databases, we obtain an image data set consisting of 690 (or 480 images if Database #2 is excluded) images compressed using different image compression methods such as JPEG, JPEG-XT, and JPEG 2000. In this section, we provide a thorough analysis of the performance of several HDR image fidelity metrics. The performance of these quality metrics was evaluated both from the point of view of prediction accuracy and of their ability to tell whether two images are actually perceived as being of different quality.



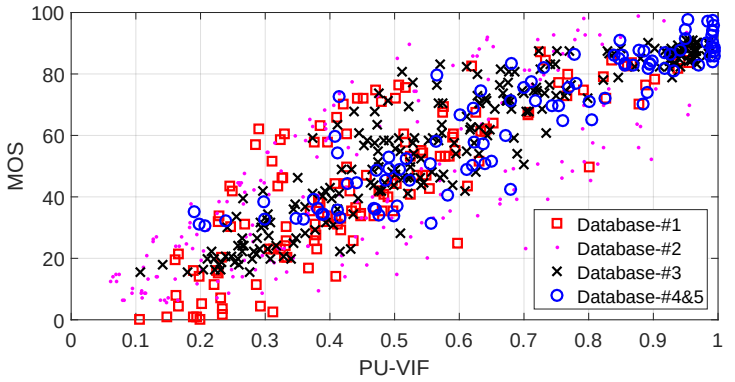
(a) HDR-VQM - Before INLSA (RMSE = 44.10)



(b) HDR-VQM - After INLSA (RMSE = 42.09)



(c) PU-VIF - Before INLSA (RMSE = 59.27)



(d) PU-VIF - After INLSA (RMSE = 57.96)

Figure 4.4 – Plots of MOS vs objective quality scores for HDR-VQM and PU-VIF before and after INLSA alignment. In order to compare the scatter plot quantitatively, the root mean squared error (RMSE) of the data is reported for each case.

4.3.1 Objective Quality Metrics under Consideration

We include in our evaluation a number of commonly used full-reference image quality metrics, including the mean squared error (MSE), peak signal to noise ratio (PSNR), structural similarity index (SSIM) [WBSS04], multi-scale SSIM (MSSIM) [WSB03], information fidelity criterion (IFC) [SBDV05], universal quality index (UQI) [WB02], and VIF [SB06]. In addition to those metrics, we consider HDR-VDP-2.2 [NMDSLC15], HDR-VQM [NDSLC15], additional full-reference metrics recently proposed for HDR video such as mPSNR, tPSNR, CIE ΔE 2000 [TS15], and spatial extension of CIE ΔE 2000 [ZW97] which is computed with S-CIELAB model.

The objective quality metrics under consideration can be grouped as following:

- **HDR-specific metrics:** *HDR-VDP-2.2* and *HDR-VQM* are recent full-reference quality metrics developed for HDR image and video, respectively. They model several phenomena that characterize the perception of HDR content and thus require some knowledge of viewing conditions (such as distance from the display, ambient luminance, etc.). The *mPSNR* is PSNR applied on an exposure bracket extracted from the HDR image, and then averaged across exposures.

- **Color difference metrics:** we use *CIE ΔE 2000* (denoted as CIE ΔE_{00}), which entails a color space conversion in order to get perceptually uniform color differences [LCR01], and its *spatial extension* [ZW97] (denoted as CIE ΔE^S_{00}). More sophisticated color appearance models were not considered in this study, as their use in quality assessment has been marginal so far.

- **SDR metrics applied after a transfer function:** SDR metrics such as *MSE*, *PSNR*, *VIF*, *SSIM*, *MSSIM*, *IFC*, and *UQI*. To compute these SDR metrics we use:
  - Physical luminance of the scene directly, denoted with the prefix *Photometric-*,
  - Perceptually uniform [AMS08] encoded pixel values, denoted with the prefix *PU-*,
  - Logarithmic coded pixel values, denoted with the prefix *Log-*, or
  - Perceptually quantized [MND12, SMP14] pixel values. For this case, only *tPSNR-YUV* was considered as in HDRtools [TS15].

Calculation of the Objective Quality Metrics

In order to calculate quality metrics, we first scaled pixel values to the range of luminance emitted by the HDR displays used in each subjective experiments. This is especially important for those metrics such as HDR-VDP 2.2 which rely on physical luminance. In order to compute these values, we converted HDR pixels into luminance emitted by a hypothetical HDR display, assuming it has a linear response between the minimum and

maximum luminance of the display. As all the experiments use the same display (i.e. SIM2 HDR47E S 4K), we selected the same parameters for all experiments, i.e., 0.03 cd/m^2 and 4250 cd/m^2 for minimum an maximum luminance, respectively.

Although the emitted luminance on HDR displays depends on many factors and is not exactly a linear function of input pixel values, we found in our previous work that it is adequately close to linear [ZVD16] and from a practical point of view, this simple linear assumption is equivalent to more sophisticated luminance estimation techniques which require a detailed knowledge of the reproduction device [VDSLD14]. Thus, the pixel values were multiplied with the luminance efficacy constant 179 [VDSLD14] and then were clipped below the minimum or above the maximum luminance values, generating a *linear model* for the HDR display used.

We used the publicly available implementation of these metrics, i.e., HDR-VDP-2.2.1 available at `http://sourceforge.net/projects/hdrvdp/files/hdrvdp/` , HDR-VQM available at `https://sites.google.com/site/narwariam/hdr-vqm` , HDRtools version 0.4 [TS15] developed within MPEG, the MeTriX MuX library for Matlab, available at `http://ollie-imac.cs.northwestern.edu/~ollie/GMM/code/metrix_mux/` .

4.3.2 Statistical Analysis

The performance of the aforementioned full-reference quality metrics was evaluated in terms of *prediction accuracy*, *prediction monotonicity*, and *prediction consistency* [DS12]. For prediction accuracy, Pearson correlation coefficient (PCC), and root mean squared error (RMSE) were computed. Spearman rank-order correlation coefficient (SROCC) was used to find the prediction monotonicity, and outlier ratio (OR) was calculated to determine the prediction consistency. These performance metrics were computed after a non-linear regression performed on objective quality metric results using a logistic function, as described in the final report of VQEG FR Phase I [RLC+00]. This logistic function is given in Equation 4.4:

$$Y_i = \beta_2 + \frac{\beta_1 - \beta_2}{1 + e^{-(\frac{X_i - \beta_3}{|\beta_4|})}}, \tag{4.4}$$

where X_i is the objective score for the i^{th} distorted image, and Y_i is the mapped objective score. It minimizes the least-square error between the MOS values and the objective results. This fitting was done using the `nlinfit` function of Matlab to find optimal β parameters for each objective quality metric. After fitting, the performance scores were computed using the mapped objective results, Y_i, and MOS values.

The results of these performance indices (SROCC, PCC, RMSE, and OR) were computed for each database separately, as well as considering all of the data together. The results are reported in Tables 4.3-4.6. The aligned data scores are denoted as **"Combined"**, and **"Except Database #2"** for the data aligned excluding Database #2 as explained in Section 4.2.

Table 4.3 – Pearson Correlation Coefficient (PCC) Results for Each Database and for Aligned Data

Metric	Database #1	Database #2	Database #3	Database #4 & 5	Combined	Except Database #2
Photometric-MSE	0.4153	0.1444	0.7080	0.5095	0.3817	0.6876
Photometric-PSNR	0.4292	0.2564	0.7132	0.5594	0.5123	0.6511
Photometric-SSIM	0.4041	0.3583	0.8655	0.6708	0.6392	0.7397
Photometric-IFC	0.7795	0.8234	0.9183	0.8195	0.8296	0.7762
Photometric-UQI	0.8090	0.8208	0.8846	0.7876	0.8414	0.7967
Photometric-VIF	0.7489	0.5076	0.8666	0.6144	0.6224	0.8297
PU-MSE	0.5146	0.3309	0.8559	0.8024	0.6316	0.7836
PU-PSNR	0.5506	0.3269	0.8606	0.8009	0.6314	0.7889
PU-SSIM	0.8178	0.7049	0.9532	0.9201	0.8316	0.8954
PU-IFC	0.8034	0.8422	0.9201	0.8566	0.8575	0.8201
PU-MSSIM	0.8567	0.7236	0.9564	0.9038	0.8480	0.9184
PU-UQI	0.8058	**0.8507**	0.8768	0.7777	0.8453	0.7925
PU-VIF	0.8212	0.7583	0.9349	0.9181	0.8624	0.8870
Log-MSE	0.4946	0.5314	0.8856	0.8820	0.6502	0.7579
Log-PSNR	0.5120	0.5624	0.8870	0.8819	0.6628	0.7575
Log-SSIM	0.6722	0.8035	0.9235	0.8255	0.7971	0.8023
Log-IFC	0.8224	0.8366	0.9167	0.8551	0.8603	0.8318
Log-UQI	0.8197	0.8268	0.8786	0.7830	0.8388	0.7933
Log-VIF	0.1858	0.6202	0.8354	0.7065	0.4803	0.5180
HDR-VDP-2.2 Q	**0.9127**	0.5482	0.9531	**0.9408**	0.7498	0.9171
HDR-VQM	0.8936	0.7932	**0.9612**	0.9332	**0.8759**	**0.9460**
mPSNR	0.5938	0.6564	0.8593	0.8587	0.7283	0.7888
tPSNR-YUV	0.5654	0.4524	0.8319	0.7789	0.6524	0.7735
$CIE\ \Delta E_{00}$	0.6165	0.2553	0.7889	0.6082	0.5042	0.7794
$CIE\ \Delta E_{00}^{S}$	0.6549	0.3331	0.8793	0.7322	0.5958	0.8154

Table 4.4 – Spearman Rank-Ordered Correlation Coefficient (SROCC) Results for Each Database and for Aligned Data

Metric	Database #1	Database #2	Database #3	Database #4 & 5	Combined	Except Database #2
Photometric-MSE	0.4294	0.1235	0.7227	0.5711	0.3423	0.7087
Photometric-PSNR	0.4341	0.2783	0.7183	0.5737	0.5006	0.6610
Photometric-SSIM	0.4436	0.3063	0.8792	0.6770	0.6274	0.7609
Photometric-IFC	0.7739	0.8254	0.9179	0.8109	0.8322	0.7811
Photometric-UQI	0.7859	0.8299	0.8686	0.8017	0.8420	0.7943
Photometric-VIF	0.7363	0.4915	0.8723	0.4864	0.5924	0.8163
PU-MSE	0.5147	0.2959	0.8617	0.8065	0.6159	0.7911
PU-PSNR	0.5147	0.2959	0.8617	0.8065	0.6157	0.7909
PU-SSIM	0.8099	0.7234	0.9503	0.9121	0.8419	0.9081
PU-IFC	0.7939	0.8433	0.9165	0.8489	0.8587	0.8226
PU-MSSIM	0.8394	0.7363	0.9517	0.8969	0.8500	0.9219
PU-UQI	0.7801	**0.8608**	0.8569	0.7932	0.8454	0.7895
PU-VIF	0.8030	0.7662	0.9306	0.9083	0.8620	0.8865
Log-MSE	0.4822	0.5843	0.8892	0.8719	0.6333	0.7458
Log-PSNR	0.4821	0.5843	0.8892	0.8710	0.6450	0.7466
Log-SSIM	0.6749	0.7869	0.9268	0.8179	0.8058	0.8122
Log-IFC	0.8080	0.8420	0.9140	0.8482	0.8610	0.8338
Log-UQI	0.7993	0.8232	0.8592	0.7960	0.8399	0.7894
Log-VIF	0.0278	0.5908	0.8385	0.6653	0.4996	0.4813
HDR-VDP-2.2 Q	**0.9077**	0.5727	0.9503	**0.9298**	0.7550	0.9268
HDR-VQM	0.8865	0.8126	**0.9572**	0.9193	**0.8733**	**0.9471**
mPSNR	0.5705	0.6496	0.8648	0.8521	0.7225	0.7948
tPSNR-YUV	0.5550	0.4342	0.8374	0.7901	0.6394	0.7782
$CIE\ \Delta E_{00}$	0.5929	0.2551	0.7824	0.5951	0.4883	0.7825
$CIE\ \Delta E_{00}^{S}$	0.6337	0.3096	0.8779	0.7430	0.5991	0.8208

Table 4.5 – Root Mean Squared Error (RMSE) Results for Each Database and for Aligned Data (Please note that, in order to have comparable results, RMSE values were calculated after all MOS values were scaled to the range of [0,100].)

Metric	Database #1	Database #2	Database #3	Database #4 & 5	Combined	Except Database #2
Photometric-MSE	23.409	27.459	22.163	25.684	23.723	17.833
Photometric-PSNR	23.242	26.791	22.000	24.742	22.043	18.641
Photometric-SSIM	23.537	25.907	15.719	22.138	19.738	16.527
Photometric-IFC	16.119	15.748	12.426	17.105	14.331	15.485
Photometric-UQI	15.125	15.850	14.635	18.392	13.871	14.842
Photometric-VIF	17.053	23.909	15.659	23.551	20.089	13.709
PU-MSE	22.063	26.187	16.232	17.814	19.898	15.259
PU-PSNR	21.481	26.225	15.984	17.874	19.904	15.091
PU-SSIM	14.808	19.683	9.489	11.688	14.254	10.934
PU-IFC	15.323	14.963	12.295	15.403	13.203	14.053
PU-MSSIM	13.273	19.153	9.165	12.775	13.605	9.719
PU-UQI	15.238	**14.586**	15.093	18.765	13.712	14.979
PU-VIF	14.683	18.089	11.142	11.828	12.994	11.342
Log-MSE	22.364	23.508	14.574	14.067	19.500	16.021
Log-PSNR	22.104	22.945	14.494	14.071	19.219	16.032
Log-SSIM	19.052	16.520	12.038	16.847	15.497	14.660
Log-IFC	14.639	15.201	12.540	15.477	13.083	13.633
Log-UQI	14.738	15.611	14.988	18.567	13.973	14.954
Log-VIF	25.284	21.769	17.249	21.126	22.513	21.007
HDR-VDP-2.2 Q	**10.517**	23.209	9.496	**10.120**	16.982	9.791
HDR-VQM	11.549	16.900	**8.657**	10.725	**12.383**	**7.962**
mPSNR	20.704	20.934	16.053	15.298	17.589	15.094
tPSNR-YUV	21.224	24.748	17.418	18.721	19.452	15.566
$CIE\ \Delta E_{00}$	20.261	26.830	19.285	23.694	22.165	15.388
$CIE\ \Delta E^{S}_{00}$	19.445	26.165	14.949	20.330	20.614	14.218

Table 4.6 – Outlier Ratio (OR) Results for Each Database and for Aligned Data

Metric	Database #1	Database #2	Database #3	Database #4 & 5	Combined	Except Database #2
Photometric-MSE	0.779	0.933	0.787	0.830	0.832	0.744
Photometric-PSNR	0.786	0.905	0.767	0.820	0.806	0.725
Photometric-SSIM	0.829	0.938	0.679	0.780	0.781	0.675
Photometric-IFC	0.743	0.871	0.546	0.610	0.654	0.621
Photometric-UQI	0.643	0.871	0.558	0.640	0.643	0.604
Photometric-VIF	0.729	0.948	0.617	0.800	0.797	0.621
PU-MSE	0.800	0.933	0.633	0.680	0.777	0.619
PU-PSNR	0.743	0.919	0.579	0.660	0.778	0.627
PU-SSIM	0.693	0.948	0.404	0.560	0.671	0.504
PU-IFC	0.707	0.886	0.500	0.610	0.633	0.592
PU-MSSIM	0.671	0.933	0.388	0.570	0.652	0.438
PU-UQI	0.621	0.848	0.583	0.680	0.645	0.602
PU-VIF	0.650	0.943	0.450	0.520	0.626	0.565
Log-MSE	0.771	0.924	0.592	0.570	0.716	0.642
Log-PSNR	0.750	0.919	0.588	0.580	0.755	0.658
Log-SSIM	0.771	0.876	0.525	0.570	0.733	0.585
Log-IFC	0.650	**0.833**	0.529	0.610	**0.625**	0.581
Log-UQI	0.643	0.843	0.579	0.630	0.645	0.606
Log-VIF	0.821	0.924	0.654	0.730	0.862	0.783
HDR-VDP-2.2 Q	**0.529**	0.938	**0.342**	**0.490**	0.741	0.471
HDR-VQM	0.636	0.890	0.392	0.530	0.638	**0.431**
mPSNR	0.750	0.895	0.667	0.610	0.722	0.635
tPSNR-YUV	0.721	0.952	0.625	0.670	0.771	0.637
$CIE\ \Delta E_{00}$	0.750	0.924	0.675	0.760	0.819	0.656
$CIE\ \Delta E^{S}_{00}$	0.700	0.933	0.613	0.710	0.796	0.615

These results show that the performance of many full-reference quality metrics may significantly vary from one database to another, due to the different characteristics of the test material and of the subjective evaluation procedure. In particular, Database #2 is the most challenging for all the considered metrics, due to its more complex distortion features, as discussed in Section 4.2. Despite the variations across databases, we can observe a consistent behavior for some metrics. Photometric-MSE is the worst correlated one, for all databases. This is expected as mean squared error is computed on photometric values, without any consideration of visual perception phenomena. On the other hand, HDR-VQM, HDR-VDP-2.2 Q, and PU-MSSIM are the best performing metrics, with the exception of Database #2.

When we analyze the objective metrics for each transfer function, we observe that Photometric-IFC is the best correlated and Photometric-MSE is the worst in the linear domain; Log-IFC is the best correlated and Log-VIF is the worst in the logarithmic domain. Among the objective metric results in PU domain, PU-MSSIM and PU-SSIM display high correlation coefficients, while PU-MSE is again the worst performer. Comparing the three transfer functions, PU is the most effective, as PU-MSSIM and PU-SSIM achieve performance very close to HDR-VDP-2.2 Q and HDR-VQM. In general, metrics which are based on MSE and PSNR (PU-MSE, Log-MSE, PU-PSNR, mPSNR, etc.) yield worse results compared to other metrics. Instead, more advanced SDR metrics such as IFC, UQI, SSIM, and MSSIM yield much better results. We also notice that mPSNR, tPSNR-YUV, and CIE ΔE 2000, which have been recently used in MPEG standardization activities, perform rather poorly in comparison to the others.

We also evaluate the significance of the difference between the considered performance indices, as proposed in ITU-T Recommendation P.1401 [ITU12c]. In this recommendation, three different tests were proposed to evaluate the significance of the difference of the correlation scores. These are the evaluation of the significance of the difference between correlation coefficients such as PCC and SROCC, evaluation of the significance of the difference between ORs, evaluation of the significance of the difference between RMSEs. The correlation scores calculated above were evaluated by these tests. The results are provided in Fig. 4.5 and Fig. 4.6 for "Combined" and "Except Database #2" cases respectively. The bars indicate statistical equivalence between the quality metrics. For example, there is not a statistically significant difference between HDR-VQM, PU-VIF, PU-IFC, and Log-IFC in terms of PCC, SROCC, OR, and RMSE.

We observe that the performance of HDR-VQM –along with PU-VIF, PU-IFC, and Log-IFC– in the combined database is significantly different from the others while PU-VIF, PU-IFC, Log-IFC and some other metrics have essentially equivalent performance across the combined databases. Although HDR-VDP-2.2 has a lower performance on combined dataset compared to its performance on individual databases, it is among the three most correlated metrics with HDR-VQM and PU-MSSIM on the case excluding Database #2. Interestingly, the HDR-VQM metric, which was designed to predict *video* fidelity, gives excellent results

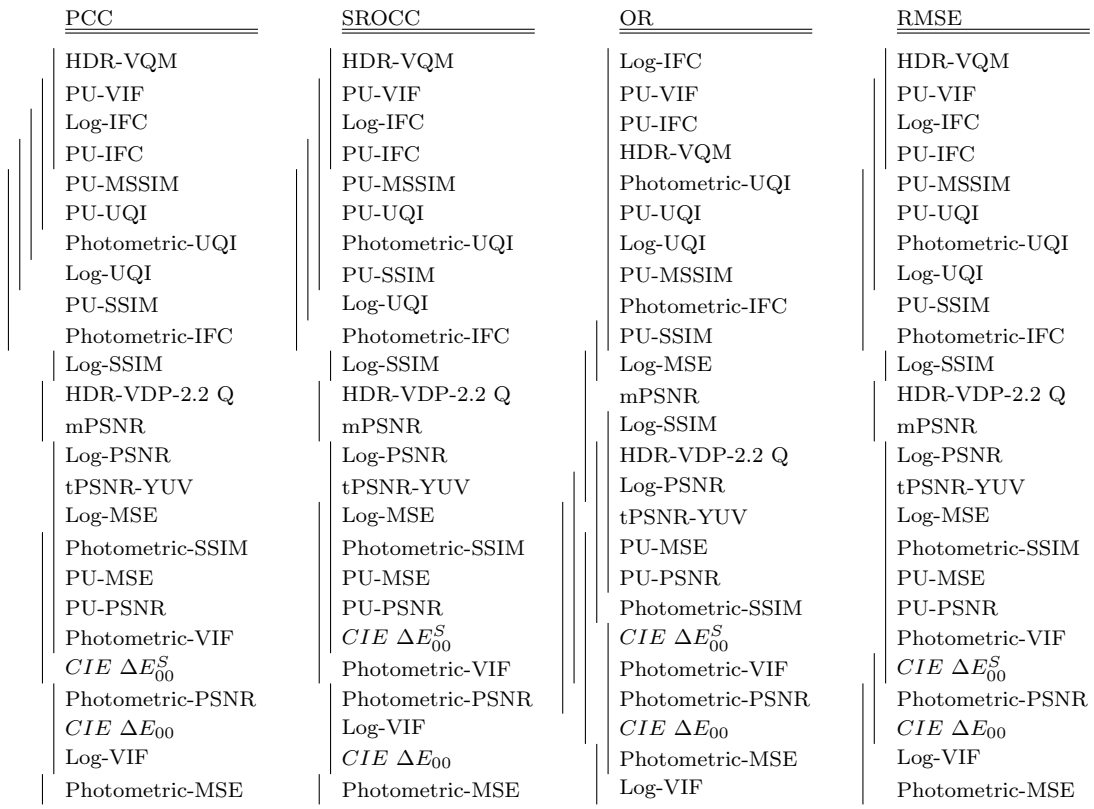


Figure 4.5 – Statistical analysis results for correlation indices for combined data according to ITU-T Recommendation P.1401 [ITU12c]. The bars signify statistical equivalence between the quality metrics if they have the same bar aligned with two quality metrics; e.g., there is not a statistically significant difference between HDR-VQM, PU-VIF, PU-IFC, and Log-IFC in terms of PCC, SROCC, OR, and RMSE.

also in the case of static images and is indeed more accurate on Database #2 than HDR-VDP-2.2. Furthermore, we notice that all metrics except CIE ΔE_{00} and CIE ΔE_{00}^{S} consider only luminance values. Although CIE ΔE_{00} and CIE ΔE_{00}^{S} have been found to be among the most relevant color difference metrics among others in a recent study [OJKP16], they have lower correlation scores when compared to luminance-only metrics. In fact, this result is not in disagreement with [OJKP16], which did not consider compression artifacts in the experiments, as the impact of those on image quality was deemed to be much stronger than color differences. Thus, our analysis confirms that luminance artifacts such as blocking, etc., play a dominant role in the formation of quality judgments, also in the case of HDR.

4.3.3 Discriminability Analysis

MOS values are *estimated* from a sample of human observers, i.e., they represent expected values of random variables (the perceived annoyance or quality). Moreover, the individual opinion scores are affected by several different factors. These factors may include many things that are known to affect the perception of human viewers such as the small physi-

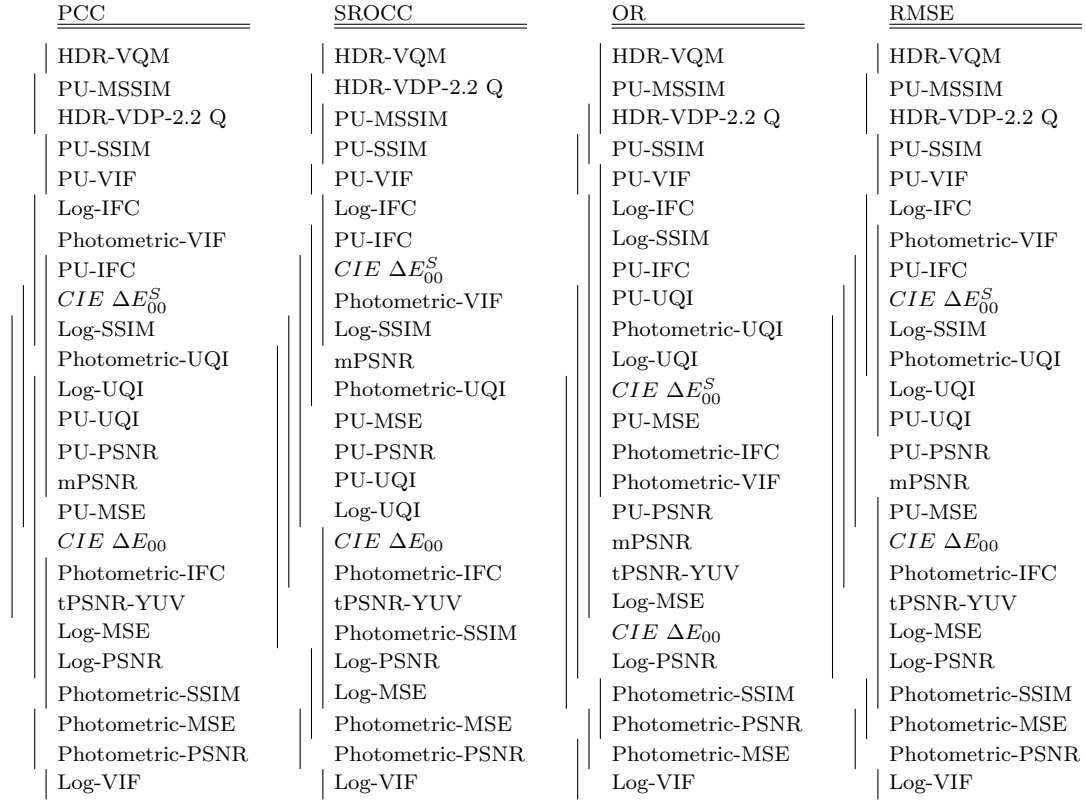


Figure 4.6 – Statistical analysis results for correlation indices for combined data excluding Database #2 according to ITU-T Recommendation P.1401 [ITU12c]. The bars signify statistical equivalence between the quality metrics if they have the same bar aligned with two quality metrics; e.g., There is a statistically significant difference between HDR-VQM and all the other metrics considered in terms of PCC, SROCC, and RMSE.

cal variations in test set-up, emotional variations of the viewers like mood and previous experiences, or the attention levels. Therefore, MOS values are as well random variables which are known with some uncertainty, which is typically represented by their confidence intervals [ITU12b]. As a result, different MOS values could correspond to the same underlying distribution of subjective scores and two images with different MOS might indeed have the same visual quality in practice (with confidence level). The performance scores considered in Section 4.3.2 assume instead that MOS values are deterministically known, and that the goal of full-reference quality metrics is to predict them as precisely as possible, without taking into account whether two different subjective scores do actually correspond to different quality. Therefore, in the following, we consider another evaluation approach, which aims at assessing if a full-reference objective quality metric is able to discriminate whether two images have significantly different subjective quality.

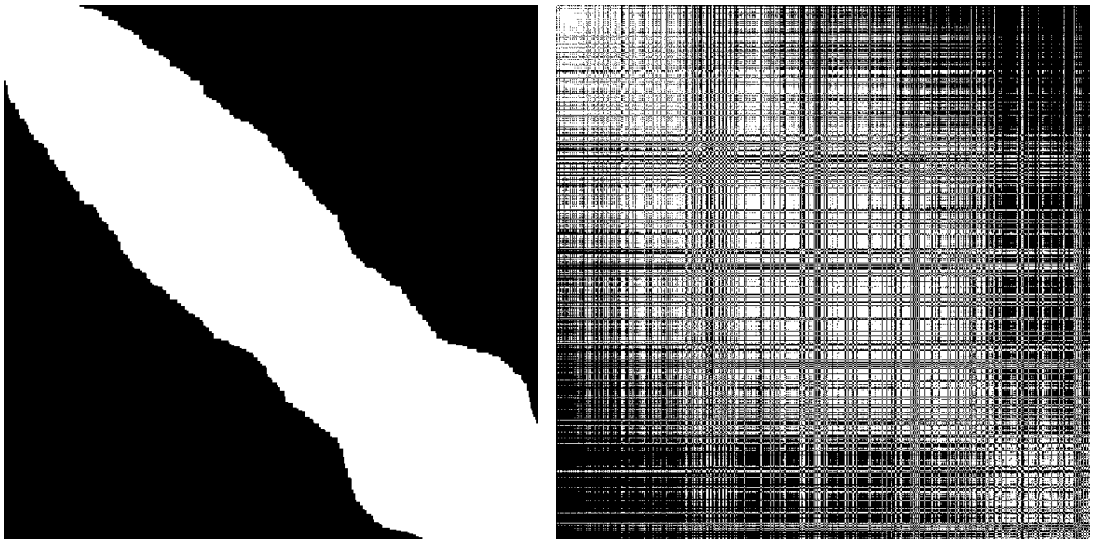
The intrinsic variability of MOS scores is not a completely new problem, and several approaches have been proposed in the literature to take this into account while evaluating objective metrics. Brill et al. [BLC+04] introduced the concept of *resolving power* of an objective metric, which indicates the minimum difference in the output of a quality

prediction algorithm such that at least $p\%$ of viewers (where generally $p = 95\%$) would observe a difference in quality between two images. This approach was also standardized in ITU Recommendation J.149 [ITU04b], and used in subsequent work [PW08, Bar09, HŘE15, NVH16]. Nevertheless, this technique has a number of disadvantages. Resolving power is computed after transforming MOS to a common scale, which requires applying a fitting function; however, the fitting problem could be ill-posed in some circumstances, yielding incorrect results. Also, the resolving power in the common scale corresponds to a variable metric resolution in the original scale, which makes it difficult to interpret. Moreover, it is not always possible to fix the level of significance p to be the same for different metrics, as there could be cases when the percentage of observers seeing a difference between image qualities is lower than p for any metric difference values. Finally, the results of this approach are generally evaluated in a qualitative manner, e.g., by considering how the number of correct decisions, false rankings, false differentiations, etc., vary as a function of objective metric differences [BLC+04, HŘE15]; conversely, a compact, quantitative measure is desirable in order to fairly compare different metrics. Another approach to this problem has been recently proposed by Krasula et al. [KFLCK16]. In their paper, Krasula et al. find the accuracy of an objective image or video quality metric by transforming the problem into a classification problem. For this purpose, they find z-score of subjective scores and the difference of objective scores for each pair of stimuli, and then find the accuracy of the metric by calculating classification rates. Two analysis are conducted: different vs. similar, and better vs. worse. They also propose a method to determine the statistical significance of the results.

Due to the factors above limiting the effectiveness of resolving power, in this chapter, we propose an alternative approach in the *original* scale of the metric similar to what has been presented in Krasula et al. [KFLCK16], which enables to evaluate its discrimination power while avoiding the shortcomings discussed above. Despite the similarities, the implementation and the data processing steps of their work and the proposed algorithm are not the same. Therefore, we give the details of the proposed algorithm below in order to clarify differences.

The basic idea of the proposed method is to convert the classical *regression* problem of accurately predicting MOS values, into a *binary classification* (detection) problem [Kay98]. We denote the subjective (MOS) and objective quality of stimulus I by $S(I)$ and $O(I)$, respectively, for a certain objective quality metric. Given two stimuli I_i, I_j, we model the detection problem as one of choosing between the two hypotheses $\mathcal{H}_0$, i.e., there is no significant difference between the visual quality of I_i and I_j, and $\mathcal{H}_1$, i.e., I_i and I_j have significantly different visual quality. Formally:

$$\begin{aligned} &\mathcal{H}_0 : S(I_i) \cong S(I_j); \\ &\mathcal{H}_1 : S(I_j) \ncong S(I_j), \end{aligned} \tag{4.5}$$



(a) MOS equivalence matrix at 95% confidence level (b) HDR-VDP-2.2 Q estimated equivalence matrix (τ fixed for maximum accuracy)

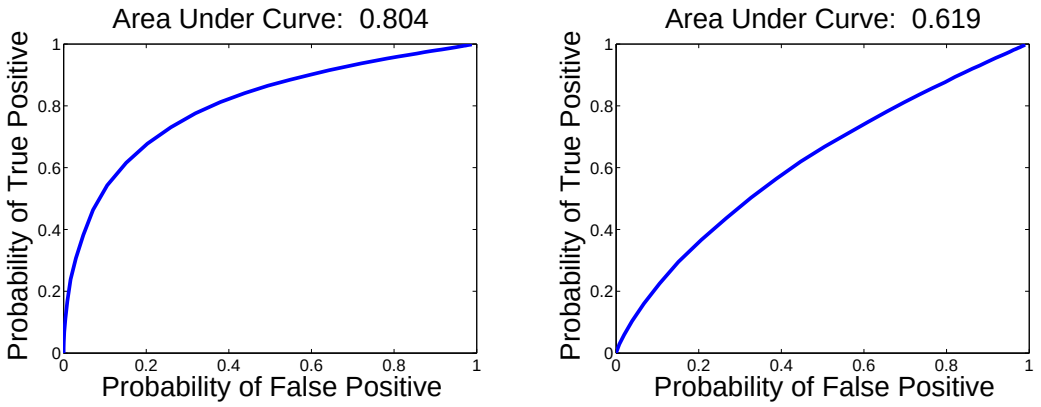
Figure 4.7 – Equivalence maps for the (sorted) combined database. White entries correspond to $S(I_i) \cong S(I_j)$, black to $S(I_i) \ncong S(I_j)$.

where we use $\cong$ (resp. $\ncong$) to indicate that the means of two populations of subjective scores (i.e., two MOS values) are the same (resp. different). Given a dataset of subjective scores, it is possible to apply a pairwise statistical test (e.g., a two-way *t-test* or *z-test*) to determine whether two MOSs are the same, at a given significance level. In our work, we employ a one-way analysis of variance (ANOVA), with Tukey's honestly significant difference criterion to account for the multiple comparison bias [HL87], as it is also stated as the ideal way to find statistical significance in [KFLCK16]. Figure 4.7.(a) shows the results of ANOVA on our combined database, thresholded at a confidence level of 95% (i.e., 5% significance). For the convenience of visualization, MOS values were sorted in ascending order before applying ANOVA. White entries represent MOS pairs which are statistically indistinguishable.

In order to decide between $\mathcal{H}_0$ and $\mathcal{H}_1$, similar to Krasula et al. [KFLCK16], we consider the simple test statistic $\Delta_{ij}^O = |O(I_i) - O(I_j)|$, i.e., we look at the difference between the objective scores for the two stimuli and compare it with a threshold τ, that is:

$$\text{Decide: } \begin{cases} \mathcal{H}_0 & \text{if } \Delta_{ij}^O \leq \tau \\ \mathcal{H}_1 & \text{otherwise.} \end{cases} \tag{4.6}$$

For a given value of τ, we can then label the set of stimuli as being equivalent or not, as shown in Figure 4.7.(b). The performance of the detector in (4.6) depends on the choice of τ. Intuitively, when τ is small (in the extreme case, equal to zero), all pairs of stimuli will be labeled as being of different quality. This maximizes the probability of detecting images



(a) HDR-VQM (b) PU-PSNR

Figure 4.8 – Example of ROC curves for two objective quality metrics, with corresponding area under the curve (AUC). Metrics with higher AUC enable better discrimination between the two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$.

that are actually of different quality; however, many pairs of stimuli having the same MOS will be misclassified as different. On the other hand, when τ is large (in the limit, when it tends to infinity), all pairs of stimuli will be labeled as being of the same quality. Thus, the opposite kind of error happens, i.e., many pairs of stimuli with different MOSs will be misclassified as being equivalent.

After finding equivalence matrices for both MOS values and objective quality metrics scores, the evaluation problem is converted to a binary classification problem, that is whether two images have the same quality. We call *true positive rate (TPR)* the ratio of images with different MOSs correctly classified as being of different quality, and *false positive rate (FPR)* the ratio of images with equal MOSs incorrectly classified as being of the different quality. By varying the value of τ, we can trace a Receiver Operating Characteristic (ROC) curve, which represents the TPR at a given value of FPR [Kay98]. An example of ROC curves for two objective metrics O is reported in Figure 4.8.

The area under the ROC curve (AUC) is higher when the overlap between the marginal distributions of Δ_{ij}^O under each hypothesis, that is, $p(\Delta_{ij}^O; \mathcal{H}_0)$ and $p(\Delta_{ij}^O; \mathcal{H}_1)$, is smaller. Therefore, the AUC is a measure of the *discrimination power* of an objective quality metric.

Table 4.7 reports the AUC values for the combined case and the combination without Database #2. In addition to the area under the ROC curve, we also compute the balanced classification accuracy, which is an extension of the conventional accuracy measure to unbalanced datasets, i.e., where the number of positive and negative samples is different [BOSB10]:

$$Acc = \frac{2 \times TP}{TP + FN} + \frac{2 \times TN}{TN + FP}. \tag{4.7}$$

In Table 4.7 we report the maximum classification accuracy, $Acc^* = \max_\tau Acc$, which characterizes the global detection performance, as well as the value of the detector threshold

Table 4.7 – Results of discriminability analysis: area under the ROC curve (AUC), threshold τ at 5% false positive rate, maximum classification accuracy. We report for comparison the fraction of Correct Decisions (CD) at 95% confidence level as proposed in [BLC$^+$04]. For CD, '–' indicates that the 95% confidence level cannot be achieved.

	Combined				**Except Database #2**			
Metric	AUC	$\tau_{.05}$	Acc*	CD [BLC$^+$04]	AUC	$\tau_{.05}$	Acc*	CD [BLC$^+$04]
Photometric-MSE	0.534	26718.659	0.531	–	0.648	30534.138	0.618	-
Photometric-PSNR	0.582	24.798	0.563	–	0.645	18.135	0.607	0.241
Photometric-SSIM	0.605	0.061	0.588	–	0.673	0.038	0.635	0.319
Photometric-IFC	0.713	6.610	0.664	0.384	0.682	7.554	0.633	0.343
Photometric-UQI	0.774	0.333	0.713	0.405	0.754	0.381	0.696	0.339
Photometric-VIF	0.602	0.730	0.582	0.203	0.703	0.730	0.644	0.414
PU-MSE	0.595	390.075	0.577	–	0.679	390.075	0.642	0.389
PU-PSNR	0.626	19.902	0.595	–	0.724	17.415	0.671	0.391
PU-SSIM	0.706	0.063	0.652	0.372	0.791	0.048	0.716	0.503
PU-IFC	0.730	6.081	0.677	0.450	0.707	6.896	0.653	0.419
PU-MSSIM	0.726	0.092	0.670	0.421	0.831	0.065	0.754	0.591
PU-UQI	0.777	0.318	0.716	0.413	0.753	0.364	0.695	0.337
PU-VIF	**0.787**	0.419	**0.725**	0.472	0.826	0.455	0.755	0.542
Log-MSE	0.557	1.423	0.549	0.248	0.554	1.090	0.551	0.401
Log-PSNR	0.638	27.767	0.595	0.255	0.674	27.767	0.625	0.406
Log-SSIM	0.681	0.169	0.626	0.341	0.701	0.135	0.654	0.396
Log-IFC	0.732	6.074	0.680	0.459	0.714	6.903	0.660	0.432
Log-UQI	0.779	0.324	0.718	0.395	0.754	0.371	0.696	0.326
Log-VIF	0.605	0.425	0.570	0.210	0.612	0.351	0.604	-
HDR-VDP-2.2 Q	0.683	23.955	0.626	0.269	0.836	20.962	0.744	0.592
HDR-VQM	0.786	1.962	0.723	**0.483**	**0.900**	1.503	**0.821**	**0.701**
mPSNR	0.678	13.557	0.638	0.277	0.721	13.164	0.668	0.399
tPSNR-YUV	0.629	17.041	0.596	0.168	0.707	17.041	0.653	0.381
$CIE\ \Delta E_{00}$	0.580	7.643	0.557	0.172	0.724	6.688	0.668	0.346
$CIE\ \Delta E^S_{00}$	0.609	6.718	0.583	0.189	0.744	5.878	0.690	0.382

at FPR = 5%, that is,

$$\tau_{.05} = \min\{\tau : p(\Delta^O_{ij} > \tau; \mathcal{H}_0) \leq 0.05\}, \tag{4.8}$$

which indicates the minimum value of τ in order to keep below 5% the probability of incorrectly classifying two stimuli as being of different quality. This latter measure provides somehow the *resolution* of an objective metric (with a 5% tolerance) in the original metric scale.

These results in Table 4.7 are complemented with the percentage of correct decisions (CD) in [BLC$^+$04], which is to be compared with Acc^*. Furthermore, we present the results of statistical significance evaluation of the reported AUC values according to the guidelines presented in Krasula et al. [KFLCK16]. The results of this statistical significance evaluation are presented in Fig. 4.9. The results show that HDR-VQM is the best performing metric, and PU-VIF and –in the case excluding Database #2– PU-MSSIM perform better than

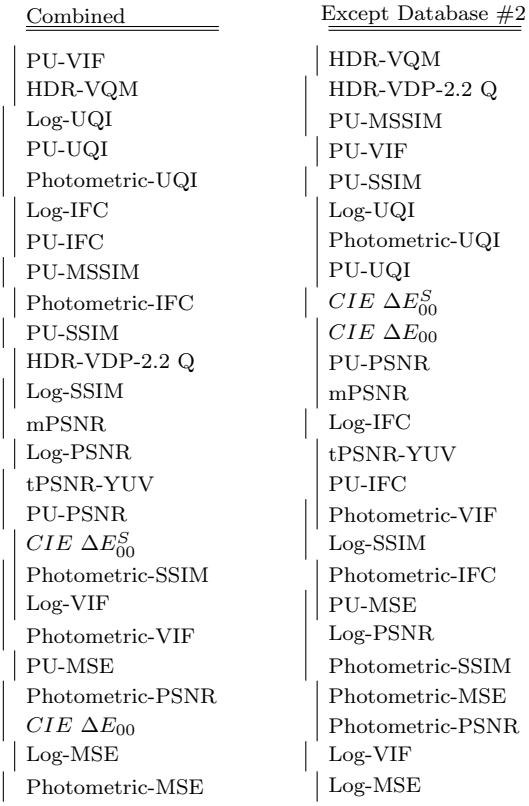


Figure 4.9 – Statistical analysis results for the discriminability analysis, according to the procedure described in Krasula et al. [KFLCK16]. The bars signify statistical equivalence between the quality metrics if they have the same bar aligned with two quality metrics. It can be said that among PU-UQI, Log-UQI, and Photometric-UQI, there is not any statistically significant difference. Whereas, there is a statistically significant difference between HDR-VQM and all the other metrics considered.

most of the considered metrics. Although its performance is reduced in the combined case, HDR-VDP-2.2 Q also is statistically better than most other metrics in the case excluding Database #2.

We notice that, in general, the values of CD are much lower than Acc^*. This is due to the fact that the method in [BLC$^+$04] not only aims at distinguishing whether two images have the same quality but also to determine which one has better quality. Thus the classification task is more difficult, as there are three classes – equivalent, better or worse – to label. Indeed, we observe a certain coherence between our approach and [BLC$^+$04], and with the statistical analysis in Section 4.3.2: the best performing metrics are HDR-VQM and those based on PU transfer function such as PU-MSSIM, PU-VIF, and PU-SSIM. Nevertheless, our analysis provides a better insight into the discrimination power of fidelity metrics compared to [BLC$^+$04], and gives practical guidelines on which should be the minimal differences between the objective scores of two images in order to claim that those have different visual quality. Finally, the fact that, even for the best performing metrics in terms of correlation with MOSs, maximum accuracy saturates at 0.8, suggesting

that there is still space for improving existing HDR objective quality measures, as far as discriminability (and not only prediction accuracy) is included in the evaluation of performance.

4.4 Discussion

An extensive evaluation of full-reference objective HDR image quality metrics was conducted, and its results were presented in this chapter. In order to conduct this evaluation, four different publicly available HDR image quality databases were collected, and a new HDR image quality database was created. These five databases were aligned using the INLSA algorithm in order to have consistent MOS values. In total, 690 compressed HDR images were evaluated using several full-reference HDR image quality assessment metrics.

The performance of these metrics was evaluated from two different aspects; statistical analysis and discriminability analysis. The statistical analysis considers the quality estimation as a regression problem and uses conventional statistical accuracy and monotonicity measures [DS12]. Discriminability analysis, on the other, focuses on the ability of objective metrics to discriminate whether two stimuli have the same perceived quality.

The analysis results show that recent metrics designed for HDR content, such as HDR-VQM and to some extent HDR-VDP-2.2, provide accurate predictions of MOSs. It is necessary to point out that these results are gathered using HDR image quality databases which have compression-like distortions. The results also confirm the findings of the previous work [VDSLD14, HBP$^+$15] as the results indicate that legacy SDR image quality metrics have a good prediction and discrimination performance, provided that a proper transformation such as PU encoding is done beforehand. This somehow suggests that the quality assessment problem for HDR image compression is similar to the case of SDR, if HDR pixels are properly preprocessed. Nonetheless, the performance results of the metrics reveal that none of the tested metrics provides highly reliable predictions, when all of the databases with heterogeneous characteristics are considered together (e.g. Database #2 in our experiments).

Except two of them, all of the considered metrics are computed on the luminance channel of the images. Interestingly, the non color-blind metrics, CIE ΔE_{00} and CIE ΔE_{00}^{S}, display poor performance in our evaluation. While other studies report different results in terms of correlation with MOSs [HRE16], we believe that a partial explanation for these results is that in the case of coding artifacts, the structural distortion (blocking, blur) in the luminance channel dominates the color differences, captured by CIE ΔE_{00} and CIE ΔE_{00}^{S} . The important aspect of color fidelity metrics for HDR content, however, is still little understood. The importance of color on the human perception of the quality is a prospective future research for HDR/WCG.

Finally, an alternative evaluation methodology is proposed in this chapter of the thesis. This evaluation methodology is based on the discriminability of a metric, and it

provides a complementary perspective on the performance of objective quality metrics. As the subjective experiments are done on a small sample set of observers, MOS values are probabilistic in their nature and are known with uncertainty. The proposed method recognizes the stochastic nature of MOS values, and it evaluates the objective quality metrics' ability to detect whether images have significantly different quality. The relevance of this alternative point of view is demonstrated by the amount of efforts to go beyond classical statistical measures such as correlation in the last decade, from the seminal work of Brill et al. [BLC+04] to the very recent work of Krasula et al. [KFLCK16], developed in parallel to our study.These analyses show that, even for metrics which can accurately predict MOS values, the rate of incorrect classifications is still quite high (20% or more). This suggests that novel and more performing object quality metrics could be designed, provided that new criteria such as discriminability are taken into account alongside the correlation indices used to find statistical accuracy.

Along with the results of this extensive evaluation, details of the proposed evaluation methodology was published in the journal article [ZVD17]. In order to support the research efforts on HDR image quality, the proposed database (merger of Database #4 and #5) of 100 HDR images have been made publicly-available over the Internet. These HDR images, along with their subjective quality scores, are available at `http://webpages.l2s.centralesupelec.fr/perso/giuseppe.valenzise/download.htm`.

Chapter 5

The Relation Between MOS and Pairwise Comparisons

Contents

Subjective quality assessment is used in many domains including psychology, medical applications, computer graphics, and multimedia. Regardless of the domain, it is regarded as a reliable method of quality assessment, and it is often employed to collect "ground-truth" quality scores.

Two of the main methods of subjective quality assessment for multimedia content are direct rating and ranking [ITU08, ITU12b]. Direct rating methods ask the observers to assign scores to observed stimuli. They may involve displaying a single stimulus (absolute category rating (ACR), single stimulus continuous quality evaluation (SSCQE)) or displaying two stimuli (double stimulus impairment scale (DSIS), double stimulus continuous quality evaluation (DSCQE)). Ranking methods ask the observers to compare two or more stimuli and order them according to their quality. The most commonly employed ranking method is pairwise comparisons (PC). Pairwise comparisons methodology was argued to be more suitable for collecting quality datasets because of the simplicity of the task and consistency of the results [PLZ+09, PJI+15]. The works of Ponomarenko et al. [PLZ+09, PJI+15], however, did not consider an important step in the analysis of pairwise comparisons data,

which is *scaling* pairs of comparisons onto an interval quality scale. Here, we analyze the importance of this step and demonstrate how it enables yielding a unified quality scale between rating and ranking methods.

In the previous chapter, the evaluation of FR HDR objective quality metrics shows that aligning subjective datasets is tricky and not straightforward. The MOS values are susceptible to many different factors such as the environmental factors of the subjective test, the instructions given to the subject in the training session before the experiment, the experimenters' understanding of quality, the range (or strength) of distortions applied to the stimuli, and even the mood of the subjects. Thus, MOS values may have a very different meaning and scale, depending on how they are collected. The objective alignment we used in the previous chapter, INLSA [PW03b], is valid under some assumptions. It is important to understand whether there is a more general method for alignment. How can we find a more robust method to align quality databases? What is a good measure to use while comparing the quality of two different stimuli? In order to answer these questions, we should first understand what causes the variance in perceived quality, and we should be able to reduce the variance.

In general, most of the subjective quality assessment studies use direct rating methods. The three mostly used direct rating methodologies are ACR [PPLC08b, GSI10, GDSE10, BPLC$^+$11], absolute category rating with hidden reference (ACR-HR) [SSB06, DSNT$^+$09, OZW10, SSBC10a, SSBC10b], and DSIS [LCA05, KHR$^+$15, HRE16]. For these methodologies, human observers evaluate the stimuli considering all the other stimuli in their mind. In order to make proper quality judgments, they need to remember how they voted for other stimuli. This necessity essentially creates a quality scale in the minds of observers and makes it harder to compare and judge as the number of stimuli grows. On the other hand, basic ranking methods, e.g. pairwise comparisons, are generally much more straightforward and simple compared to direct rating methodologies. Especially in the case of pairwise comparisons, we can argue that comparing between two stimuli is truly simpler than comparing among many. In their work, Mantiuk et al. [MTM12] found that the forced-choice pairwise comparisons methodology was the most accurate, as well as the most time-efficient, among four methodologies compared: single-stimulus categorical rating (i.e. ACR-HR), double-stimulus categorical rating, forced-choice PC, and similarity judgments.

The PC experiment and the results of PC scaling are much less influenced by the human factors by their nature. Thus, it can be used as a "universal" scale which you can align your datasets to. In this chapter, we try to understand the relation between the direct rating –i.e. MOS– and ranking –i.e. PC scaling– results, and we compare the MOS values and the scaling results of the PC experimental data.

The vast majority of studies employing the pairwise comparisons method compare only the images depicting the same content, for example comparing different distortion levels applied to the same original image. This "apple-to-apple" comparison simplifies the observers' task, making results consistent within content. However, it also comes with

some limitations. On one hand, assessing and scaling each content independently makes it difficult to obtain scores that correctly capture quality differences between conditions *across different contents* on a common quality scale. On the other hand, pairwise comparisons capture only relative quality relations. Therefore, in order to assign an absolute value to such relative measurements, the experimenter needs to assume a fixed quality for a certain condition which is then used as the reference for the scaling. As a result, the scaling error accumulates as conditions get perceptually farther from the reference.

In this chapter, we also study the effect of adding cross-content comparisons, showing that this not only allows unifying the quality scale across content but also improves the accuracy of scaled quality scores significantly. In order to understand the effect of cross-content pairwise comparisons, four different experiments were conducted using pairwise comparisons and double stimulus impairment scale methodologies. There are three major findings of the study described in this chapter:

- There is a strong linear relation between the mean opinion scores (MOS) obtained by direct rating, and scaled PC results;
- The addition of cross-content comparisons to the traditional PC reduces error accumulation and increases accuracy when scaling PC results;
- Cross-content comparisons align the PC scaling results of different contents to a common quality scale, reducing content dependency.

For this study, we use the high dynamic range (HDR) video quality dataset, presented in Section 3.2. Detailed information on scaling, the video quality database used, and the results are presented in the following sections.

5.1 Scaling Pairwise Comparisons Data

The results of a pairwise comparison experiment can be gathered in a preference matrix, also known as a comparison matrix. Its elements contain the counts of how many times one condition is voted as better than the other. These preference matrices can be used to find a quality score for each condition using one of several *scaling* methods [BT52, Thu27, LDSE11, TG11].

Commonly, pairwise comparison experiments are described by either of the two models: Bradley-Terry model [BT52] or Thurstone's model [Thu27]. Bradley-Terry model finds the quality, or rating, of each stimulus which satisfies $\sum_{i=1}^{N} \pi_i = 1$ and $P(i > j) = \frac{\pi_i}{\pi_i + \pi_j}$, where N is the total number of stimuli and π_i is the quality of stimulus i. It assumes that the difference between the quality of two stimuli i and j, $\pi_i - \pi_j$, has a logistic distribution. Thurstone's model, on the other hand, assumes that people may have different opinions about each stimulus and the quality, or rating, of each stimulus can be estimated with a

Gaussian distribution. Thurstone [Thu27] considers five different cases which have different properties. The most commonly used case is Case V which assumes that each option has equal variance and equal (or zero) correlations.

Other scaling methods proposed are generally based on these two models. Lee et al. [LDSE11] proposed Paired Evaluation via Analysis of Reliability (PEAR) which is based on Bradley-Terry model. It computes the quality scores and their confidence intervals using the distribution of winning frequencies and ties. The scores are then found by maximizing the log-likelihood function. Tsukida and Gupta [TG11] compares several methods based on both Bradley-Terry and Thurstone's model, such as least-square estimation, maximum likelihood estimation, and maximum a posteriori estimation.

In this chapter, we use *pwcmp*, an open source software[1] for scaling pairwise comparison results [POM17]. As also described in Section 3.2.2, this software estimates the quality scores using a Bayesian method, which employs a maximum-likelihood-estimator to maximize the probability that the collected data explains the quality scores under the Thurstone Case V assumptions. It is robust against incomplete and unbalanced designs, and it can scale pairs which have a unanimous agreement. The preference probabilities are converted to quality scores considering that the probability of 0.75 (mid-point between random guess (0.5) and certainty (1)) maps to 1 just objectionable difference (JOD). The concept of JOD and its difference from JND is better explained in the 'JNDs and JODs' part below. The *pwcmp* software also computes the confidence intervals using bootstrapping. Due to the relative nature of the pairwise comparison experiment, JOD values are relative. Therefore, we always fix the undistorted reference image at 0, and the distorted stimuli have negative JOD values.

JNDs and JODs

The results of paired comparisons are typically scaled in *Just-Noticeable-Difference (JND)* units [Eng00, SF01]. Two stimuli are 1 JND apart if 50% of observers can see the difference between them. However, we believe that considering measured differences as "noticeable" leads to an incorrect interpretation of the experimental results. Let us take as an example the two distorted images shown in Figure 5.1: one image is distorted by noise, the other by blur. Both images are definitely noticeably different, and intuitively they should be more than 1 JND apart. However, the question we ask in an image quality experiment is not whether they are different but rather which one is closer to the perfect quality reference. Note that a reference image does not need to be shown to answer this question as we usually have a mental notion of how a high quality image should look. Therefore, the data we collect does not measure visual differences between images, but rather it measures image quality difference in relation to a perfect quality reference. For that reason, we describe this quality measure as *Just-Objectionable-Differences (JODs)* [AVS+17] rather than JNDs.

[1] *pwcmp* toolbox for scaling pairwise comparison data `https://github.com/mantiuk/pwcmp`

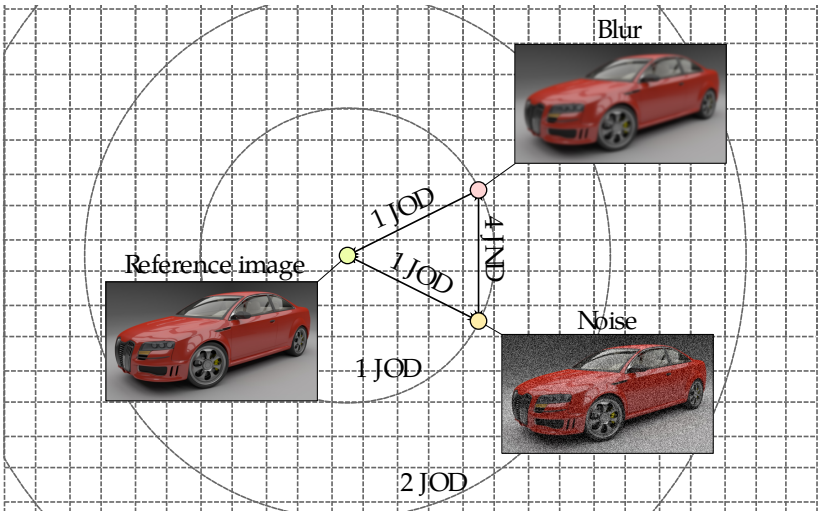


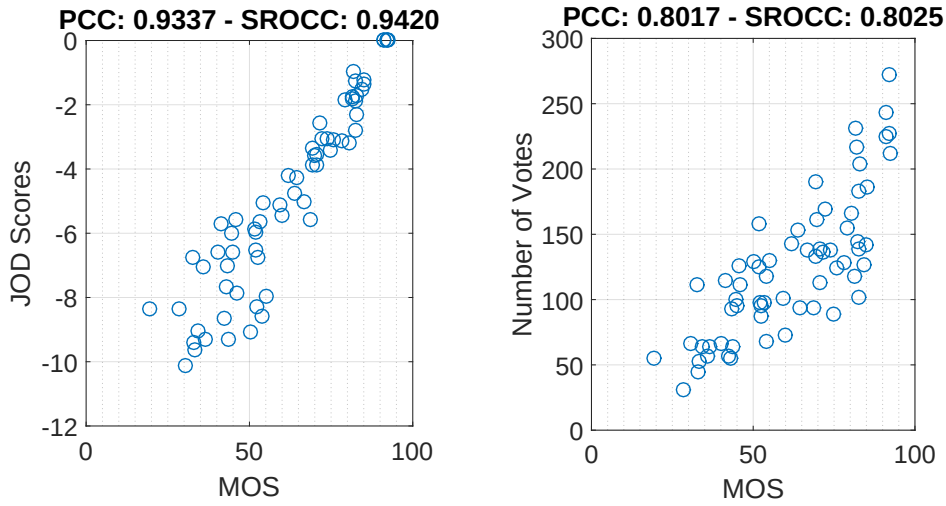
Figure 5.1 – Illustration of the difference between just-objectionable-differences (JODs) and just-noticeable-differences (JNDs). The image affected by blur and noise may appear to be similarly degraded in comparison to the reference image (the same JOD), but they are noticeably different and therefore several JNDs apart. The mapping between JODs and JNDs can be very complex and the relation shown in this plot is just for illustrative purposes.

Note that the measure of JOD is more similar to the quality expressed as a difference mean opinion score (DMOS) rather than to JNDs.

Is Scaling Necessary?

Scaling methods are not always used to convert a preference matrix into quality, and some researchers use alternative methods. In [PLZ+09] and [PJI+15], the quality values were estimated by counting the times one stimulus was preferred over another. However, this approach requires a complete experiment design, in which all pairs are compared, or a heuristic that would infer missing comparisons. In contrast to vote counts, scaling methods introduce an additional step of converting preference probabilities into an interval quality scale. In order to understand the difference between vote counts and the results of scaling, we compared both to the collected MOS values. We converted the results of the first pairwise comparison experiment to vote counts by counting how many times one condition was preferred over another.

To simulate how it was done in [PLZ+09, PJI+15], the missing comparisons were populated by the following operations: $V(A,C) = min(V(A,B), V(B,C))$ and $V(C,A) = min(V(B,A), V(C,B))$ where $V(x,y)$ is the number of votes in the preference matrix, provided that comparison of A and C was missing, but they were both compared to B. The resulting scores are presented in Figure 5.2. The plots show that PC scaling (in this case, JOD) scores are better correlated to MOS values compared to the quality estimates according to the number of votes. Considering this result, it can be claimed that using a



(a) PC scaling (JOD) scores vs. MOS values (b) Number of votes vs. MOS values

Figure 5.2 – Comparison of two different quality score estimation methods. The results of the first experiment is used to find the preference matrix. PC scaling done by *pwcmp* software (a) yields a better correlation to the MOS values than the quality score estimation via counting the number of votes (b).

scaling method yields results which are better correlated with MOS values.

Difference between MOS and PC Scaling

Although the mean opinion scores (MOS) are commonly used for the analysis of the subjective quality experiment results, there are several drawbacks of MOS values. The outcome of the MOS experiments strongly depends on the training procedure used to familiarize participants with the quality scale. Because of the differences in this training phase, measured scores are relative and are different for each session. The MOS values can result in different scales according to the instructor who does the training and also according to the experiment design. As it has been noticed in the Chapter 4, MOS values coming from different datasets may not be comparable with each other. While combining different datasets, an alignment step is often necessary; however, this is usually overlooked.

Pairwise comparison scaling in general, and JOD scaling used here in particular, does not require training and, in principle, should give consistent results for each session. Since pairwise comparison is a much more straightforward procedure, JOD values should be comparable between different datasets.

5.2 The Relation Between MOS and Pairwise Comparisons

In this section, we present the results of the subjective test conducted in order to understand the relation between MOS values and PC scaling results.

5.2.1 Details of the Subjective Experiments

In order to compare the MOS values to PC scaling results, two subjective tests were conducted. In these subjective tests, we used the HDR video quality dataset created in Section 3.2. This dataset consists of 60 compressed HDR videos. 5 original video sequences were compressed using HEVC Main 10 profile with 3 different color space conversions (RGB $\rightarrow$ Y'CbCr, ITP, and Ypu'v') and 4 different bitrates which are reported in Table 3.2. Each video sequence was 10 seconds long, composed of two identical 5-second long video segments played twice in succession.

The experiments were conducted in a quiet and dark room conforming to ITU Recommendations [ITU12b, ITU98]. The ambient illumination of the room was set to 2.154 lux, and the luminance of the screen when turned off was 0.03 cd/m^2. A calibrated HDR SIM2 HDR47 S 4K 47" display with 1920×1080 pixel resolution was used in its native built-in rendering mode. The subjects' distance from the screen was fixed to three heights of the display, with the observers' eyes positioned zero degrees horizontally and vertically from the center of the display [ITU98].

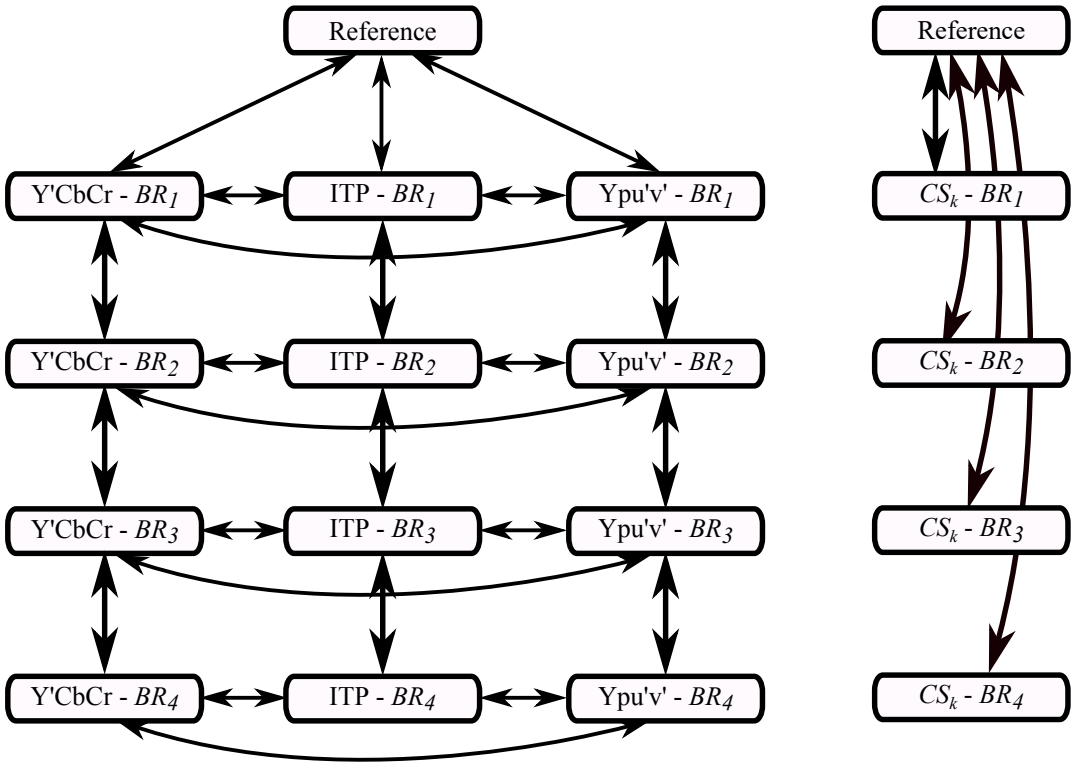
The conducted experiments share a common set of parameters in addition to those of the test room. The stimuli were presented as pairs with a side-by-side representation. A gray screen was shown before each pair for 2 seconds. The stimuli were presented, and the viewers were asked to vote. The duration of voting was not limited. A training session was conducted before each test, and the duration of the tests was less than 30 minutes including the training. All of the observers were screened and reported normal or corrected-to-normal visual acuity.

Pairwise Comparisons Experiment

The first experiment conducted was a pairwise comparisons experiment with incomplete design. In this experiment, a pair of videos with two consecutive bitrates from the same color space or with the same bitrate from two different color spaces was compared, as shown in Figure 5.3.(a). In order to keep the experiment short, other pair combinations were not included in this test. These comparisons were made only within the same content.

In total, 65 videos (5 contents $\times$ 3 color spaces $\times$ 4 bitrates + 5 reference sequences) were compared in 240 pairs (including mirrored versions). In order to keep each session under 30 minutes, the tests were conducted in two sessions. The order of the pairs was randomized for each session and the second session comprised of the mirrored versions of the videos of the first session. The duration of each session was approximately 30 minutes. There were 18 participants (14 men and 4 women) with an average age of 29.44. Further explanations on the experiment design are given in Section 3.2.

The confidence intervals get more precise (or narrower) with the increasing number of subjects [PPLC08b]. Therefore, in order to be able to compare the confidence intervals in a fair manner, it is desirable to have the same number of subjects. Since the DSIS experiment



(a) Standard Pairwise Comparisons - Incomplete design

(b) Double Stimulus Impairment Scale

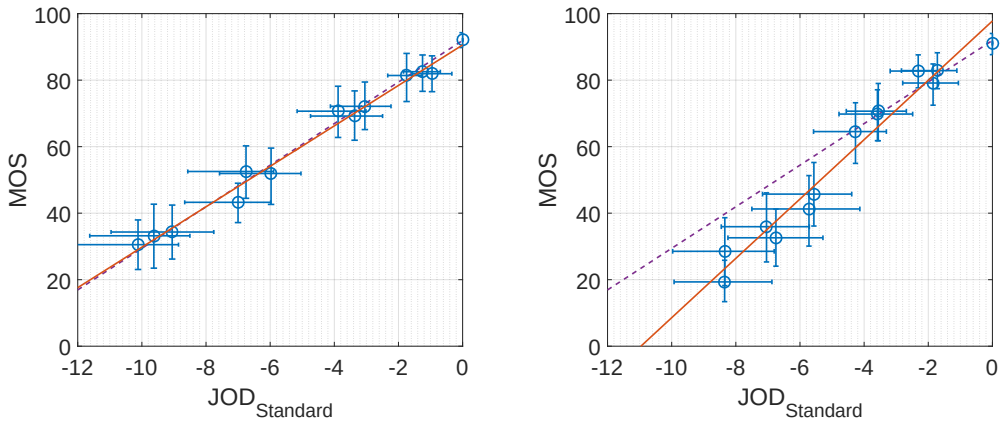
Figure 5.3 – Compared pairs for the (a) pairwise comparisons and (b) DSIS experiments. To avoid cluttering, comparisons for DSIS experiment are shown for each color space CS_k where k is the index of $CS = \{$Y'CbCr, ITP, Ypu'v'$\}$.

has 15 participants, in order to keep the number of the participants the same in all of the experiments, opinion scores of 3 random participants were removed from the results of this experiment.

Double Stimulus Impairment Scale Experiment

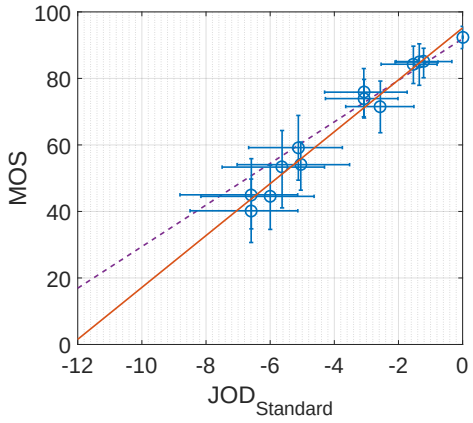
In order to analyze the pairwise comparisons scaling results and understand whether these scaling results are comparable to the quality scores, a second experiment was conducted following the double stimulus impairment scale (DSIS) methodology. In this second experiment, DSIS Variant I methodology with a side-by-side presentation was used, as in [HRE16]. A continuous scale ([0,100], 100 corresponding to "Imperceptible") was used instead of a categorical one (5 point impairment scale). All of the distorted videos were compared with the non-distorted reference video, as shown in Figure 5.3.(b).

A total of 120 pairs were compared (including mirrored versions). In order not to distract the viewers, left or right side was selected, and original videos were always placed on the selected side for each viewer. To avoid any contextual effects, the original videos

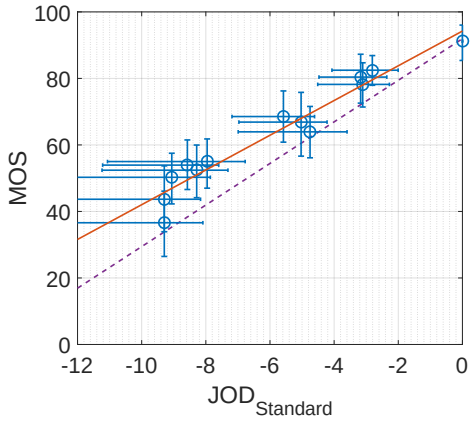


(a) Balloon

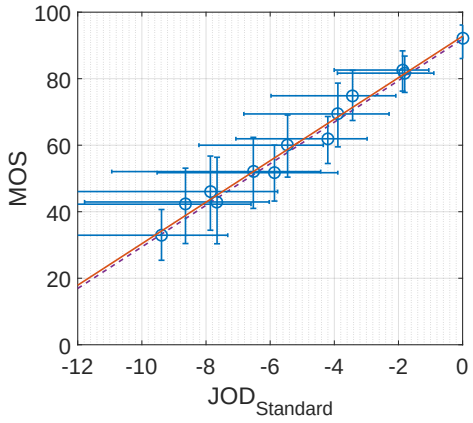
(b) Bistro



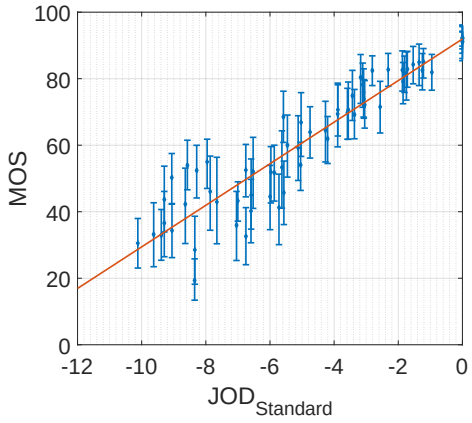
(c) Hurdles



(d) Market



(e) Starting



(f) All Together

Figure 5.4 – $JOD_{Standard}$ vs. MOS. Solid red line indicates the best linear fit to the data, and the dashed violet line indicates the best linear fit line of the case 'All Together'.

were presented on the left side of the display for half of the viewers and on the right side for the other half of the viewers. The duration of the DSIS tests was approximately 18 minutes. In total, 15 people (8 men and 7 women) with the mean age of 26.87 participated in the test.

5.2.2 Comparison of MOS and Pairwise Comparisons

The preference matrices of the PC experiments were found, and JOD scores were estimated using *pwcmp* software. For the DSIS experiment, the MOS values were calculated by taking the mean of opinion scores. Confidence intervals (CI), on the other hand, were calculated using bootstrapping in order to compare them to the CIs of JOD scores. The resulting JOD scores (denoted as $JOD_{Standard}$ to indicate that the standard pairwise comparisons methodology was used) were plotted against MOS values. The plots are shown in Figure 5.4.

The results show that there is a strong relation between MOS values and JOD scores. As presented in Figure 5.4, JOD scores and MOS values show almost linear behavior for all contents. This relation was also verified with PCC and SROCC computations. Reported in Table 5.1, PCC and SROCC values show that the relation is almost perfectly linear for each video sequence.

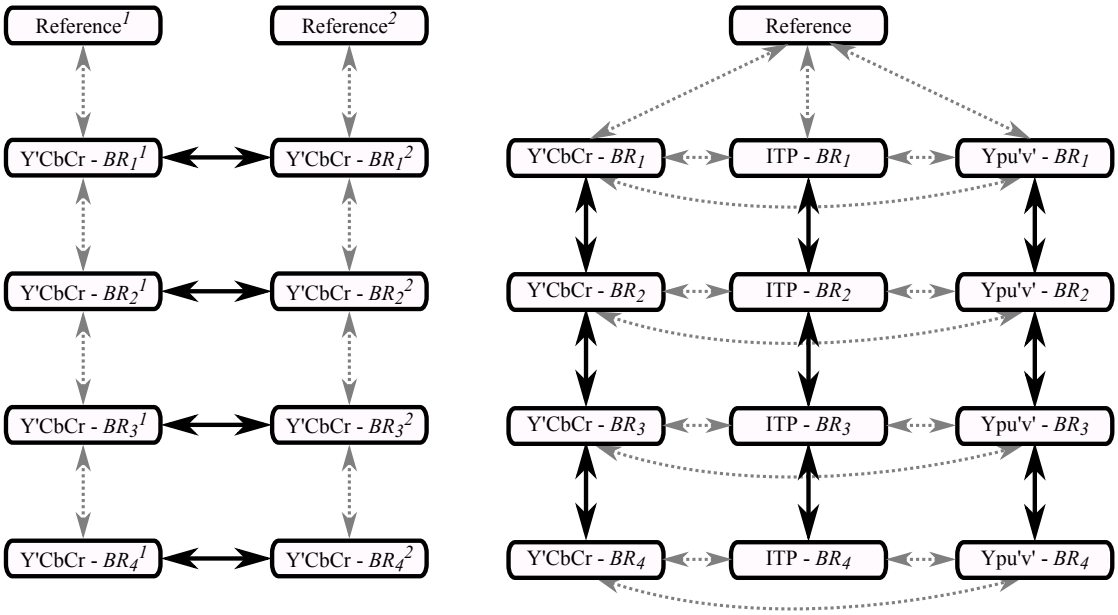
Table 5.1 – Linearity of the relation between MOS and JOD

Sequence	PCC	SROCC
Balloon	0.9936	0.9835
Bistro	0.9824	0.9890
Hurdles	0.9864	0.9670
Market	0.9696	0.9615
Starting	0.9897	0.9835
All Contents	0.9337	0.9420

We noticed that the MOS values have CIs close to uniform; however, the CIs of JOD values increase as absolute JOD values themselves increase. This was caused by the accumulation of the estimation errors, which results from comparing consecutive pairs. Note that running the full design, in which all pairs are compared, will not decrease such error to a large extent as the comparison for conditions that differ more than 2 JODs do not contribute much to the estimation.

5.3 Extending Pairwise Comparisons: Cross-Content Comparisons

The pairwise comparisons experiments are designed to compare conditions coming from the same content so that “apples” are compared to “apples”. Because only the reference condition is anchored and the quality of all other conditions is estimated from paired-



(a) Additional cross-content experiment (b) Additional same-content experiment

Figure 5.5 – Experiment design for two additional experiments. Selected additional pairs are shown with black arrows, where Reference^i is the reference (original) for video content i, BR_j^i is video content i compressed with the j-th bitrate ($j = 1$ is the highest bitrate). The pairs shown with dashed gray arrows are the pairs (shown in Figure 5.3.(a)) compared for the standard pairwise comparisons, as described in Section 5.2.1. To avoid cluttering, the comparisons between color spaces are not shown in subfigure (a).

relations, the estimation error is accumulated while moving away from the anchor. Instead, comparing "apples" to "oranges" may introduce new information, improve the scaling, and reduce cross-content variance.

In this section, we propose to extend the standard pairwise comparisons methodology to include cross-content comparisons. It is found that including cross-content comparisons improves the accuracy of PC scaling and reduces error accumulation.

5.3.1 Cross-Content Pairwise Comparisons Experiment

In addition to the subjective experiments described in the previous section, two additional experiments were conducted in order to analyze and understand the effects of cross-content pairwise comparisons. All of the variables except the selected pair of stimuli were kept the same. We were motivated to run such a cross-content comparison experiment after observing that such comparisons are indirectly performed in the DSIS methodology. When the viewers rate sequences, they judge the quality in relation to all other sequences they have seen, also the sequences presenting different content.

To keep the additional experiments short, for cross-content experiment, we paired videos with different contents and same bitrate, as shown in Figure 5.5.(a) using the comparisons shown with solid black arrows. The obtained results were combined with the

standard (same-content) pairwise comparison experiment results (shown with grey arrows in Figure 5.5) and scaled again using the same *pwcmp* software. The results are presented in the corresponding section below. The new JOD scores obtained from the combination of standard PC and cross-content experiment are called $JOD_{CrossContent}$.

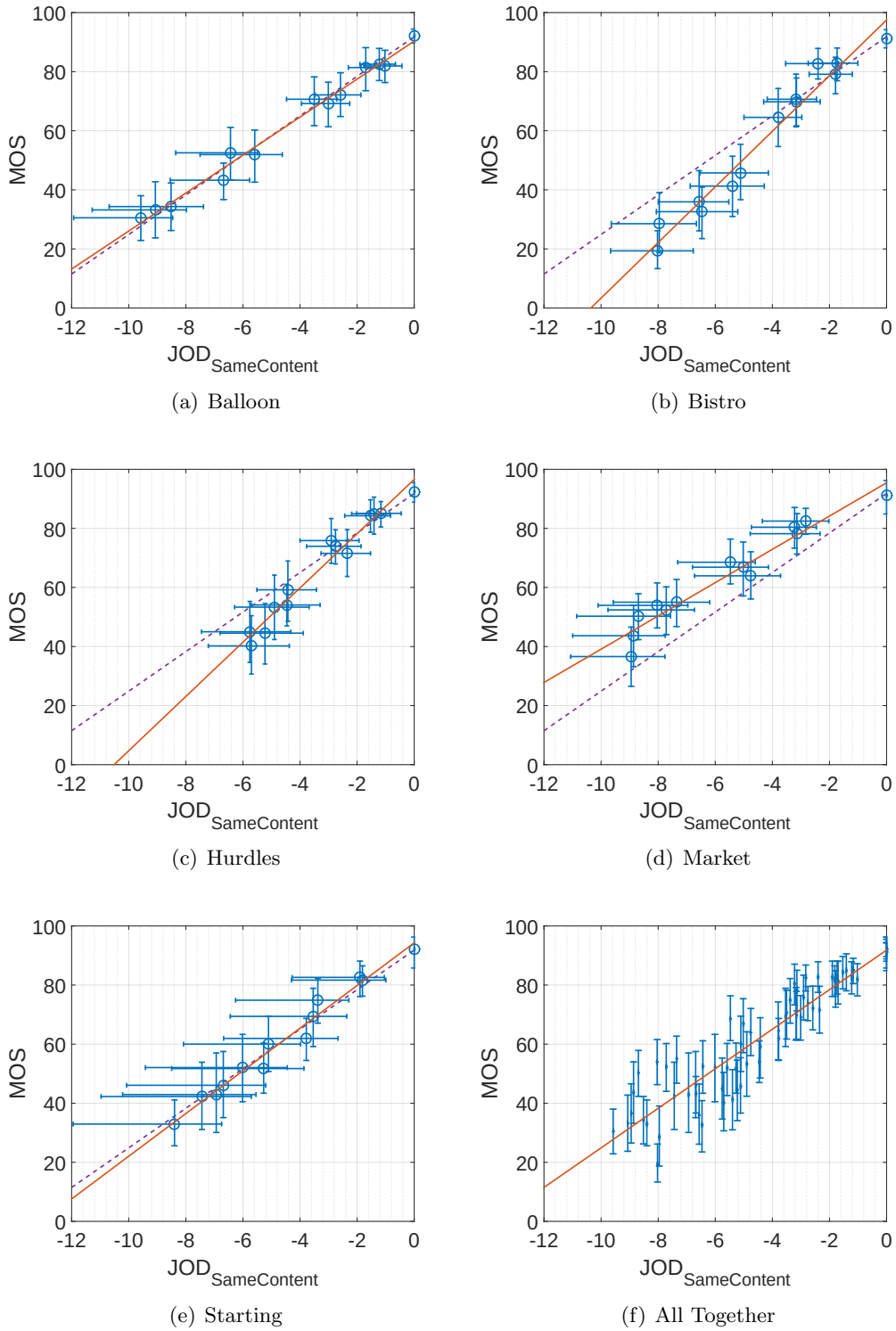
In order to conduct the test in one session and within 30 minutes, only videos encoded using Y'CbCr color space were compared, and the test set consisted of a total of 80 pairs (including mirrored versions). The duration of the tests was approximately 20 minutes. 15 people (8 men and 7 women) with an average age of 28 years took part to the test. Viewers were introduced the compression artifacts in the training part, and they were asked "Which one of the pairs have a better quality in terms of compression artifacts?".

In order to make a fair comparison in terms of the number of total comparisons, an additional same-content experiment was also conducted. To keep the additional number of pairs in a similar range, the test consisted of a total of 90 pairs (including mirrored versions). For this purpose, we selected pairs with consecutive bitrates and same color spaces, as shown with solid black arrows in Figure 5.5.(b). They are essentially additional observations of some of the pairs of the standard PC test described in the previous section. These pairs were compared by 15 people (8 men and 7 women) with an average age of 29. These additional same-content pairs were again combined with the standard PC experiment results and scaled again with *pwcmp* software. The JOD scores obtained from the combination of standard PC and additional same-content experiment are called $JOD_{SameContent}$.

5.3.2 Impact of Cross-Content Comparisons

The JOD values we use were found using three different sets of PC data. As described in the previous section, $JOD_{Standard}$ was found using the data acquired in the within-content PC experiment shown in Figure 5.3.(a). $JOD_{SameContent}$ was found using the combination of standard PC experiment results and additional same-content experiment results, and $JOD_{CrossContent}$ was found using the combination of standard PC experiment results and the cross-content experiment results.

Although the cross-content comparisons were made only for the videos with Y'CbCr color space, the JOD scores for all videos with all three color spaces are updated after re-scaling. The updated JOD values, $JOD_{SameContent}$ and $JOD_{CrossContent}$, were plotted against MOS and CI in Figures 5.6 and 5.7, respectively. The updated results show that the high CI values for high JOD scores are now significantly reduced with the addition of cross-content pairs. The slopes of the best linear fit change. Therefore, the relationship between JOD and MOS becomes much more linear and the correlation between JOD and MOS becomes much higher.

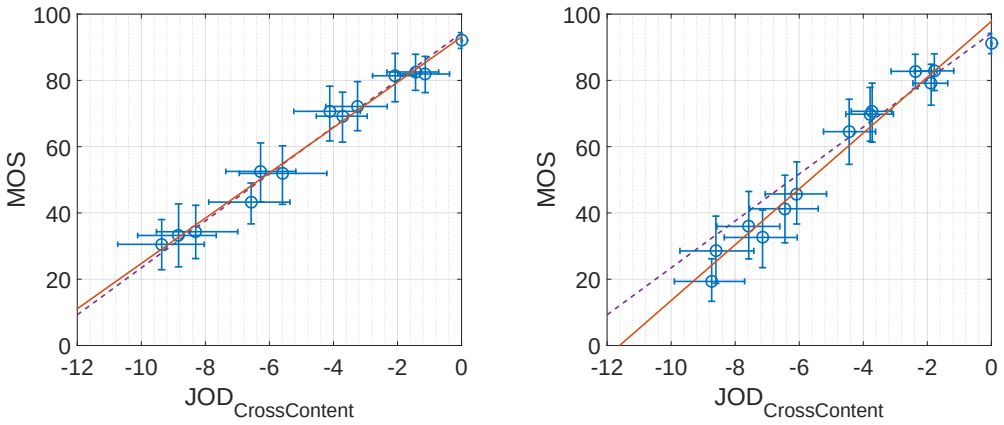


(a) Balloon (b) Bistro

(c) Hurdles (d) Market

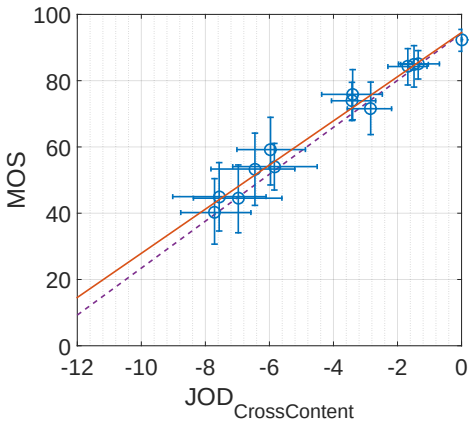
(e) Starting (f) All Together

Figure 5.6 – $JOD_{SameContent}$ vs. MOS. $JOD_{SameContent}$ is found using a combination of standard PC experiment (shown as in Figure 5.3.(a)) and additional same-content pairs as shown in Figure 5.5.(b). Solid red line indicates the best linear fit to the data, and the dashed violet line indicates the best linear fit line of the case 'All Together'.

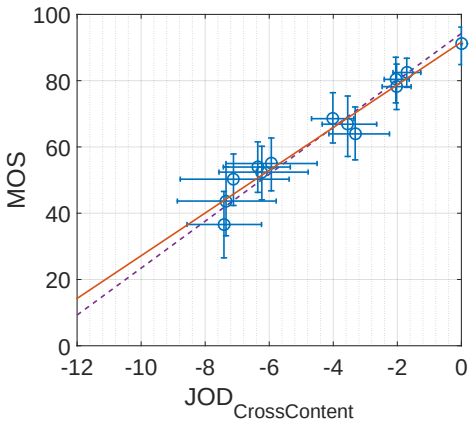


(a) Balloon

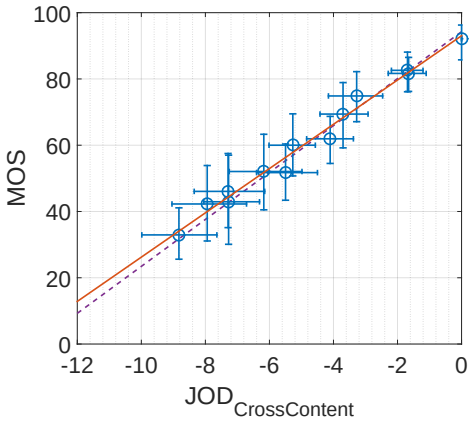
(b) Bistro



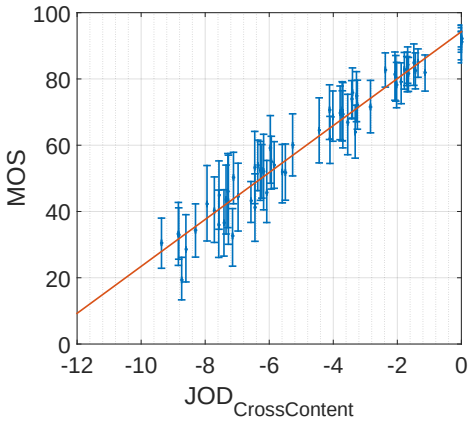
(c) Hurdles



(d) Market



(e) Starting



(f) All Together

Figure 5.7 – $JOD_{CrossContent}$ vs. MOS. Instead of only same-content pairs, a combination of same-content (shown as in Figure 5.3.(a)) and cross-content pairs were used to find $JOD_{CrossContent}$. Solid red line indicates the best linear fit to the data, and the dashed violet line indicates the best linear fit line of the case 'All Together'.

Table 5.2 – Linearity of the relation between MOS and JODs of two different cases: standard PC experiment with additional same-content pairs, JOD_{SC}, and the proposed PC experiment with same-content and cross-content pairs, JOD_{CC}.

Sequence	Case	PCC	SROCC	Slope	Std_{p2l}
Balloon	$JOD_{SameContent}$	0.9939	0.9835	6.43	0.3582
	$JOD_{CrossContent}$	0.9907	0.9835	6.84	0.4018
Bistro	$JOD_{SameContent}$	0.9836	0.9835	9.42	1.4544
	$JOD_{CrossContent}$	0.9860	0.9890	8.43	0.8499
Hurdles	$JOD_{SameContent}$	0.9833	0.9615	9.19	0.9639
	$JOD_{CrossContent}$	0.9882	0.9725	6.66	0.4885
Market	$JOD_{SameContent}$	0.9721	0.9670	5.63	1.5853
	$JOD_{CrossContent}$	0.9772	0.9670	6.45	0.5154
Starting	$JOD_{SameContent}$	0.9883	0.9890	7.23	0.4369
	$JOD_{CrossContent}$	0.9914	0.9835	6.68	0.3649
All Together	$JOD_{SameContent}$	0.9248	0.9324	6.69	1.0841
	$JOD_{CrossContent}$	0.9788	0.9733	7.08	0.5516

Linear Relationship Between MOS and PC Scaling

The linear behavior observed between the $JOD_{Standard}$ and MOS values holds for the cases of $JOD_{SameContent}$ and $JOD_{CrossContent}$, as well. Furthermore, the introduction of cross-content pairs increases the correlation and linearity of the relationship between JOD and MOS. The JOD scores become more linear after the combination of same-content and cross-content pairs, as can be seen in Figure 5.6.(f), Figure 5.7.(f), and in Table 5.2.

Reduced Content Dependency

In both Figure 5.6 and 5.7, the slopes of the best fitted line are found for each content. These slopes are reported in Table 5.2. In order to find the effect of the addition of cross-content pairs, the variance of these slopes was found. Variance of the slopes in the case of $JOD_{SameContent}$ was 2.7972 and in the case of $JOD_{CrossContent}$ was 0.6445. This significant reduction in the variance of the slopes implies that the best linear fit for each content is much closer, and there is less variance across different contents.

Another metric, Std_{p2l}, was computed for the points plotted in each sub-figure presented. It is calculated as:

$$Std_{p2l} = \sqrt{mean(d(P,l)^2)} \tag{5.1}$$

where $d(P,l)$ is the perpendicular distance from point P to line l. Std_{p2l} was computed for the best linear fit of the case 'All Together'. The best linear fit corresponds to the dashed violet line in the sub-figures (a)-(e). The results of Std_{p2l} are reported in Table 5.2.

It is clear that the addition of cross-content pairs decreases the variance of the slopes of the best fitted line for each content and Std_{p2l} as well, thus bringing JOD scores closer

Contents		$CI_{Standard}$	$CI_{SameContent}$	$CI_{CrossContent}$	$Ratio_{CC/SC}$
Balloon	BR_1	1.23	1.23	1.53	1.25
	BR_2	2.21	1.68	1.86	1.11
	BR_3	3.03	2.84	2.48	0.87
	BR_4	3.93	3.36	2.56	0.76
Bistro	BR_1	1.60	1.70	1.25	0.73
	BR_2	2.12	1.91	1.46	0.76
	BR_3	2.92	2.49	2.00	0.81
	BR_4	3.34	2.91	2.26	0.78
Hurdles	BR_1	1.45	1.50	1.12	0.75
	BR_2	2.31	1.90	1.55	0.82
	BR_3	3.12	2.36	2.46	1.04
	BR_4	3.43	2.96	2.62	0.89
Market	BR_1	2.12	2.35	0.85	0.36
	BR_2	3.05	2.80	1.63	0.58
	BR_3	4.32	3.18	2.57	0.81
	BR_4	4.73	3.28	2.94	0.90
Starting	BR_1	3.52	3.50	1.29	0.37
	BR_2	4.45	4.06	1.47	0.36
	BR_3	5.61	4.76	1.97	0.41
	BR_4	6.04	5.11	2.29	0.45

Table 5.3 – Average confidence intervals of the videos with different bitrates (BR_1 is the highest) for the considered experiments. The last column is the ratio of the CI of the combined PC data with additional cross-content pairs ($CI_{CrossContent}$, CI of $JOD_{CrossContent}$) to the CI of the combined PC data with additional same-content pairs ($CI_{SameContent}$, CI of $JOD_{SameContent}$). CI of standard PC experiment ($CI_{Standard}$, CI of $JOD_{Standard}$) are also reported for completeness.

on a common quality scale.

Reduced Error Accumulation

In order to analyze the change in CI, average CI values are reported in Table 5.3. Since the CI does not change with respect to the color space much, the CI values were averaged for the same bitrate. The last column of Table 5.3 shows that the CIs decrease for almost every case up to 60%, especially at higher bitrates where scaling error would instead accumulate in the standard PC. With cross-content comparisons, the CI size becomes more uniform across different levels of quality. Even though the total number of comparisons for $JOD_{CrossContent}$ is very close to those of $JOD_{SameContent}$, we observe a reduction in confidence intervals. These results show that the decrease in confidence intervals is not due to the increase in the total number of comparisons, but through the new information introduced by the comparison of cross-content pairs.

All the results indicate that the scaling of the pairwise comparisons data yields JOD scores that are highly correlated to MOS values acquired in the DSIS experiment. The

introduction of cross-content pairs make JOD more uniform and reduce the confidence intervals.

5.4 Discussion

Subjective quality assessment is considered as the most reliable approach for multimedia quality assessment. Although there are several different methodologies for measuring the subjective quality, pairwise comparisons methodology is considered to be one of the simplest, yet most precise, of all the well-known methodologies. The results of pairwise comparisons experiments can also be converted to numerical quality scores after a process called *scaling*.

In this chapter, we proposed to add cross-content comparisons in pairwise comparisons methodology to reduce the error accumulation that occurs during scaling. We present the results of three different experiments and analyze the effect of the proposed cross-content comparisons. Results show that the scaling performance improves and the confidence intervals reduce when cross-content pairs are introduced.

Pairwise comparisons methodology does not suffer from the quality scale difference as MOS experiments do, and JOD scores can be used as a more robust representation of subjective quality. This study serves as a preliminary study towards finding a more effective method to align datasets and develop novel hybrid methodologies where one can fuse MOS values and PC scaling results in order to have a better scale of quality scores. The results and findings of this study on the relation between MOS and pairwise comparisons was published in [ZHV+18].

Chapter 6

Conclusion and Future Work

Summary

In this thesis, we addressed some of the limitations and challenges of quality assessment in the context of high dynamic range image and video. Specifically, the goal of this thesis was to study the new conditions of HDR display technology and provide insight into the assessment and analysis of HDR video quality. For this purpose, we investigated three aspects of HDR quality assessment.

First, we analyzed the parameters affecting the subjective and objective HDR quality assessment in order to understand the influence of the new conditions introduced by HDR technology, and for this purpose, we developed an HDR frame rendering algorithm. In this part, we focused on the effects of display rendering (related to the brightness and contrast of the display) and color on HDR quality assessment.

Second, based on our findings, we evaluated the objective HDR image quality assessment methods using a 690-images dataset created by aligning MOS values of different databases, and we proposed a novel classification-based discriminability analysis method for the evaluation of objective metric performance.

Third, we compared pairwise comparisons scaling results to MOS values, with the intention of finding a common representation to align quality datasets and eliminating the need for the alignment step which was found to be necessary. Additionally, we proposed to include cross-content comparisons to pairwise comparisons methodology in order to reduce the cross-content variance and confidence intervals of PC scaling results.

The details of these three aspects are further discussed in the following sections.

Effects of the Display Rendering and Color

We first analyzed the characteristics of SIM2 HDR47 display, and we developed an HDR frame rendering algorithm. The developed algorithm estimates the LED and LCD values by calculating the convolution of the LED values with the point spread function and scaling

the backlight values iteratively. The experimental results show that the proposed algorithm both reproduces the intended luminance values and estimates the emitted luminance values accurately.

In order to understand the effects of display rendering, a subjective experiment was conducted by displaying the HDR images using the proposed rendering. The subjective results of this experiment were then compared to the results of another experiment in which the same HDR images were displayed using the built-in rendering of SIM2 display. Results show that there is no significant difference in subjective quality scores, as the MOS values do not change much. The reason for this was found after qualitative inspection of the images. Although the artifacts in darker regions became visible, artifacts in brighter regions became saturated and invisible to human eye. Overall, the subjective quality of the images stayed similar.

The comparison in the objective quality scores also showed no significant difference between the two rendering methods. The results show that a simple linear model is able to provide reliable results as if a detailed knowledge of the reproduction display were available, and it can be used to compute objective quality metrics. Moreover, the measurement results show that this simple model is actually very close to the real display response of the SIM2 display (see Figure 2.6.(a)).

In order to understand the effects of color, we created an HDR video compression dataset consisting of 60 videos in total, which were compressed with HEVC (HM 16.5) Main 10 Profile using three different color spaces: Y'CbCr, ITP, and Ypu'v'. These compressed videos were displayed using the built-in rendering of SIM2 display for a pairwise comparisons subjective experiment. The gathered PC data were scaled to find quality scores in terms of JOD. The results show that the influence of color space on coding performance is, in general, little and content dependent.

The quality of the compressed videos were also assessed using a luminance-only quality metric, i.e. HDR-VQM, and a color difference metric, i.e. ΔE_{00}. The results show that HDR-VQM can predict the general trend of the subjective quality scores, but it is not precise enough to predict absolute quality levels of JOD. This difference in quality levels motivates further studies in this direction.

The results of ΔE_{00} are not in agreement with either the subjective JOD scores or the objective HDR-VQM scores. Both this disagreement and the HDR-VQM's ability to predict the general trend of the subjective results indicate that the perceived quality for HDR video compression is dominated by the structural distortion caused by the changes in the luminance channel.

Objective Quality Metric Evaluation

We evaluated 25 objective quality metrics using five different HDR image quality databases consisting of 690 compressed HDR images in total. Due to the characteristics of the

experiments and the test material of each database, the MOS values can be in different ranges, and a similar level of impairment in the subjective scale may correspond to very different values of objective metrics. The MOS values of these considered databases had the same problem (see Figure 4.2), and we aligned the MOS values using INLSA algorithm prior to the computation of the objective quality metrics.

In addition to the statistical analysis techniques used, we proposed a novel classification-based discriminability analysis method. The significance of the aligned MOS values are found using multiple comparison test and are fed into the proposed method as ground-truth subjective quality. Then, the proposed method calculates the classification rates by sweeping for the objective quality score difference threshold, τ from 0 to the maximum, in order to find the performance indicators: area under curve (AUC) and the maximum balanced accuracy. Having high rate of incorrect classification hints that there may be still room for improvement, and the proposed method can be a useful indicator of performance in addition to the statistical analysis methods.

The objective quality estimations were compared to the MOS values, and the metrics were evaluated using both the proposed discriminability analysis and the statistical analysis methods. In addition to the numerical results, the proposed method and statistical analysis methods were analyzed for the significance of difference. The results indicate that although HDR-specific metrics yield the highest correlation scores, legacy SDR image quality metrics also have a good prediction and discrimination performance, provided that a proper transformation such as PU encoding is done beforehand. Moreover, the higher performance of luminance-only quality metrics compared to color difference metrics, ΔE_{00} and ΔE_{00}^{S}, supports the claim that HDR image and video compression is dominated by the structural distortion.

In addition to the results of the evaluation, the database consisting of 100 compressed HDR images, which was created as a merger of Database #4 and #5, has been made publicly available over the Internet in order to support the research efforts on HDR image quality assessment.

Comparison of Subjective Quality Scores

As the evaluation of the quality metrics show, the MOS values may be different for different databases, and this result affects not only subjective quality assessment but also objective quality assessment as the training and adjustments of objective quality metrics depend on the subjective quality scores. In order to find a common representation for different quality databases, we compared pairwise comparisons scaling results to MOS values, conducting a series of subjective experiments.

For the comparison of MOS values to JOD values, we conducted a DSIS subjective experiment with the same material used in the PC subjective experiment described in Section 3.2. The results show that there is an almost perfectly linear relationship between

MOS and JOD values.

In order to improve scaling performance and reduce cross-content variance, we proposed to include cross-content comparisons in pairwise comparisons methodology. In order to analyze the effect of adding cross-content pairs, the variance between the best-fit slopes and standard deviation of point to best-fit line distances were calculated. Results show that the inclusion of cross-content pairs reduces the confidence intervals of the PC scaling results (i.e. JOD values) and the cross-content variance of the relationship between JOD and MOS values. It also improves the overall scaling performance.

Since pairwise comparisons methodology is easier for subjects compared to the direct rating methods, it is expected to result in more reliable quality scores. Moreover, we showed that inclusion of cross-content pairs improve scaling performance. Presenting these results, this study serves as a preliminary study towards finding a more effective method to align datasets and develop novel hybrid methodologies where one can fuse MOS values and PC scaling results in order to have a better scale of quality scores.

Future Research Directions

Rapid commercialization of HDR/WCG technology and increasing volume of HDR content bring about new perspectives for future research. We believe that some aspects of the HDR/WCG technology need further investigations, and we describe a number of possible extensions of this thesis.

Estimation of Emitted Color

Although the proposed rendering method is able to estimate the emitted luminance accurately, it cannot estimate the emitted chrominance at the moment. Its output includes color information, but the color information is acquired simply by dividing the color HDR image to the backlight value.

For accurate chrominance estimation, the LCD panel response and color primaries of the SIM2 display should be better studied. This new rendering method which can estimate emitted color may be useful for possible subjective experiments related to color perception in HDR and wide color gamut studies.

Understanding Color Artifacts for HDR/WCG

For the case of compression, the structural distortions created by the differences in luminance channel are found to be dominant over the human perception of HDR image and video quality. However, in cases other than compression, changes in color may still influence the perceptual quality. Color artifacts can be created by several reasons such as color space or color gamut conversions and EOTF conversions. Hence, a wider range of color distortions may be studied to understand the effects of color artifacts in the sense of HDR/WCG. The

findings can be used to develop a more suitable color fidelity (or color difference) metric for HDR content.

Additionally, as we noticed from the results of Chapter 3 and Chapter 4, the color difference metrics are not able to predict the quality of compressed HDR content. The findings of the proposed color artifacts study can also be used for the development of a quality metric which takes color into account for the case of HDR image and video compression.

Evaluation of Objective Metrics

The evaluation of HDR image quality metrics in Chapter 4 has important results and conclusions, some of which can be extended for the case of video. In this thesis, we could not carry out such an evaluation for HDR video due to scarcity of publicly available HDR video quality databases at the time of this study. Therefore, a similar evaluation can be done for the case of HDR video, in order to take the temporal characteristics into account.

New technologies such as 4/8K and high frame rate can also be considered in combination with HDR/WCG technologies. This combination may introduce new challenges for objective quality assessment algorithms.

JOD as a Universal Subjective Quality Score

As we discussed previously, we believe that JOD can be a universal subjective quality score thanks to the easiness and robustness of its calculation. Although the initial results are promising, this claim needs to be validated with a larger set of data with various distortions and quality levels. Provided that it is validated, the JOD values can be used for aligning different databases, and this can improve both the evaluation and development of objective HDR quality metrics.

By its nature, the proposition of JOD as a universal subjective quality score is not limited to HDR quality assessment, and JOD can be used in almost all multimedia quality assessment applications.

Chapter 7

Publications

Journal articles

1. Emin Zerman, Giuseppe Valenzise, and Frédéric Dufaux, "An Extensive Performance Evaluation of Full-Reference HDR Image Quality Metrics", *Quality and User Experience*, volume 2, April 2017.

2. Yi Liu, Naty Sidaty, Wassim Hamidouche, Olivier Deforges, Giuseppe Valenzise, Emin Zerman, "An Adaptive Quantizer for High Dynamic Range Content: Application to Video Coding", *IEEE Transactions on Circuits and Systems for Video Technology*, volume PP, issue 99, December 2017.

Conference papers

1. Emin Zerman, Giuseppe Valenzise, Francesca De Simone, Francesco Banterle, Frédéric Dufaux, "Effects of Display Rendering on HDR Image Quality Assessment", *SPIE Optical Engineering+ Applications, Applications of Digital Image Processing XXXVIII*, San Diego, CA, USA, August 2015.

2. Emin Zerman, Giuseppe Valenzise, and Frédéric Dufaux, "A Dual Modulation Algorithm for Accurate Reproduction of High Dynamic Range Video", *IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, Bordeaux, France, July 2016.

3. Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, Rafał Mantiuk, Frédéric Dufaux, "Effect of Color Space on High Dynamic Range Video Compression Performance", *9th International Conference on Quality of Multimedia Experience (QoMEX)*, Erfurt, Germany, June 2017.

4. David Kane, Vedad Hulusic, Giuseppe Valenzise, Emin Zerman, Antione Grimaldi, Marcelo Bertalmio, "Subjects Prefer to View a Linear Image When Both the Image and the Display Have the Same Dynamic Range", *40th European Conference on Visual Perception ECVP 2017*, Berlin, Germany, June 2017.

5. Kutan Feyiz, Fatih Kamışlı, Emin Zerman, Giuseppe Valenzise, Alper Koz, Frédéric Dufaux, "Statistical Analysis and Directional Coding of Layer-based HDR Image Coding Residue", *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Luton, UK, October 2017.

6. Yi Liu, Naty Sidaty, Wassim Hamidouche, Olivier Deforges, Giuseppe Valenzise, Emin Zerman, "An Adaptive Perceptual Quantization Method for HDR Video Coding", *IEEE International Conference on Image Processing (ICIP)*, Beijing, China, September 2017.

7. David Kane, Antione Grimaldi, Emin Zerman, Marcelo Bertalmio, Vedad Hulusic, Giuseppe Valenzise, "The Preferred System Gamma is Primarily Determined by the Ratio of Dynamic Range of the Original Scene and the Displayed Image", *IS&T/SPIE Electronic Imaging, Human Vision and Electronic Imaging XXII*, San Francisco, California, USA, January 2018.

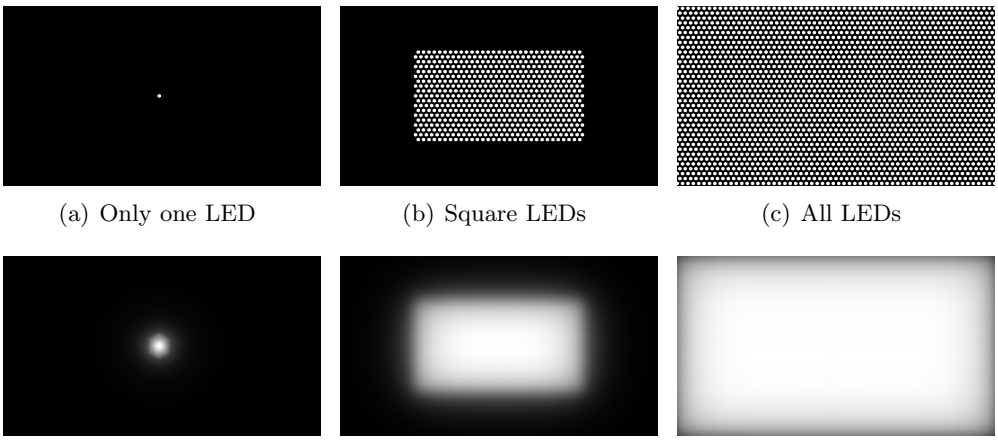
8. Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, Rafał Mantiuk, Frédéric Dufaux, "The Relation Between MOS and Pairwise Comparisons and the Importance of Cross-Content Comparisons", *IS&T/SPIE Electronic Imaging, Human Vision and Electronic Imaging XXII*, San Francisco, California, USA, January 2018.

Annex A

SIM2 Display Measurements

In order to understand the characteristics of the SIM2 HDR47E S 4K display used in our studies, detailed and extensive measurements were made. For the sake of reproducible research and in order to help other researchers, these measurements are reported in this chapter. These measurements were taken in a room which was sealed to block all external light. Konica Minolta LS-100 was used as a light meter. The focus of the light meter was set to 1-meter, and all of the measurements were taken at a 1-meter distance to the SIM2 display, perpendicularly.

The following sections explain the LED measurements, LCD measurements, and the measurements made to understand the point spread function and the "border effect".



(a) Only one LED (b) Square LEDs (c) All LEDs

(d) Luminance of one LED (e) Luminance of square LEDs (f) Luminance of all LEDs

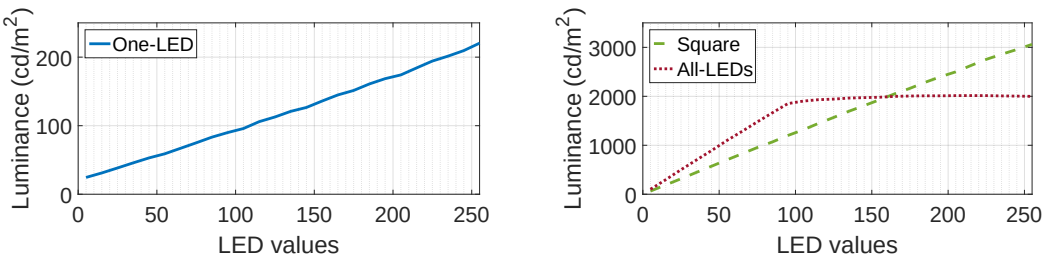
Figure A.1 – Test patterns created for LED measurements for three different cases using (a) only one LED, (b) a collection of LEDs that covers 30% of the display area, and (c) all the LEDs of the backlight layer. LCD values for these test patterns were set to 255 in order to ensure that all of the backlight passes the LCD layer. Estimated luminance values (presented in (d)-(f)) were normalized for representation. The emitted luminance values were measured at the center of the display for each case.

A.1 LED measurements

As described in the Section 2.1.1, the SIM2 display has three different layers: backlight layer, light diffuser layer, and LCD layer. To understand how the LEDs of the backlight layer work, the duality of the 'dual modulation' of the SIM2 display has to be suspended. For this purpose, LCD values for these test patterns were set to 255 in order to ensure that all of the backlight passes the LCD layer. We created a few test patterns to be displayed through a custom dual modulation input provided by the user, also known as DVI Plus (or DVI+) Mode. These test patterns are shown in Figure A.1.

These three test patterns let us understand *i)* the relationship between the LED value stated in the DVI+ header and the luminance of the LED, *ii)* the same relationship in the case of multiple LEDs, and *iii)* the relationship between the physical (i.e. power) limitations of the display and the emitted luminance. For the case *(i)*, we used a single LED and measured its luminance for different values of the LED $\in \{5, 15, 25, \ldots, 255\}$. The representation of the position of the selected LED and its estimated luminance are shown in Figure A.1.(a) and (d), respectively.

In the technical manual of the SIM2 display, it is stated that the brightness of the LEDs is limited by the power limitation of the display which is around 1500 W. It is also stated that this limit is reached when approximately 40% of the LEDs are lit at their maximum brightness. Therefore, we used a collection of LEDs forming a rectangle that covers 30% of the display area for the case *(ii)*, in order to be within the power limitation. The representation of the positions of the selected LEDs and their estimated luminance are shown in Figure A.1.(b) and (e), respectively. Lastly, we try to understand how the emitted luminance changes when the power consumption of LEDs exceeds the power limitation of the display hardware. For this last case, case *(iii)*, all the LEDs of the backlight layer were set to the same value. The representation of the positions of all the LEDs and their estimated luminance are shown in Figure A.1.(c) and (f) respectively.



(a) Only one LED

(b) Multiple LEDs

Figure A.2 – Plots of the measured luminance with respect to the LED value for the case of (a) only one LED and (b) multiple LEDs. The plots show a linear (or close to piecewise linear in the case of all LEDs) relationship between the LED values and the emitted luminance. Since the scale of luminance values are not comparable in the case of only one LED and multiple LEDs, they are presented in different sub-figures.

The emitted luminance values were measured using the Konica Minolta LS-100 light meter. The luminance measurements taken are reported in Table A.1, and the variation of the emitted luminance with respect to the LED value is shown in Figure A.2. Considering the case of only one LED and square LEDs, it is clear that there is a linear relationship between the LED value and the emitted luminance. In the case of all LEDs, the slope of all LEDs case is steeper than the case of square LEDs, and this is expected due to the increased number of LEDs. When LED values pass some threshold, the emitted luminance values are saturated due to the power limitation of the display.

These results show us that the response of individual LEDs is very close to linear and the power limitation is needed to be taken into consideration for an accurate reproduction of the HDR images and video frames.

Table A.1 – The luminance measurements (cd/m^2) for the LED values for three distinct cases. Only one LED was used in the case of 'One-LED', a collection of LEDs forming a rectangle that covers 30% of the display area was used for the case of 'Square', and all of the LEDs were used for the case of 'All-LEDs'.

LED Value	**One-LED**	**Square**	**All-LEDs**
5	24.51	64	101.4
15	31.05	191	297.3
25	38.33	316	501
35	45.8	445	697.3
45	52.98	572	891.7
55	59	701	1096
65	67	827	1284
75	75.03	954	1478
85	83.31	1070	1666
95	90	1200	1851
105	95.88	1315	1902
115	105.9	1446	1931
125	112.7	1568	1943
135	120.9	1695	1964
145	126.6	1810	1973
155	136	1928	1981
165	144.9	2051	1998
175	151.4	2163	2006
185	160.9	2287	2010
195	168.6	2400	2010
205	174.1	2500	2014
215	184.3	2630	2015
225	194.3	2755	2014
235	201.5	2858	2008
245	209.6	2956	2004
255	220.4	3061	1999

A.2 LCD Measurements and the Analysis of the Display Gamma

Similar to the previous section, the duality of the 'dual modulation' of the SIM2 display was suspended by setting all the LEDs same value and creating a constant backlight. We created seven different test patterns for the LCD measurement and displayed these patterns on SIM2 using DVI+ mode. All the pixels of the LCD was set to the same value –or same color– and these colors were chosen as the primary and secondary colors, i.e. Red, Green, Blue, Yellow, Cyan, and Magenta. We also added 'White' to this list to act as the reference luminance values.

The emitted luminance values were measured using the Konica Minolta LS-100 light meter. The measured luminance values were reported in Table A.2. Using these luminance values, we estimated the display gamma value of the SIM2 display. The display gamma –denoted by γ– was calculated using the three color channels; red, green, blue.

Table A.2 – The luminance measurements for the analysis of the display gamma. Measurements were made (in cd/m^2) for each LCD pixel value p where $p \in \{0, 15, 30, \ldots, 255\}$. LED values were kept same during the whole measurement. The luminance values were measured using different color schemes where only the pixel values of the indicated channels were changed, i.e. the first and third channels were changed in the 'Magenta' color scheme, and the second channel pixels were kept as zero.

	Luminance Values (cd/m^2) for Different Color Schemes						
Pixel Value p	**White** $[p\ p\ p]$	**Red** $[p\ \mathbf{0}\ \mathbf{0}]$	**Green** $[\mathbf{0}\ p\ \mathbf{0}]$	**Blue** $[\mathbf{0}\ \mathbf{0}\ p]$	**Yellow** $[p\ p\ \mathbf{0}]$	**Cyan** $[\mathbf{0}\ p\ p]$	**Magenta** $[p\ \mathbf{0}\ p]$
0	0.35	0.355	0.356	0.354	0.349	0.339	0.341
15	2.5	0.757	1.711	0.532	2.034	1.827	0.988
30	13.2	2.838	9.176	1.222	11.48	9.967	3.961
45	34.55	7.36	24.4	2.602	30.84	26.27	9.862
60	68.32	14.51	49	5.01	62.31	52.79	19.28
75	115.8	24.91	83.48	8.462	106.9	89.91	32.75
90	180	39.15	130.7	13.03	168.2	140.5	51.18
105	263	56.93	191	18.91	245.7	205.7	74.05
120	364.3	78.47	266.1	26.6	343.1	286.1	104.1
135	477	102.4	348.6	35.65	450.1	376.5	133.7
150	606	129.4	441.8	46.24	570.5	477.7	170.1
165	744	158.6	543.1	57.21	700	588.2	206.5
180	899	192.7	653.1	69.84	848.4	707.7	250.1
195	1058	229.3	779.1	83.82	1003	842.1	297.6
210	1250	265.3	919.6	100.1	1177	992.5	346.1
225	1436	307.2	1068	118.9	1362	1148	402.8
240	1650	349.9	1223	141	1548	1318	463.4
255	1857	397.1	1379	174.7	1740	1502	541

The γ values were found using the following equation:

$$V_{\text{out},k} = A \times {V_{\text{in},k}}^{\gamma} \tag{A.1}$$

where $A = 1$ and $k \in \{R, G, B\}$ corresponding to 'Red', 'Green', and 'Blue' channels. The luma values $V_{\text{in},k}(p)$ for pixel value p are found by normalizing the luminance value $Y_k(p)$. This normalization was carried out using the following equation:

$$V_{\text{out},k}(p) = \frac{Y_k(p) - \min(Y_k)}{\max(Y_k) - \min(Y_k)} \tag{A.2}$$

where $k \in \{R, G, B\}$ corresponding to the color channels, $\max(Y_k)$ and $\min(Y_k)$ is the maximum and minimum luminance levels for the color channel k, respectively. The maximum and minimum luminance levels are reported in the Table A.2 corresponding to $p = 255$ and $p = 0$, respectively, for each color channel considered (i.e. for red channel, $V_{\text{out},R}(p) = (Y_R(p) - 0.355)/396.745$). The luma values are presented in Table A.3. After this operation, γ values were found by

$$\gamma_k(p) = \frac{\log V_{\text{out},k}(p)}{\log V_{\text{in},k}(p)} \tag{A.3}$$

where p is the pixel value $p \in \{0, 15, 30, \ldots, 255\}$. The resulting $\gamma_k(p)$ values are presented in Table A.4.

Table A.3 – Normalized channel values for the analysis of the display Gamma.

Input Luma V_{in}	**Red** V_R	**Green** V_G	**Blue** V_B
0.0000	0.0000	0.0000	0.0000
0.0588	0.0010	0.0010	0.0010
0.1176	0.0063	0.0064	0.0050
0.1765	0.0177	0.0174	0.0129
0.2353	0.0357	0.0353	0.0267
0.2941	0.0619	0.0603	0.0465
0.3529	0.0978	0.0945	0.0727
0.4118	0.1426	0.1383	0.1064
0.4706	0.1969	0.1928	0.1505
0.5294	0.2572	0.2526	0.2024
0.5882	0.3253	0.3202	0.2632
0.6471	0.3989	0.3937	0.3261
0.7059	0.4848	0.4735	0.3986
0.7647	0.5771	0.5649	0.4787
0.8235	0.6678	0.6668	0.5721
0.8824	0.7734	0.7744	0.6799
0.9412	0.8810	0.8868	0.8067
1.0000	1.0000	1.0000	1.0000

Table A.4 – The γ values found by each pixel value $p \in \{0, 15, 30, \ldots, 255\}$ considered and each color channel. The average value for each channel is given at the bottom row.

Input Luma V_{in}	**Red** γ_R	**Green** γ_G	**Blue** γ_B
0.0588	2.4335	2.4442	2.4308
0.1176	2.3709	2.3606	2.4778
0.1765	2.3271	2.3342	2.5084
0.2353	2.3037	2.3113	2.5039
0.2941	2.2736	2.2950	2.5072
0.3529	2.2325	2.2648	2.5170
0.4118	2.1951	2.2297	2.5248
0.4706	2.1560	2.1841	2.5121
0.5294	2.1351	2.1635	2.5115
0.5882	2.1166	2.1461	2.5157
0.6471	2.1114	2.1415	2.5740
0.7059	2.0786	2.1466	2.6411
0.7647	2.0495	2.1291	2.7458
0.8235	2.0796	2.0875	2.8761
0.8824	2.0529	2.0425	3.0819
0.9412	2.0893	1.9808	3.5430
Average	2.1878	2.2039	2.6544

The relationship between V_{in} and V_{out} for each color channel is shown in Figure A.3.(a). It is clear from both Figure A.3.(a) and Table A.4 that the gamma values for each color channel are different. The luma values are gamma corrected as the following:

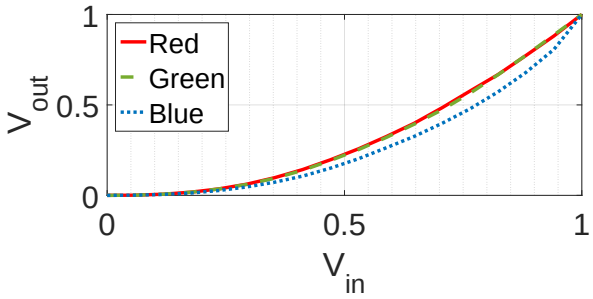
$$\begin{aligned} V_{\text{in},k}^{corrected} &= V_{\text{in},k}^{1/\hat{\gamma}_k} \\ V_{\text{out}}^{corrected} &= A \times (V_{\text{in}}^{corrected})^{\gamma} \\ V_{\text{out},k}^{corrected} &= A \times V_{\text{in},k}^{\gamma/\hat{\gamma}_k} \\ V_{\text{out},k}^{corrected} &= A \times V_{\text{out},k}^{1/\hat{\gamma}_k} \end{aligned} \tag{A.4}$$

where $\hat{\gamma}_k$ is the average gamma per color channel as found in the Table A.4, which are in agreement with the gamma values reported in the work of Nam [Nam10]. The resulting $V_{\text{out}}^{corrected}$ is shown in Figure A.3.(b).

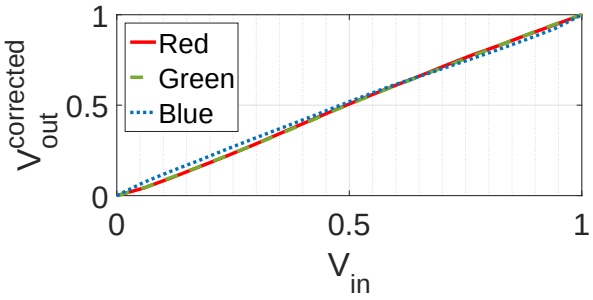
The variation between the γ values presented in Table A.4 are quite high. In addition to that, gamma corrected values are found to have high variation both objectively and subjectively. Subjectively, viewers were able to notice that the gamma corrected images had some bluish or yellowish color artifacts. This finding is supported objectively by the plot shown in Figure A.3.(c). For each color channel, $V_{\text{out}}^{corrected}$ deviates from V_{in}. The plot in Figure A.3.(c) shows us that the gamma correction step is not properly handled. Furthermore, the plot suggests that the gamma value is not constant even for each color

channel.

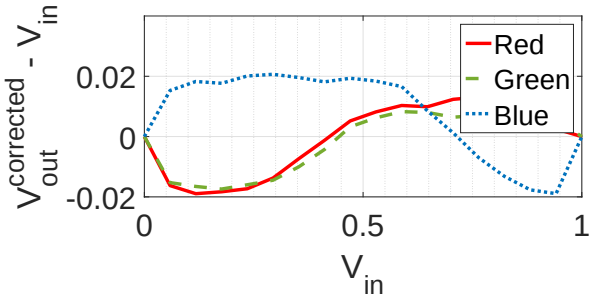
In order to remove this effect and fix gamma correction part, each color channel was measured again with the light meter using pixel values $p \in \{0, 1, 2, 3, ..., 255\}$. The values found were used to create a look-up table for the gamma correction, and this look-up table was used in the gamma correction step in Section 2.1.2 for each pixel value p.



(a) The relationship between V_{in} and V_{out} for each color channel

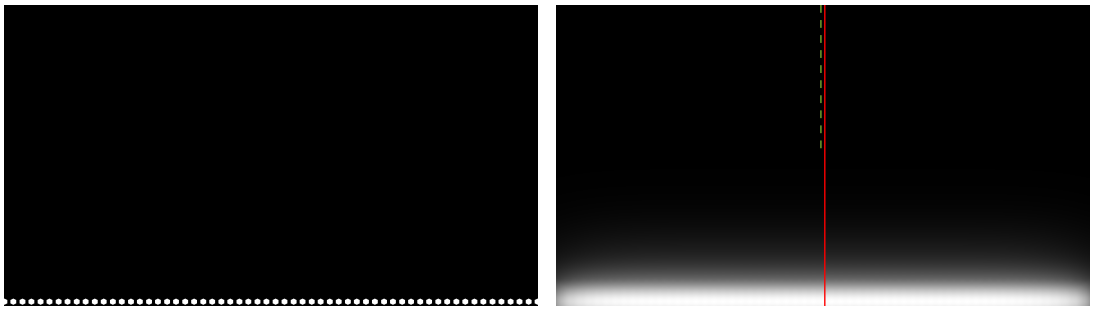


(b) Gamma correction



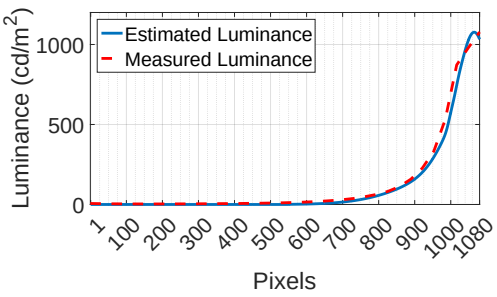
(c) Deviation for each color channel

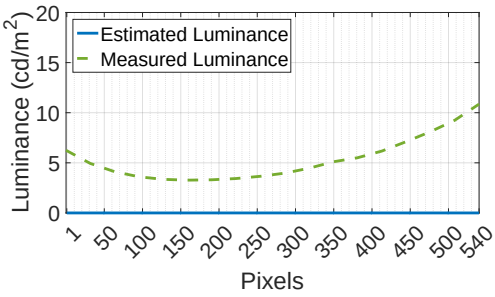
Figure A.3 – Plots for the computation of display gamma. (a) The relationship between V_{in} and V_{out} for each color channel. The plots show that each color channel has different γ values. (b) Plot of input luma and output luma after gamma correction. (c) Deviation of $V_{\text{out}}^{corrected}$ from V_{in} for each color channel.



(a) Only the bottom-most LEDs

(b) Luminance of the selected LEDs





(c) Luminance of pixels at the solid-red cross-section

(d) Luminance of pixels at the dashed-green cross-section

Figure A.4 – Figures of a measurement test where (a) only the bottom-most LEDs were selected. The (b) luminance of the selected LEDs were measured following the solid-red and dashed-green lines. (c) The measured luminance values for the solid-red cross-section are in agreement with the estimated luminance values, and it shows that the estimation of the developed algorithm is accurate. However, (d) the measured luminance values for the dashed-green cross-section reveals a strange phenomenon. The luminance values increase as we move away from the light source and get closer to the edge.

A.3 Point Spread Function & "Border Effect"

In order to fully understand the working principles of the SIM2 display, it is crucial to understand the point spread function (PSF) introduced by the light diffuser layer of the display. Even though a single LED cannot be defined as a point source, within the context of this thesis, we define the light spread caused by a single LED as the *point spread function*. For the custom rendering made available through the DVI+ mode of the display, the knowledge of the PSF is imperative in order to estimate the generated and emitted light.

The point spread function of the SIM2 display was measured by Dr. Francesco Banterle. During the measurement of the PSF, a DSLR camera and a light meter were used. The pixel values acquired by the DSLR camera were normalized using the measurements taken using the light meter. After the normalization, the PSF of the SIM2 display's light diffuser layer has been found with the approximate size of 1000×1000 pixels.

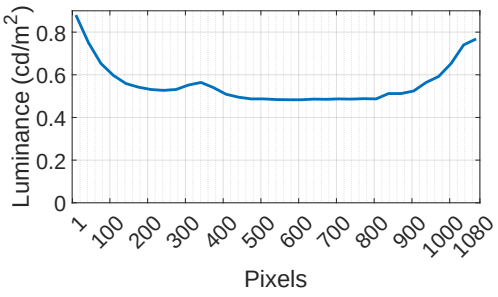
During the development process of the display rendering algorithm proposed in Section 2.1, several measurements were taken to ensure that the reproduction of the developed

algorithm is accurate as well as its estimation of emitted luminance. These tests included several combinations of LED and LCD values as well as different LED patterns. One of the patterns was turning only the bottom-most LEDs on, of which a representation is shown in Figure A.4.(a). The estimated luminance of these LEDs is normalized for the presentation in Figure A.4.(b). A series of measurements was taken on the pixels indicated with a solid-red line on Figure A.4.(b). The measured luminance values were compared to the estimated luminance values, and the estimation results matched the measured luminance values. These two sets of luminance values are shown in the Figure A.4.(c). Nevertheless, a strange phenomenon was observed during these measurements. It was expected that the luminance values should decrease as the measurement point moves away from the LEDs. Contrarily, the measured luminance was increasing. In order to better show this phenomenon, the measured luminance values are re-plotted for the upper-half of the screen, and this plot is shown in the Figure A.4.(d).

According to these measurements, it can be hypothesized that the light generated by the LEDs can bounce back from the chassis of the display and create an imaginary light source at the other end of the display. The measurement results shown in Figure A.4.(d) support this hypothesis.

(a) Only the left-most LEDs

(b) Luminance of the selected LEDs



(c) Luminance of pixels at the solid red cross-section

Figure A.5 – Figures of a measurement test where (a) only the left-most LEDs are selected. The (b) luminance of the selected LEDs were measured following the solid-red line. (c) The measured luminance values for the solid-red cross-section show that the edges of the display has more light compared to the center part.

Moreover, this phenomenon was also observed in another measurement session where

only the left-most LEDs were turned on, as shown in Figure A.5.(a). Again, a series of measurements was taken on the pixels indicated with a solid-red line on Figure A.5.(b). The light coming from the left side of the display was expected to decay until the other edge of the display. It was assumed that the light would have a near-constant luminance for all the pixels on the solid-red line indicated. The measurements made on those pixels reveal that not only the other end but also all edges of the display have this behavior. The luminance values seem to increase while getting close to the edges as it is shown in the Figure A.5.(c).

These measurements were replicated in several other tests. Similar experiments were conducted with different patterns of LEDs, and the same phenomenon was observed over and over again in all the other experiments.

These results update the hypothesis. Although we do not know the exact cause of this circumstance, it is hypothesized that the light generated by the LEDs can bounce back from the chassis of the display, and it affects all of the display's edges. We call this effect "border effect". After comprehensive measurements, this border effect was integrated within the HDR frame rendering algorithm developed.

The results of these measurements helped us to understand the characteristics of the LEDs, the pixels of the LCD panel, and the overall working principles of the SIM2 display. We were able to model some key parameters related to these characteristics, and we integrated the findings into the display rendering algorithm proposed in Section 2.1.

Annex B

Résumé de thèse

B.1 Introduction

Le système visuel humain (HVS - Human Visual System) est capable de percevoir une gamme beaucoup plus large de couleurs et d'intensités lumineuses présentes dans notre environnement que les systèmes d'imagerie à dynamique standard (SDR - Standard Dynamic Range) traditionnels peuvent capturer et reproduire. Avec les développements de la technologie à dynamique haute (HDR - High Dynamic Range), nous sommes maintenant en mesure de capturer, stocker, transmettre et afficher des images et des vidéos d'une manière plus réaliste [BADC11, DLCMM16]. Pouvoir reproduire des scènes HDR a accéléré les efforts de standardisation pour la compression d'images et de vidéos HDR [Ric13, LFH15, HRE16] en tant que parties d'une chaîne de livraison de contenu HDR de bout en bout. Afin de s'assurer que la compression est effectuée avec la meilleure qualité possible, une évaluation de la qualité est nécessaire pour les images et les vidéos HDR.

Cette thèse se concentre sur l'évaluation et l'analyse de l'image et de la vidéo à haute gamme dynamique. Le problème de l'évaluation de la qualité de l'image et de la vidéo est un problème largement étudié dans la communauté du traitement du signal [SSB06, SSBC10a, PJI$^+$15] pour le cas du SDR. La perception humaine de la lumière n'est pas proportionnelle à la magnitude physique de la lumière. Pour ce faire, les valeurs des pixels de l'image sont traitées à l'aide d'une courbe de loi de puissance, appelée *fonction de correction de gamma* [ITU11], pour les affichages SDR. Après cette opération, les valeurs de pixels SDR deviennent perceptiblement linéaires où le changement de magnitude correspondra à un changement proportionnel de la perception. Ainsi, les mesures objectives de qualité SDR supposent que les pixels de l'image sont perceptiblement uniformes. Ce n'est pas le cas pour les images HDR, car les images HDR stockent généralement des valeurs de pixels proportionnelles aux valeurs de luminance physique. De même, on s'attend à ce que l'évaluation subjective de la qualité HDR soit différente puisque le niveau et le ratio de luminosité sont différents. Pour une évaluation correcte, ces nouvelles conditions doivent être prises en compte.

Bien que l'estimation et l'évaluation de la qualité de la vidéo soient essentielles pour de nombreuses autres applications, la compression de l'image et de la vidéo est considérée comme la principale source de distorsion tout au long de la thèse, car il s'agit du scénario le plus pratique et le plus réaliste. Sur la base de ces considérations, nous posons la question suivante : ***Quels sont les paramètres qui affectent l'estimation de la qualité objective avec référence et la perception de la qualité subjective dans le cas de la compression d'image et de vidéo HDR ?***

Pour tenter de répondre à cette question, nous identifions d'abord deux aspects principaux qui peuvent avoir un impact sur l'évaluation objective et subjective de la qualité de l'image et de la vidéo HDR :

- Afin de rendre les valeurs de pixels HDR perceptiblement uniformes, plusieurs méthodes de codage des pixels ont été proposées [AMS08, SMP14, MND12, Bor14]. Cependant, l'impact de la connaissance des valeurs de luminance émises et les effets des différents rendus d'affichage sur l'évaluation de la qualité HDR n'ont pas encore été étudiés. C'est pourquoi nous essayons de répondre à la question suivante : *Comment le rendu de l'affichage HDR affecte l'évaluation de la qualité HDR, à la fois subjectivement et objectivement ?*

- La luminance accrue dans les conditions HDR peut changer la façon dont nous percevons la qualité, et la couleur peut influencer la qualité perceptuelle en raison de certains aspects des phénomènes d'apparence des couleurs, par exemple l'effet Hunt, le changement de teinte Bezold-Brücke, etc. [Fai13]. Comme nous considérons la compression comme notre principale distorsion tout au long de la thèse, nous essayons d'analyser et de comprendre l'impact de la couleur sur la compression et donc de poser la question suivante : *Quels sont les effets de la transformation de l'espace couleur et des distorsions spécifiques à la couleur sur l'évaluation de la qualité HDR ?*

En analysant les effets de ces deux aspects et en évaluant les mesures de qualité dans les chapitres suivants, nous constatons que les résultats subjectifs de l'évaluation, c.-à-d. les notes d'opinion moyennes (MOS - Mean Opinion Scores), ont des fourchettes différentes pour les bases de données de qualité annotées subjectivement que nous considérons. Même si la qualité objective des stimuli est la même, le score de qualité subjective d'un stimulus peut être différent pour différentes bases de données. Cette observation a des résultats importants pour l'évaluation subjective de la qualité.

Comme la 'qualité' est subjective par définition, la plupart des algorithmes d'évaluation objective de la qualité utilisent les valeurs MOS comme vérité terrain. Afin d'utiliser ces bases de données pour l'évaluation ou l'élaboration de mesures objectives, les valeurs MOS doivent être alignées. De cette façon, les scores subjectifs de qualité de deux stimuli avec les mêmes scores objectifs de qualité seraient similaires. Pour s'attaquer à ce problème,

nous essayons de répondre à la question suivante : *Comment pouvons-nous mieux définir une échelle de qualité qui ne serait pas affectée par les facteurs environnementaux et quelle méthodologie subjective d'évaluation de la qualité devrions-nous utiliser ?*

Tout au long de la thèse, nous visons à répondre à ces questions et à comprendre les facteurs sous-jacents qui affectent l'évaluation de la qualité HDR, par une série d'expériences subjectives et d'analyses approfondies.

B.2 Notions de base

Dans le chapitre 1, les études précédentes sur l'évaluation subjective et objective de la qualité, les étapes de la distribution du contenu HDR et l'état de l'art de l'évaluation de la qualité HDR ont été discutées. Premièrement, les méthodes subjectives d'évaluation de la qualité, leur comparaison et leur utilisation dans l'évaluation de la qualité de l'image et de la vidéo ont été discutées, ainsi que les méthodes objectives d'évaluation de la qualité de l'image et de la vidéo couramment utilisées. En outre, des méthodes statistiques et d'autres méthodes d'évaluation objective de la qualité ont été décrites. Deuxièmement, l'imagerie HDR et la distribution du contenu ont été expliquées en détail, en commençant par les méthodes d'acquisition et de stockage et en incluant les méthodes de compression d'images et de vidéo HDR jusqu'à la reproduction et l'affichage. Enfin, les recherches de pointe pour l'évaluation subjective et objective de la qualité ont été discutées pour le contenu HDR.

B.3 Effets du rendu d'affichage sur l'évaluation de la qualité d'image HDR

Les écrans HDR ont une luminance de crête plus élevée et un contraste élevé par rapport aux écrans SDR. Dans le chapitre 2, nous examinons le fonctionnement du SIM2 et analysons les conditions de visualisation et leurs effets sur l'évaluation subjective et objective de la qualité. Cette évaluation des effets sur la qualité a été faite en comparant deux méthodes de rendu d'affichage différentes : la méthode de rendu SIM2 intégrée et une méthode de rendu d'affichage que nous proposons dans ce chapitre. Dans ce qui suit, nous décrivons en détails la méthode de rendu d'affichage proposée, présentons les résultats de la validation expérimentale et discutons des effets de l'utilisation de différentes méthodes de rendu d'affichage sur la qualité subjective et objective de l'image HDR.

B.3.1 Reproduction précise d'image à haute gamme dynamique

La méthode la plus populaire pour la production d'écrans HDR est l'utilisation de différentes couches pour le rétroéclairage et l'ajustement des couleurs, une méthode connue sous le nom de *dual modulation* [SHS+04, NDSLC16a]. Cela se fait en couplant une source de lumière à gradation locale, comme un panneau de LEDs, avec un écran LCD frontal. Mais,

les valeurs des pixels LED et LCD doivent être calculées afin de reproduire une image HDR dans ce cadre.

Caractéristiques d'affichage

L'algorithme de rendu que nous proposons est conçu pour fonctionner sur les affichages SIM2 HDR47E S 4K [SIM14]. L'écran est un double affichage modulé qui comprend un réseau de LED pour le rétro-éclairage, une couche de diffuseur de lumière et un panneau LCD dans cet ordre. Il y a 2202 lumières LED contrôlables indépendamment et un panneau LCD de 1920×1080 pixel qui peut être contrôlé séparément. L'affichage SIM2 peut être contrôlé à l'aide du rendu automatique intégré ou d'une entrée double modulation personnalisée fournie par les utilisateurs.

Un algorithme de double modulation pour la reproduction d'images

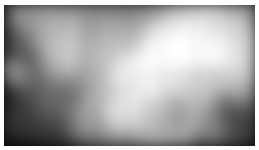
Dans cette thèse, nous proposons et développons une méthode de rendu à double modulation sur mesure afin de reproduire les images HDR avec une très grande précision et fidélité par rapport aux valeurs de luminance prévues. L'algorithme se compose des parties suivantes :

- *Prétraitement :* Tout d'abord, nous trouvons les valeurs de luminance cibles *se référant à moniteur (display-referred)* à partir de l'image HDR *se référant à scène (scene-referred)*. En supposant que les images d'entrée ont été préalablement graduées à l'affichage –manuellement ou par un processus automatique [MDK08]–, nous saturons juste des valeurs de luminance supérieures à la luminance maximale de l'affichage, c.-à-d. 4250 cd/m^2. Nous notons I l'image prétraitée.

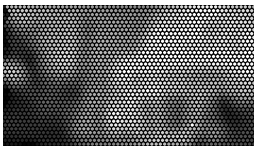
- *Calcul du rétroéclairage cible :* Ensuite, nous trouvons le rétroéclairage optimal cible, BL_{target}, qui minimise la luminance du rétroéclairage requis. Pour trouver BL_{target}, nous définissons deux autres images rétroéclairées : BL_{min} et BL_{max}. Comme les cellules à cristaux liquides ne peuvent que bloquer la lumière et ne peuvent pas générer de lumière, au moins BL_{min} est nécessaire pour s'assurer que le rétroéclairage est suffisant pour tous les pixels. Pour trouver BL_{min}, on calcule les maxima locaux de la luminance cible sur des fenêtres d'un rayon de 30 pixels correspondant à la surface d'une seule LED. Les cellules à cristaux liquides ne sont pas idéales et laissent échapper de la lumière même si elles sont complètement fermées. Afin de contrôler les effets de fuite de l'écran LCD, la luminance maximale pour chaque pixel, BL_{max}, est obtenue en divisant les valeurs de luminance de l'image de ce pixel par le facteur de fuite estimé de l'écran LCD $\epsilon = 0,005$, ce qui est empirique. BL_{target} est alors trouvé en filtrant et en combinant les BL_{min} et BL_{max}.

  Après le calcul du rétroéclairage cible, les LEDs et le rétroéclairage sont initialisés en échantillonnant BL_{target} sur les emplacements des LEDs et en prenant la convolution

(a) L'image HDR (Tone mappée en utilisant [MDK08] pour la représentation)



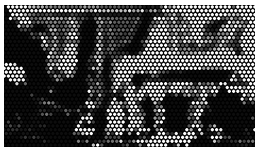
(b) BL_{target} – Rétroéclairage cible



(c) LED_0 – Réseau de LED initialisé par l'échantillonnage du rétroéclairage cible

(d) S_0 – Échelle pour la première itération

(e) BL_2 – Rétroéclairage trouvé lors de la seconde itération



(f) LED_{final} – Réseau de LED trouvé

(g) BL_{final} – Rétroéclairage trouvé

(h) Valeurs des pixels de l'écran LCD

Figure B.1 – Étapes de l'algorithme de rendu d'image HDR pour l'image HDR *Market3*

avec la PSF, respectivement. C'est à dire :

$$BL_t = LED_t * PSF \tag{B.1}$$

où t est l'index d'itération et $t = 0$ pour l'initialisation. LED_0 correspond aux valeurs initiales des LEDs trouvées en échantillonnant BL_{target}, et BL_0 est le rétroéclairage de LED_0.

- *Mise à l'échelle itérative :* Une carte à l'échelle est générée afin de mettre à jour les valeurs des LED à l'aide de l'équation suivante :

$$S_t = \frac{BL_{target}}{BL_t} \tag{B.2}$$

où t est le nombre d'itération. Les valeurs des LED sont multipliées par la carte à l'échelle trouvée comme suit :

$$LED_t = LED_{t-1} \times S_{t-1} = LED_{t-1} \times \left(\frac{BL_{target}}{BL_{t-1}} \right) \tag{B.3}$$

LED_t est ensuite clippé pour prendre des valeurs en $[0, 1]$, c.-à-d. qu'il est projeté sur l'ensemble des valeurs LED réalisables à chaque itération. Une fois les valeurs LED trouvées, les valeurs de rétroéclairage sont également trouvées à l'aide de l'équation B.1.

Les opérations dans les équations B.1 et B.3 sont effectuées consécutivement en augmentant le nombre d'itération jusqu'à ce que $\sum ||PU(BL_t) - PU(BL_{t-1})||^2$ tombe en dessous d'un seuil. Lorsque la mise à l'échelle itérative converge, les valeurs

LED_{final} résultantes peuvent être mises à l'échelle linéairement pour répondre aux contraintes d'alimentation de l'affichage.

- *Calcul des valeurs des pixels de l'écran LCD :* Les valeurs de pixels LCD sont trouvées en divisant (par pixels) chaque canal de couleur de l'image HDR originale par l'estimation finale du rétroéclairage, et en appliquant une correction gamma, c.-à-d.:

$$LCD_k = \left(\frac{I_k}{BL_{final}}\right)^{1/\gamma_{k,p}} = \left(\frac{I_k}{LED_{final} * PSF}\right)^{1/\gamma_{k,p}} \tag{B.4}$$

où I est l'image HDR, $k \in \{R, G, B\}$ est l'indicateur de canal RGB, $p \in \{0, 1, 2, ..., 255\}$ est la valeur du pixel LCD, et $\gamma_{k,p}$ est le facteur de correction gamma, déterminé expérimentalement comme expliqué dans l'annexe A.2.

Des exemples de résultats étape par étape de l'algorithme de reproduction proposé peuvent être vus dans la figure B.1, y compris l'image de rétroéclairage cible, une carte de l'échelle, le rétroéclairage de la seconde itération, le réseau de LED final et le rétroéclairage final –BL_{final}–, et les valeurs des pixels de l'écran LCD.

- *Estimation de la luminance émise :* Connaissant les valeurs des LEDs et des pixels LCD, nous pouvons estimer la luminance émise. Les pixels de l'image HDR produits par l'écran sont le produit du rétroéclairage et des valeurs LCD. C'est-à-dire, pour chaque canal de couleur k, l'image rendue I'_k est :

$$I'_k = (LED_{final} * PSF) \times LCD_k. \tag{B.5}$$

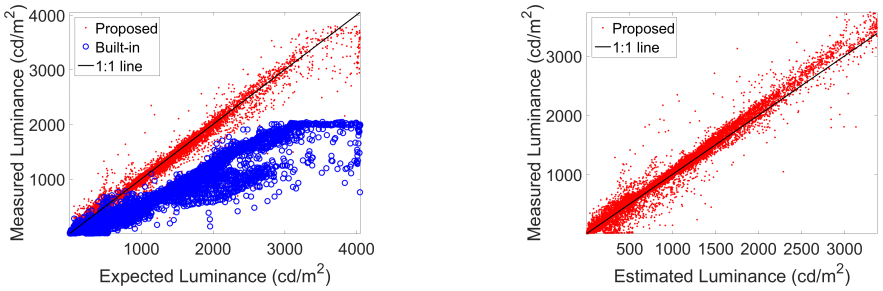
En supposant que l'on utilise les primaires de l'ITU-R BT.709 [ITU15a], nous pouvons calculer la luminance émise comme suit :

$$L = 0.2126 \times I'_R + 0.7152 \times I'_G + 0.0722 \times I'_B, \tag{B.6}$$

Un algorithme de double modulation pour la reproduction vidéo

Afin de réduire l'impact du scintillement, nous envisageons deux solutions. Tout d'abord, nous initialisons les valeurs des LED pour l'image actuelle f en utilisant celles de l'image précédente, c.-à-d. $LED_0^f = LED_{final}^{f-1}$. Deuxièmement, nous lissons les rétroéclairage cible dans le temps, sur des fenêtres consécutives qui se chevauchent, comme décrit ci-dessous.

Étant donné l'image f d'une vidéo, son rétroéclairage cible initial BL_{target}^f est calculé comme expliqué dans la section précédente. Ensuite, pour chaque image, nous considérons une fenêtre d'anticipation de N images, et nous arrangeons leur rétroéclairage correspondant dans une pile A^f. Par la suite, nous visons à lisser la trajectoire des valeurs des pixels rétroéclairés sur la fenêtre, en adoptant une approche simple qui consiste à convolutionner chaque échantillon indépendamment par une fenêtre gaussienne et à prendre le maximum



(a) Luminance mesurée par rapport à la luminance attendue

(b) Luminance mesurée par rapport à la luminance estimée

Figure B.2 – Résultats de validation expérimentale pour l'image "AirBellowsGap".

à chaque instant. Plus précisément, nous trouvons $T^f_{i,j,l} = [0 \dots A^f_{i,j}(l) \dots 0]^T$ et $B^f_{i,j} = [W^f_{i,j,1} \dots W^f_{i,j,N}]$ où $W^f_{i,j,l} = T^f_{i,j,l} * w_\sigma$ est le vecteur résultant obtenu par le filtre de lissage w_σ. Ensuite, les valeurs lissées, $M^f_{i,j}$, peuvent être trouvées en prenant la valeur maximale à travers les colonnes de $B^f_{i,j}$.

Cette procédure est répétée en utilisant une approche par fenêtre coulissante. Une fois que le rétroéclairage cible lissé a été calculé, le reste de la partie rendu suit l'algorithme décrit précédemment.

Validation expérimentale

La performance de l'algorithme de rendu proposé a été validée par différentes expériences :

- *Réponse de luminosité linéaire et luminance de crête :* À l'aide d'un photomètre, nous avons mesuré la luminance au centre du motif (une boîte blanche couvrant 30% de la surface d'affichage, entourée d'un fond noir) pour le rendu intégré et le rendu proposé. Le rendu proposé est plus précis et atteint une luminosité de crête plus élevée.

- *Contraste local :* À l'aide d'un photomètre et d'un autre modèle (une boîte blanche de 30% de la surface de l'écran avec une tache noire carrée de 64×64 pixels carrés au milieu), la luminance sur la tache noire centrale et la zone blanche juste à l'extérieur ont été mesurées. Les résultats montrent que le rendu proposé est meilleur pour traiter les fuites de l'écran LCD et permet d'obtenir un contraste plus élevé.

- *Mesure de la luminance en pixels :* Pour mesurer la luminance en pixels, nous avons capturé 7 images à l'aide d'un appareil photo reflex numérique avec différentes expositions, nous les avons fusionnées en une seule image HDR et normalisé les valeurs d'image HDR par les mesures du photomètre. Ensuite, nous avons vérifié la fidélité de l'algorithme proposé en termes de reproduction et d'estimation de la luminance émise. Les résultats montrent que l'algorithme proposé est capable de

reproduire l'image HDR et d'estimer très précisément sa luminance émise, et qu'il génère également des images plus lumineuses comme le montre la figure B.2.

- *Variation temporelle :* Bien que la méthode de rendu vidéo proposée n'ait pas été validée subjectivement, dans cette partie, nous avons mesuré la variation temporelle à l'aide d'un calcul objectif. Puisque la principale source de la variation temporelle sur la vidéo est le rétroéclairage, nous avons calculé les différences d'image sur le rétroéclairage au lieu de l'image, et nous avons mesuré la variation temporelle en utilisant l'indice de perception temporelle de l'information perceptuelle (TI) [ITU08]. Les résultats montrent que l'approche fenêtrée réduit objectivement le scintillement dans le rétroéclairage et l'augmentation de la longueur de la fenêtre N réduit l'écart type de la différence d'image, comme prévu.

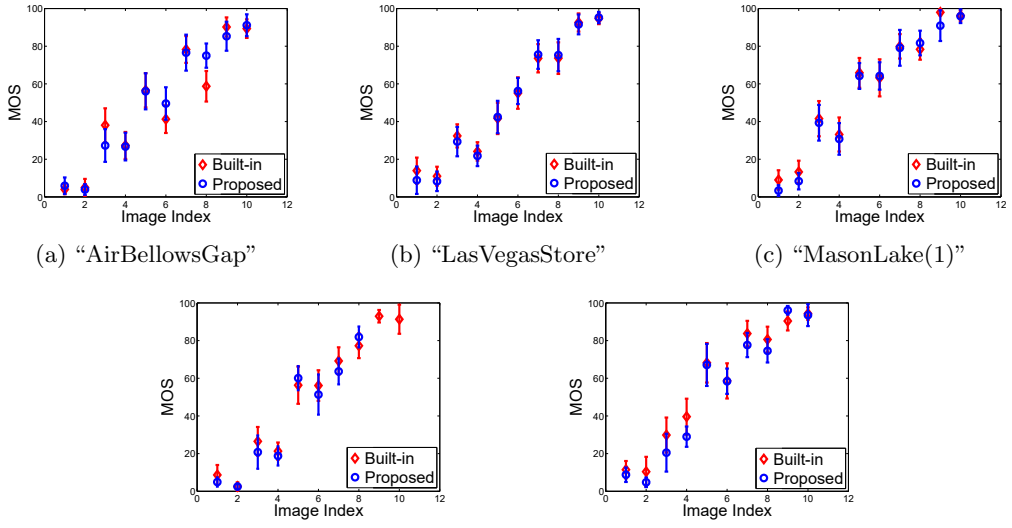
B.3.2 Effets du rendu d'affichage

Différents rendus peuvent avoir un impact potentiel sur le calcul de la qualité objective, en effet le HDR [MDMS05, AMMS08, NMDSLC15, NDSLC15] et le SDR [AMS08, VDSLD14] requièrent des valeurs de luminance par pixel (exprimées en cd/m^2) en entrée. Un rendu d'affichage différent peut également avoir un impact sur l'expérience et la perception de la qualité par les téléspectateurs. Dans cette partie, nous avons comparé deux méthodes de rendu différentes et évalué l'impact du rendu des images HDR sur les scores subjectifs et objectifs.

Impact sur l'évaluation subjective

Equipé du rendu que nous avons proposé dans la section précédente, nous avons mené une étude subjective pour comparer la qualité perçue des images HDR compressées rendues par deux algorithmes différents. Pour ce faire, nous avons utilisé les mêmes paramètres, méthodologie et environnement de test que dans le travail précédent de Valenzise et al. [VDSLD14], sauf que nous avons affiché des images avec l'algorithme de rendu proposé. Les scores de qualité subjective recueillis ont été comparés à l'aide d'analyses comparatives multiples en plus de l'analyse qualitative des images HDR obtenues. Les détails de cette expérience ont été décrits en détail dans la section 2.2.1.

Les valeurs MOS résultantes pour chaque contenu sont indiquées dans la figure B.3. Les résultats du rendu proposé ont été comparés aux valeurs MOS collectées à l'aide du rendu intégré [VDSLD14]. Ces graphiques montrent un niveau substantiel de concordance entre les scores obtenus avec les deux rendus. Ce résultat est corroboré par les résultats de l'analyse comparative multiple sur les valeurs MOS effectuée après une analyse de la variance unidirectionnelle. Les résultats montrent que, dans l'ensemble, les valeurs MOS ne sont pas affectées de façon spectaculaire par le rendu employé.



(a) "AirBellowsGap" (b) "LasVegasStore" (c) "MasonLake(1)"

(d) "RedwoodSunset" (e) "UpheavalDome"

Figure B.3 – Notes moyennes d'opinion par différents rendus pour les contenus testés. Les points indiquent les valeurs MOS et les barres indiquent les intervalles de confiance.

Impact sur l'évaluation objective

Nous avons également comparé les performances de différents cas de calcul de mesures objectives de qualité. Afin de comprendre l'effet du rendu sur l'évaluation objective de la qualité HDR, nous avons estimé la luminance par pixel produite par l'affichage à l'aide de notre algorithme de rendu et l'avons utilisé comme entrée dans les mesures de qualité pour le contenu vierge et compressé. Les résultats calculés à l'aide de la luminance estimée n'étaient pas significativement différents des résultats calculés à l'aide d'une fonction linéaire très simple, appelée *modèle linéaire*. Les résultats de l'analyse objective de la qualité montrent qu'un modèle linéaire simple, qui n'exige que la luminosité maximale de l'écran, peut fournir des résultats fiables comme si une connaissance détaillée de l'écran de reproduction était disponible.

B.4 Effets de l'espace colorimétrique sur la compression et la qualité vidéo HDR

La couleur peut influencer la qualité perceptuelle dans les conditions HDR, du à ses niveaux de luminosité et de contraste accrus, en raison de certains aspects des phénomènes d'apparence des couleurs, par exemple l'effet Hunt, le changement de teinte Bezold-brucke, etc. [Fai13]. Dans le chapitre 3, nous essayons de comprendre l'effet de la couleur sur la qualité perçue. Pour ce faire, nous avons choisi un scénario d'application pratique et réaliste, la compression vidéo HDR, et nous avons comparé les effets de trois espaces de couleurs différents sur les performances de compression vidéo HDR : Y'CbCr [ITU15a],

ITP (ICtCp) [LPY+16] et Ypu'v' – un espace couleur basé sur LogLuv [Lar98a] modifié pour le contenu HDR.

B.4.1 Sélection des stimuli de test

La méthodologie des comparaisons par paires a été choisie pour comparer les effets de différents espaces colorimétriques, car les différences entre les vidéos compressées avec différents espaces colorimétriques sont subtiles. Afin d'acquérir des données significatives sans décisions unanimes, nous avons d'abord mené une expérience préliminaire pour sélectionner les stimuli. Cette expérience préliminaire a été conçue pour trouver des distances perceptuellement uniformes entre des séquences vidéo HDR compressées à différents niveaux, mesurées en unités de différence juste perceptible (JND - Just Noticeable Difference). Pour chaque contenu, quatre échelons de JND ont été trouvées.

Détails de l'expérience subjective pour la sélection des stimuli

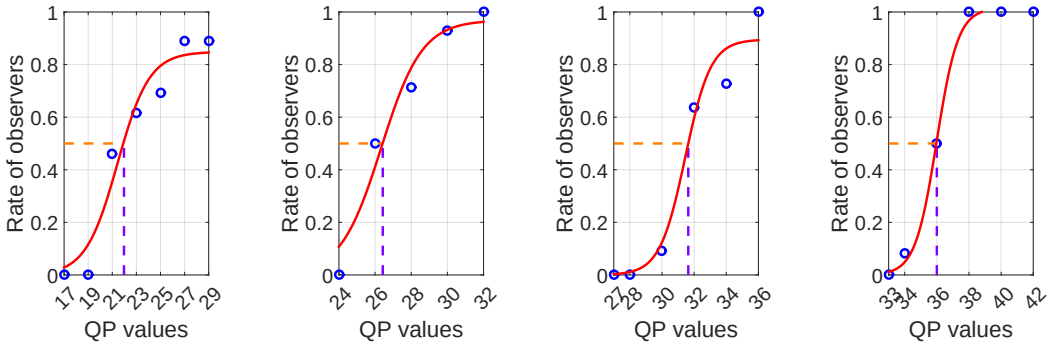
Cette expérience a été menée en quatre séances à l'aide d'évaluation de comparaisons par paires (PC - pairwise comparisons) à choix forcé à deux alternatives (2AFC - two alternative forced-choice), où la question était : *"Pouvez-vous observer une différence de qualité entre les deux vidéos affichées ?"*, et les sujets ont été en mesure de répondre 'Oui' ou 'Non'. Dans chaque essai, deux vidéos ayant le même contenu mais des niveaux de compression différents ont été affichées côte à côte.

Les images RGB HDR ont été encodées en utilisant le quantificateur perceptuel (PQ - perceptual quantizer) [SMP14] EOTF puis transformées en espace couleur Y'CbCr et encodées en utilisant le profil HEVC Main-10 avec HM 16.5 [SOHW12, BFSS17]. Les flux de bits codés ont ensuite été décodés et la transformation des couleurs et le codage EOTF ont été inversés. Les images résultantes ont été stockées dans un fichier AVI sous forme d'images vidéo non compressées.

33 personnes (20 hommes et 13 femmes) d'un âge moyen de 33,6 ans se sont portées volontaires pour l'expérience. Les expériences ont été menées dans une pièce sombre et silencieuse, avec la luminance de l'écran éteint à 0,03 cd/m^2. Les stimuli ont été présentés sur un écran calibré SIM2 HDR47 avec une résolution de 1920×1080 pixels, une luminosité maximale de 4250 cd/m^2, utilisé dans son mode de rendu intégré natif.

Sélection de stimuli pour l'expérience de l'espace colorimétrique

Afin de trouver les stimuli séparés par 1 JND, l'expérience a été menée de manière itérative. Après chaque séance, les données résultantes ont été recueillies et examinées pour en assurer la cohérence. Les résultats de chaque séquence vidéo pour chaque participant ont été regroupés et analysés. Les cas où toutes les réponses étaient "Même" ou "Différentes" ont été considérés comme des valeurs aberrantes, et les résultats de ce participant particulier pour cette scène particulière ont été rejetés, car ces conditions sont extrêmement improbables.



(a) $\widehat{QP}_1 = 22$ (b) $\widehat{QP}_2 = 26$ (c) $\widehat{QP}_3 = 32$ (d) $\widehat{QP}_4 = 36$

Figure B.4 – Le processus de recherche des vidéos de distance 1 JND pour la séquence vidéo Balloon. Les valeurs de QP correspondantes sont trouvées, où seulement 50% des participants ont pu voir la différence entre les vidéos. Les valeurs zéro et un sur l'axe vertical indiquent que la différence peut être observée par aucun ou tous les observateurs, respectivement.

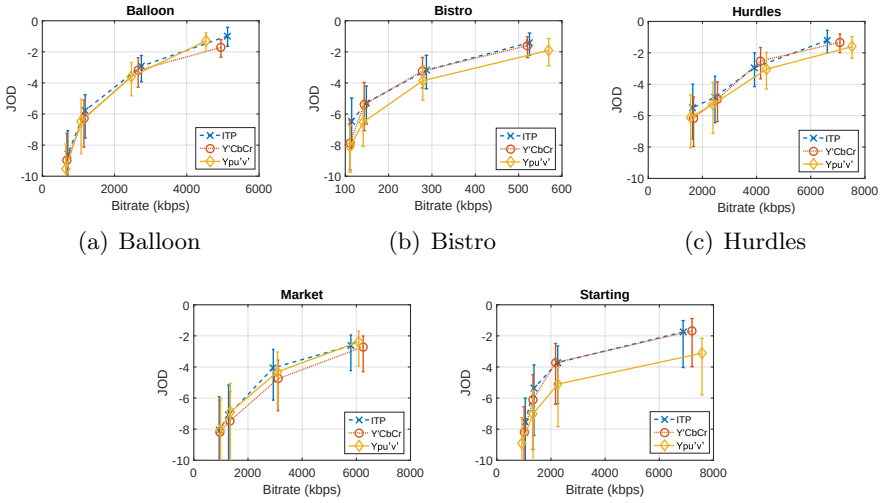
Les résultats ont été analysés plus en profondeur pour en vérifier la cohérence. Les résultats rassemblés pour chaque participant et chaque séquence vidéo devaient suivre un modèle simple : les utilisateurs ne verraient aucune différence dans les vidéos jusqu'à un certain point de seuil et verraient la différence dans tous les stimuli après ce point. Les résultats qui n'ont pas suivi ce comportement attendu ont été considérés comme incohérents, et ils ont été modifiés pour être cohérents comme décrit en section 3.1.2.

Les résultats modifiés $\widehat{R}_{k,l,m,i}$ des différents participants ont été additionnés, et le résultat a été tracé. Le résultat de cette somme ressemble –comme attendu– à une fonction de distribution cumulative (CDF) de la probabilité de voir une différence. La paramètre QP correspondant à un écart de 1 JND avec la vidéo de référence, c.-à-d. $\widehat{QP}_k$, a été déterminée en trouvant la valeur de QP la plus proche correspondant aux 50% d'observateurs qui voient la différence. Des exemples de cette opération sont montrés dans la figure B.4, et le processus de détermination de $\widehat{QP}_k$ est indiqué par des lignes pointillées. Les valeurs de QP pour les vidéos compressées avec d'autres espaces de couleur, à savoir ITP et Ypu'v', ont été déterminée en prenant les valeurs de QP minimisant la différence de débit par rapport à la vidéo Y'CbCr.

B.4.2 Effet de l'espace de couleur sur la compression

Détails de l'expérience subjective

Pour la tâche expérimentale principale, nous avons choisi la méthodologie des comparaisons par paires qui offre une plus grande sensibilité et simplifié l'expérimentation par rapport aux méthodes de notation directe. Cependant, cette méthode peut nécessiter la comparaison d'un nombre excessif de paires lorsqu'un grand nombre de conditions est impliqué [MTM12]. Afin d'éviter des résultats évidents et inutiles, un schéma incomplet dans lequel seules les



(a) Balloon (b) Bistro (c) Hurdles

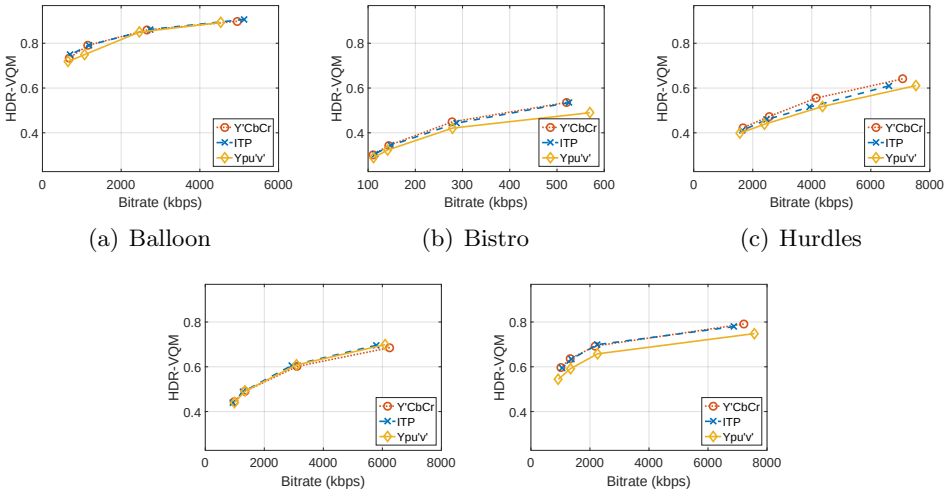
(d) Market (e) Starting

Figure B.5 – Les scores d'image obtenus en rééchelonnant les préférences en fonction des distances de qualité relative (en unités JOD) pour les trois espaces colorimétriques testés.

paires pertinentes sont comparées a été utilisé. Les séquences vidéo HDR ont été comparées à travers des débits consécutifs pour le même espace couleur, et à travers des espaces couleur utilisant le même débit. 18 personnes (14 hommes et 4 femmes), avec une moyenne d'âge de 29,44 ans, se sont portées volontaires pour l'expérience principale.

Les résultats d'un test de comparaison par paires sont généralement rassemblés dans une matrice de préférence ou de comparaison. Cette matrice comprend les ratios de préférence des stimuli, et ces ratios de préférence peuvent être convertis en scores de qualité par une procédure appelée *"rééchelonnement"* (*"scaling"* en anglais). Il existe plusieurs méthodes pour effectuer le rééchelonnement [BT52, Thu27, LDSE11, TG11], et ces méthodes utilisent deux modèles : le modèle Bradley-Terry [BT52], et le modèle Thurstone [Thu27]. Les résultats de comparaison par paires obtenus ont été rééchelonnés à l'aide du logiciel *pwcmp* accessible au public[1]. Le logiciel utilise une méthode bayésienne, qui utilise un estimateur du maximum de vraisemblance pour maximiser la probabilité que les données recueillies expliquent les scores de qualité à l'échelle sous les hypothèses du scénario V de Thurstone. Il peut rééchelonner le résultat d'un schéma incomplet et déséquilibré par paire, ainsi que les cas où il y a accord unanime. Les paramètres de distribution du logiciel sont ajustés de manière à ce que la différence d'une valeur de qualité corresponde au taux de préférence de 75%. Comme les comparaisons par paires ne peuvent fournir que des informations relatives sur la qualité, les valeurs JOD ('just objectionable difference' en anglais) sont également relatives. Pour maintenir la cohérence entre les séquences vidéo, nous fixons toujours le point de départ de l'échelle JOD à 0 pour différentes distorsions et la dégradation de la

[1] *pwcmp* toolbox pour le rééchelonnement des données de comparaison par paires `https://github.com/mantiuk/pwcmp`



(a) Balloon (b) Bistro (c) Hurdles

(d) Market (e) Starting

Figure B.6 – Les résultats obtenus en comparant toutes les scènes pour les trois espaces couleurs en utilisant la métrique HDR-VQM. Tous les scores sont normalisés, où 1 signifie une qualité parfaite et des scores plus faibles représentent une diminution de la qualité.

qualité se traduit par des valeurs JOD négatives.

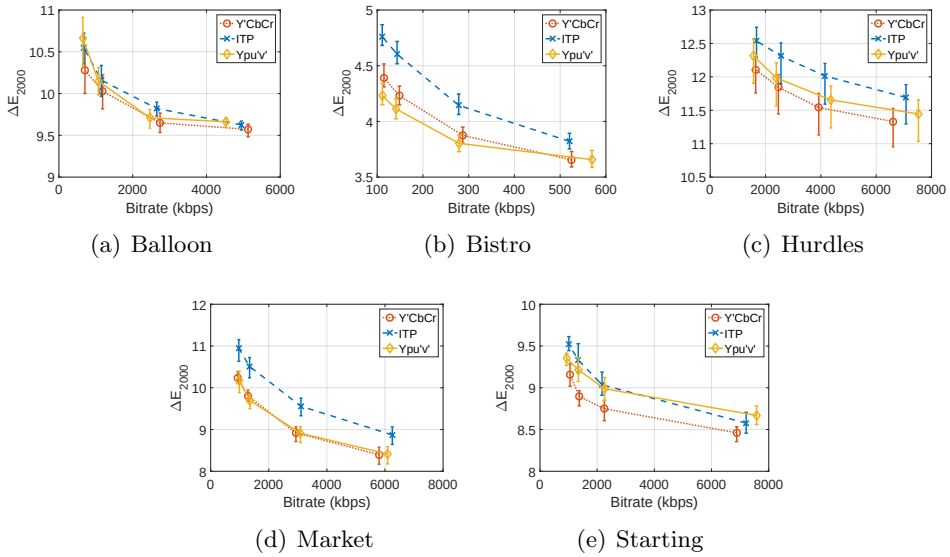
Analyse des résultats subjectifs

La matrice de comparaison pour chaque séquence vidéo a été formée séparément puisque chaque stimulus a été comparé à un autre stimulus avec le même contenu. Pour chaque séquence vidéo, la vidéo originale non compressée a été fixée à zéro JOD afin de fixer la relativité à la vidéo originale. Ensuite, les valeurs de JOD ont été trouvées pour les stimuli en utilisant le logiciel *pwcmp*. Les intervalles de confiance ont été trouvés à l'aide du bootstrapping.

Les valeurs JOD résultantes sont rapportées dans figure B.5 pour chaque séquence vidéo. Les vidéos compressées avec trois espaces de couleur ont des valeurs JOD très similaires. En regardant les données rééchelonnées, on peut dire que, dans l'ensemble, il n'y a pas de différence significative entre les performances de compression vidéo en utilisant les espaces colorimétriques testés malgré les différences numériques. De même, les résultats du test de signification statistique et du test binomial conviennent qu'il n'y a pas de différence significative entre les espaces colorimétriques comparés.

Comparaison des scores de qualité objective

En plus des résultats subjectifs, la qualité vidéo a été prédite en utilisant deux mesures objectives de qualité : une mesure objective de qualité pour la vidéo HDR, c.-à-d. HDR-VQM [NMDSLC15], et une mesure de différence de couleur, c.-à-d. ΔE_{2000} [LCR01]. Le HDR-VQM a été calculé en utilisant uniquement le canal de luminance. Les résultats de la



(a) Balloon (b) Bistro (c) Hurdles

(d) Market (e) Starting

Figure B.7 – Les résultats obtenus en comparant toutes les scènes pour les trois espaces couleurs en utilisant la métrique ΔE_{2000}. Plus ΔE_{2000} scores représentent une augmentation de la différence de couleur. Les scores faibles correpondent à des stimuli proches de la vidéo originale.

métrique de qualité HDR-VQM sont montrés dans la figure B.6, et les résultats de ΔE_{2000} sont montrés dans la figure B.7.

En comparant les mêmes stimuli à l'aide de la métrique objective HDR-VQM, nous avons trouvé des résultats presque identiques à l'expérience subjective. Les résultats de ΔE_{2000} sont différents de ce qui a été observé à la fois dans les résultats d'expériences subjectives et dans les résultats métriques objectifs HDR-VQM. Meme si la métrique HDR-VQM est insensible à la couleur, elle donne de meilleures prédiction de la qualité des vidéo compressées que ΔE_{2000}. Le désaccord de ΔE_{2000} avec les scores subjectifs et la capacité du HDR-VQM à prédire la tendance générale des résultats subjectifs indiquent que la qualité perçue de la compression vidéo HDR est dominée par la distorsion structurelle causée par les changements dans le canal de luminance.

B.5 Évaluation des performances des métriques de qualité d'image HDR avec référence

Par rapport à l'évaluation de la qualité SDR, de nouveaux défis apparaissent pour l'évaluation de la qualité visuelle HDR [NdSLC+16b]. Pour l'évaluation de la qualité du contenu HDR, les métriques développées exclusivement pour le contenu HDR [MKRH11, NDSLC15] et les métriques SDR [WSB03, WBSS04, SB06] avec linéarisation perceptuelle sont comparés aux notes d'opinion moyennes (MOS) des sujets dans plusieurs études subjec-

Table B.1 – Nombre d'observateurs, méthodologie subjective, nombre de stimuli, type de compression et correspondance des tons (TMO) utilisés dans les bases de données (DB) de qualité d'image HDR utilisées dans cet article. Légende des TMO : *AS* : Ashikmin, *RG* : Reinhard Global, *RL* : Reinhard Local, *DR* : Durand, *Log* : Logarithmique, *MT* : Mantiuk.

Numéro de DB	**Observateurs**	**Méthodologie**	**Stimuli**	**Compression**	**TMO**
#1 [NDSLCP13]	27	ACR-HR	140	JPEG†	iCAM-06
#2 [NDSLCP14a]	29	ACR-HR	210	JPEG 2000†	AS, RG, RL DR,Log
#3 [KHR+15]	24	DSIS	240	JPEG-XT	RG, MT
#4 [VDSLD14]	15	DSIS	50	JPEG† JPEG 2000† JPEG-XT	Mai
#5	15	DSIS	50	JPEG† JPEG 2000†	Mai PQ

† Les images altérées sont générées par un système de codage scalable [WS06] : l'image HDR est convertie en SDR en utilisant un TMO ; ensuite, l'image SDR est codée & décodée par un codec existant ; enfin, l'image est convertie en gamme HDR.

tives pour des scénarios de compression [VDSLD14, HBP+15, NDSLCP13, NDSLCP12].

Dans le chapitre 4, nous visons à apporter plus de clarté dans ce domaine, en fournissant un benchmark étendu, fiable et cohérent des métriques de fidélité d'image HDR les plus populaires. Nous alignons les valeurs MOS de 5 bases de données et fusionnons les valeurs MOS. Cette base de données alignée se compose d'un total de 690 images HDR compressées. À notre connaissance, il s'agit du plus grand ensemble sur lequel les mesures HDR ont été testées jusqu'à présent au meilleur de nos connaissances. À l'aide de ce vaste ensemble de données, nous analysons la précision de prédiction et la discriminabilité (c.-à-d. la capacité de détecter lorsque deux images ont une qualité perçue différente) de 25 mesures de fidélité, y compris celles qui ont été testées dans le cadre de la normalisation MPEG. Pour l'analyse de la discriminabilité, nous proposons une nouvelle méthode basée sur une approche de classification.

B.5.1 Les bases de données subjectives considérées

Par rapport à la disponibilité des images HDR de haute résolution et de haute qualité sans distorsion [DM04, Fai07, DM08, EMP13, pfs15], le nombre de bases de données (DB - database) de qualité d'image HDR subjectivement annotées et accessibles au public est très faible. Pour cette étude, nous utilisons 5 bases de données différentes. Nous avons sélectionné quatre bases de données d'évaluation de la qualité d'image HDR accessibles au public pour cette analyse. En outre, nous proposons une nouvelle base de données. Chacune de ces bases de données contient des images HDR compressées et des valeurs MOS pour ces images HDR. Les algorithmes de compression, le nombre d'observateurs, le nombre de stimuli utilisés et les méthodologies d'expérimentation sont différents, et ces paramètres

sont résumés dans le tableau B.1.

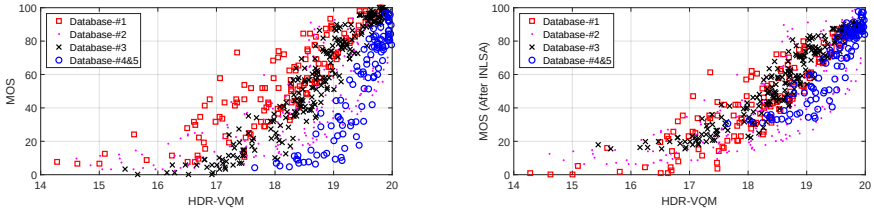
En plus des bases de données mentionnées, nous construisons une nouvelle base de données d'images HDR subjectives de 50 images, dans le prolongement des travaux précédents de Valenzise et al. [VDSLD14]. La nouvelle base de données comprend 5 images originales, sélectionnées de manière à être représentatives des différentes caractéristiques de l'image, y compris la gamme dynamique, la clé d'image et l'information spatiale. Les images de test sont obtenues en utilisant un schéma de codage HDR rétrocompatible [WS06], en utilisant JPEG et JPEG 2000 comme codecs SDR. Pour convertir de HDR à SDR, nous avons utilisé soit le TMO de Mai et al. [MMM+11] soit la courbe PQ-EOTF [MND12, SMP14]. L'environnement de test et la méthodologie ont été soigneusement contrôlés pour être les mêmes que dans DB #4 (Valenzise et al. (2014)) [VDSLD14], et la méthodologie DSIS a été employée. Un panel de 15 personnes (3 femmes, 12 hommes ; âge moyen de 26,8 ans), principalement des doctorants naïfs à la technologie HDR et à la compression d'image, ont participé au test. Les deux bases de données, #4 et #5, sont alignées à l'aide d'un ensemble commun, et l'ensemble de données combiné est appelé DB #4 & 5 tout au long de la thèse.

B.5.2 Alignement des valeurs MOS

Dans de nombreuses expériences subjectives, on demande aux sujets d'utiliser toute la gamme des notes de l'échelle lors de l'évaluation. Toutefois, la qualité du matériel de test pour différentes expériences peut ne pas être la même lorsqu'elles sont comparées les unes aux autres. Dans la figure B.8, nous observons la distribution du MOS pour les bases de données non alignées en fonction de la métrique HDR-VQM. En raison des caractéristiques des expériences et du matériel de test de chaque base de données, un niveau similaire de détérioration d'après l'échelle subjective peut correspondre à des valeurs très différentes d'après les metriques objectives. Par conséquent, afin d'utiliser de manière cohérente les valeurs MOS de différentes bases de données subjectives, celles-ci doivent être ajustées sur une échelle de qualité commune.

Afin d'aligner les valeurs MOS des cinq bases de données d'images HDR, nous utilisons l'algorithme des moindres carrés imbriqués itéré (INLSA - *Iterated Nested Least Square Algorithm*) proposé dans [PW03b]. L'INLSA aligne les valeurs subjectives de qualité collectées dans différentes expériences subjectives en utilisant des variables externes communes, c.-à-d. des résultats métiques objectifs.

L'INLSA exige des paramètres objectifs pour l'alignement, en supposant que ceux-ci sont linéaires et suffisamment bien corrélés par rapport aux MOS. Par conséquent, nous avons analysé les paramètres considérés (décrits ci-dessous) afin de sélectionner les meilleurs candidats pour cette opération. Puisque PCC est un indice de corrélation montrant la linéarité des données et SROCC est un indice de corrélation montrant la monotonie des données, nous avons trouvé les 5 métriques qui ont la valeur la plus élevée pour le produit



(a) HDR-VQM - Avant INLSA (RMSE = 44.10)

(b) HDR-VQM - Après INLSA (RMSE = 42.09)

Figure B.8 – Diagrammes des scores MOS par rapport aux scores objectifs de qualité pour le HDR-VQM avant et après l'alignement INLSA. Afin de comparer quantitativement le diagramme de dispersion, l'erreur quadratique moyenne (RMSE - root mean squared error) des données est rapportée pour chaque cas.

de PCC et SROCC : HDR-VDP-2.2, HDR-VQM, PU-IFC, PU-UQI, PU-UQI et PU-VIF. Les valeurs MOS de l'ensemble des 5 bases de données ont été rassemblées et alignées à l'aide de l'algorithme INLSA à l'aide de cinq métriques d'ancrage sélectionnées. Les diagrammes de dispersion des valeurs MOS par rapport aux valeurs objectives de qualité estimées par HDR-VQM après alignement sont présentés en figure B.8.

Après les observations initiales des images de test, nous remarquons que les images de DB #2 [NDSLCP14a] ont des caractéristiques très différentes par rapport aux autres, et les valeurs MOS sont beaucoup plus dispersées que les autres bases de données après l'alignement. Les déteriorations visuelles dans DB #2 sont très difficiles à prédire pour toutes les métriques de qualité que nous avons pris en compte dans ce travail. Par conséquent, afin de fournir une vue d'ensemble complète de la performance des mesures de fidélité HDR, nous présentons les résultats avec et sans DB #2 dans les évaluations.

B.5.3 Analyse des métriques de qualité objectives

Mesures de qualité objectives à l'étude

Nous incluons dans notre évaluation un certain nombre de mesures de qualité d'image avec référence (FR - full-reference), y compris l'erreur quadratique moyenne (MSE), le rapport signal sur bruit (PSNR), l'indice de similarité structurelle (SSIM) [WBSS04], multi-échelle SSIM (MSSIM) [WSB03], critère de fidélité à l'information (IFC) [SBDV05], indice universel de qualité (UQI) [WB02], et VIF [SB06]. En plus de ces mesures, nous considérons HDR-VDP-2.2 [NMDSLC15], HDR-VQM [NDSLC15], des métriques de référence supplémentaires récemment proposées pour les vidéos HDR telles que mPSNR, tPSNR, CIE ΔE 2000 [TS15], et l'extension spatiale de CIE ΔE 2000 [ZW97] qui est calculée avec le modèle S-CIELAB. Les mesures objectives de la qualité à l'étude peuvent être regroupées comme le montre le tableau B.2.

Pour calculer les mesures de qualité, nous avons d'abord rééchelonné les valeurs des pixels en fonction de la plage de luminance émise par les écrans HDR utilisés dans chaque

Table B.2 – Les mesures de qualité d'image HDR avec référence sont regroupées. Les mots en italique indiquent l'encodage des pixels et les tirets indiquent le préfixe.

	HDR	Écart de couleur	SDR	
Mesures	HDR-VDP-2.2 HDR-VQM mPSNR	$CIE\ \Delta E_{00}$ $CIE\ \Delta E^S_{00}$	*Photometric-*	MSE, PSNR, VIF SSIM, MSSIM IFC, UQI
			PU-	
			Log-	
			PQ	tPSNR-YUV

expérience subjective, une plage de luminance de 0,03 à 4250 cd/m^2. Ensuite, nous utilisons le *modèle linéaire* tel qu'il a été trouvé car la fonction réelle se trouve être suffisamment proche d'être linéaire. [ZVD16].

Analyses statistiques

La performance des mesures de qualité avec référence considérée dans cette étude a été évaluée en termes d'*exactitude*, de *monotonie*, et de *constance* de la prédiction [DS12]. Pour ces catégories, le coefficient de corrélation de Pearson (PCC), l'erreur quadratique moyenne (RMSE), le coefficient de corrélation de Spearman (SROCC) et le rapport des valeurs aberrantes (OR) ont été calculés [ITU12c]. Ces mesures de performance ont été calculées après une régression non linéaire effectuée sur des résultats de mesures objectives de qualité en utilisant une fonction logistique, tel que décrit dans le rapport final de VQEG FR Phase I [RLC+00]. Cette fonction logistique est donnée dans l'équation 4.4 :

Les résultats de ces indices de performance (SROCC, PCC, RMSE et OR) ont été calculés pour chaque base de données séparément, ainsi que pour l'ensemble des données et pour le cas excluant DB #2. Pour déterminer les performances des métriques, nous évaluons l'importance de la différence entre les indices de performance considérés, comme proposé dans la Recommandation ITU-T P.1401 [ITU12c]. Les résultats sont fournis dans les figures B.9 et B.10 pour les cas "Combiné" et "Sauf DB #2" respectivement. Les barres indiquent l'équivalence statistique entre les mesures de qualité.

Nous observons que la performance de HDR-VQM –ainsi que PU-VIF, PU-IFC et Log-IFC– dans la base de données combinée est significativement différente des autres. De plus, nous remarquons que toutes les métriques à l'exception de CIE ΔE_{00} et CIE ΔE^S_{00} ne prennent en compte que les valeurs de luminance. Une étude récente sur les mesures de différence de couleur [OJKP16] ne tient pas compte des artefacts de compression dans les expériences, car l'impact de ceux-ci sur la qualité de l'image a été jugé beaucoup plus fort que les différences de couleur. Ainsi, notre analyse confirme que les artefacts de luminance tels que le blocage, etc. jouent un rôle dominant dans la formation des jugements de qualité, y compris dans le cas HDR.

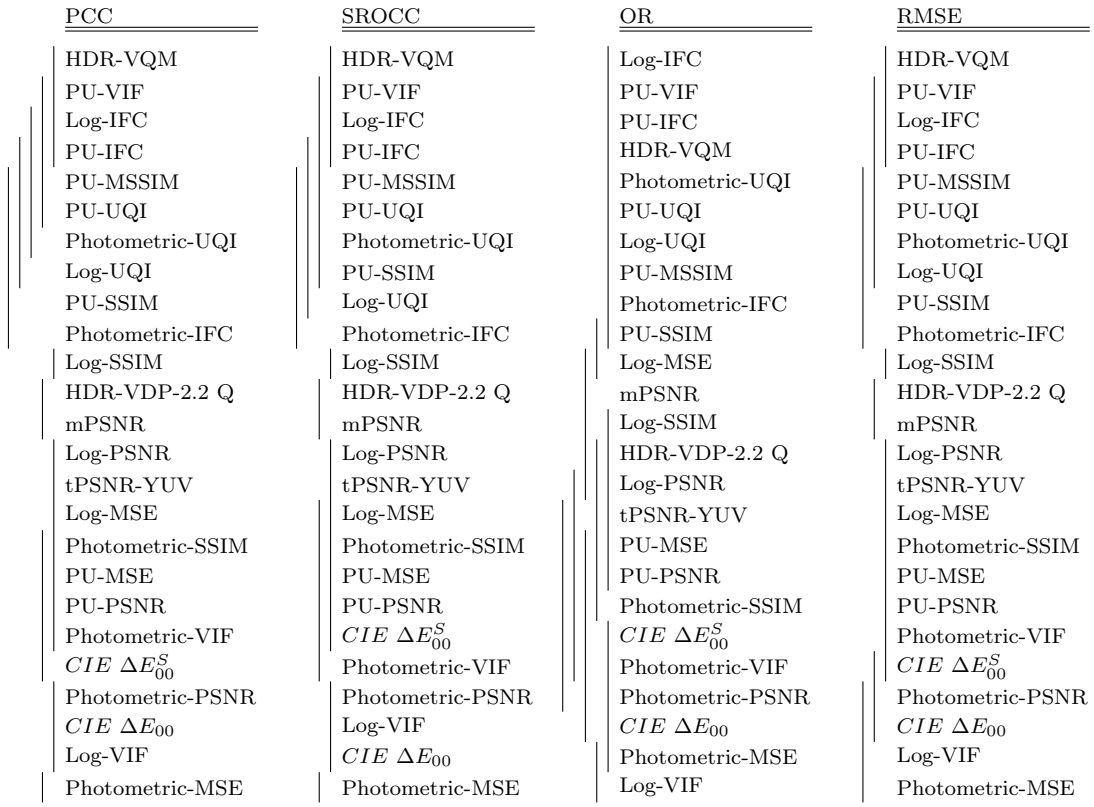


Figure B.9 – Résultats de l'analyse statistique pour les indices de corrélation des données combinées selon la Recommandation ITU-T P.1401 [ITU12c]. Par exemple, il n'y a pas de différence statistiquement significative entre HDR-VQM, PU-VIF, PU-VIF, PU-IFC et Log-IFC en termes de PCC, SROCC, OR et RMSE.

Analyse de la discrimination

Les valeurs MOS sont sont *estimées* à partir d'un échantillon d'observateurs humains, c.-à-d. qu'elles représentent les valeurs attendues de variables aléatoires (la gêne ou la qualité perçue). Par conséquent, les valeurs MOS sont également des variables aléatoires qui sont connues avec une certaine incertitude, qui est typiquement représentée par leurs intervalles de confiance [ITU12b]. Les méthodes d'analyse statistique supposent plutôt que les valeurs MOS sont connues de façon déterministe. Par conséquent, dans ce qui suit, nous considérons une autre approche d'évaluation, qui vise à évaluer si une métrique de qualité objective avec référence est capable de discriminer si deux images ont une qualité subjective significativement différente.

Pour y remédier, Brill et al. [BLC+04] ont introduit le concept de *pouvoir de résolution* d'une métrique objective, qui est la différence minimale dans la qualité prédite de telle sorte qu'au moins $p\%$ de téléspectateurs (par exemple $p = 95\%$) différencient deux images. Cette approche a également été normalisée [ITU04b], et utilisée dans les travaux ultérieurs [PW08, Bar09, HŘE15, NVH16]. Une autre approche de ce problème a été récemment proposée par Krasula et al. [KFLCK16]. Dans leur récent article, Krasula et al. trouvent la précision

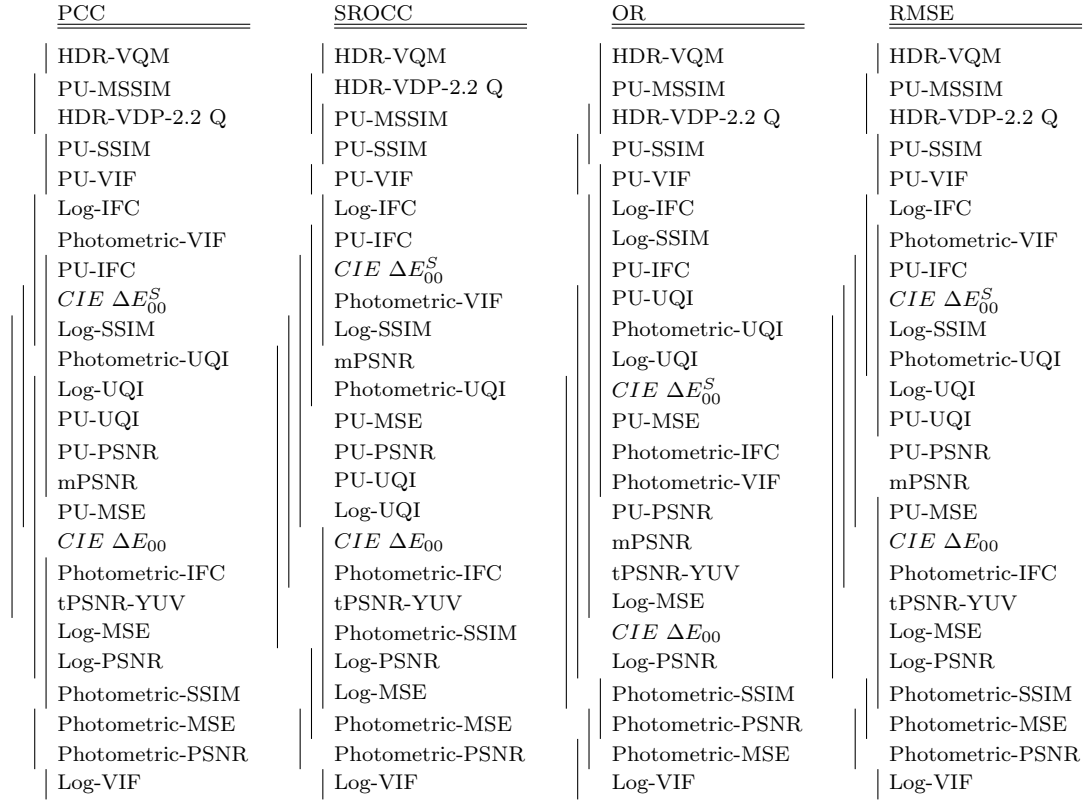


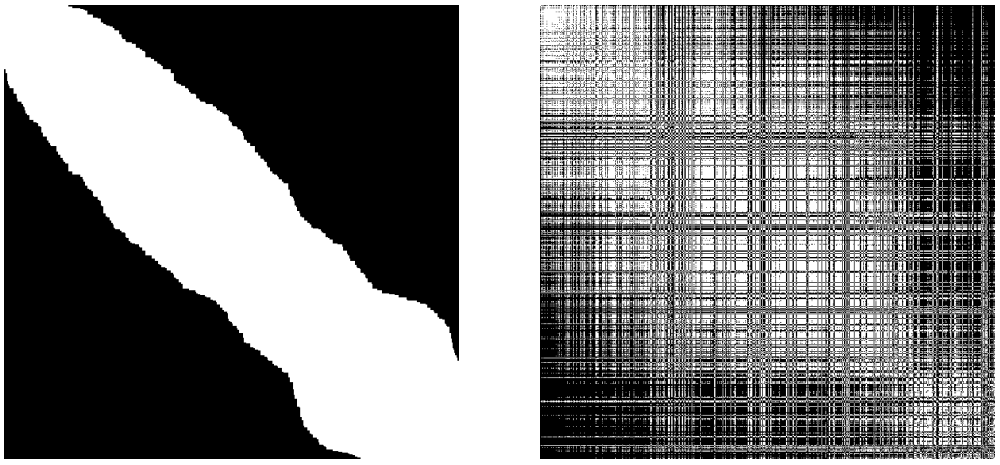
Figure B.10 – Résultats de l'analyse statistique pour les indices de corrélation pour les données combinées à l'exclusion de DB #2 selon la Recommandation ITU-T P.1401 [ITU12c]. Il y a une différence statistiquement significative entre le HDR-VQM et toutes les autres mesures considérées en termes de PCC, SROCC et RMSE.

d'une image objective ou d'une métrique de qualité vidéo en transformant le problème en un problème de classification, et deux analyses sont effectuées : différente vs. similaire, et meilleure vs. pire.

Dans ce chapitre, nous proposons une approche alternative similaire à celle présentée dans Krasula et al. [KFLCK16], qui permet d'évaluer son pouvoir de discrimination. L'idée de base de la méthode proposée est de convertir le problème classique de *régression* de prédiction précise des valeurs MOS, en un problème de *classification binaire* (détection) [Kay98].

Nous dénotons la qualité subjective (MOS) et objective du stimulus I par $S(I)$ et $O(I)$, respectivement. Étant donné deux stimuli I_i, I_j, nous modélisons le problème de détection comme l'une des deux hypothèses $\mathcal{H}_0$, c.-à-d., il n'y a pas de différence significative entre la qualité visuelle de I_i et I_j, et $\mathcal{H}_1$, c.-à-d., I_i et I_j ont une qualité visuelle significativement différente.

Dans notre travail, nous utilisons une analyse de variance à sens unique (ANOVA), avec le critère de différence honnêtement significative de Tukey pour tenir compte du biais de comparaison multiple [HL87], car c'est aussi le moyen idéal pour trouver la signification



(a) Matrice d'équivalence MOS à un niveau de confiance de 95%

(b) HDR-VDP-2.2 Q matrice d'équivalence estimée (τ fixe pour une précision maximale)

Figure B.11 – Cartes d'équivalence pour la base de données combinées (triées). Les entrées blanches correspondent à $S(I_i) \cong S(I_j)$, noir à $S(I_i) \ncong S(I_j)$.

statistique dans [KFLCK16]. La figure B.11.(a) montre les résultats de l'analyse de variance sur notre base de données combinée, avec un seuil de confiance de 95% (c.-à-d. une signification de 5%). Pour faciliter la visualisation, les valeurs MOS ont été triées par ordre croissant avant d'appliquer ANOVA. Les entrées blanches représentent des paires MOS qui sont statistiquement indiscernables.

Pour choisir entre $\mathcal{H}_0$ et $\mathcal{H}_1$, de même que Krasula et al. [KFLCK16], nous considérons la statistique de test simple $\Delta_{ij}^O = |O(I_i) - O(I_j)|$, c.-à-d, nous regardons la différence entre les scores objectifs pour les deux stimuli et nous la comparons avec un seuil τ, c'est-à-dire $\mathcal{H}_0$si $\Delta_{ij}^O \leq \tau$. Pour une valeur donnée de τ, nous pouvons alors étiqueter l'ensemble des stimuli comme étant équivalents ou non, comme le montre la Figure B.11.(b).

Après avoir trouvé des matrices d'équivalence pour les valeurs MOS et les scores des métriques de qualité objective, le problème d'évaluation est converti en un problème de classification binaire. En faisant varier la valeur de τ, on peut tracer une courbe caractéristique de fonctionnement du récepteur (ROC) [Kay98]. L'aire sous la courbe ROC (AUC - area under curve) est plus élevée lorsque le chevauchement entre les distributions marginales de Δ_{ij}^O sous chaque hypothèse $\big(p(\Delta_{ij}^O; \mathcal{H}_0)$ et $p(\Delta_{ij}^O; \mathcal{H}_1)\big)$, est plus petit. Par conséquent, l'AUC est une mesure du *pouvoir de discrimination* d'une mesure objective de la qualité.

Les résultats de cette évaluation de la signification statistique sont présentés figure B.12. Les résultats montrent que HDR-VQM est la métrique la plus performante, et PU-VIF –dans le cas excluant DB #2– PU-MSSIM donnent de meilleurs résultats que la plupart des métriques considérées.

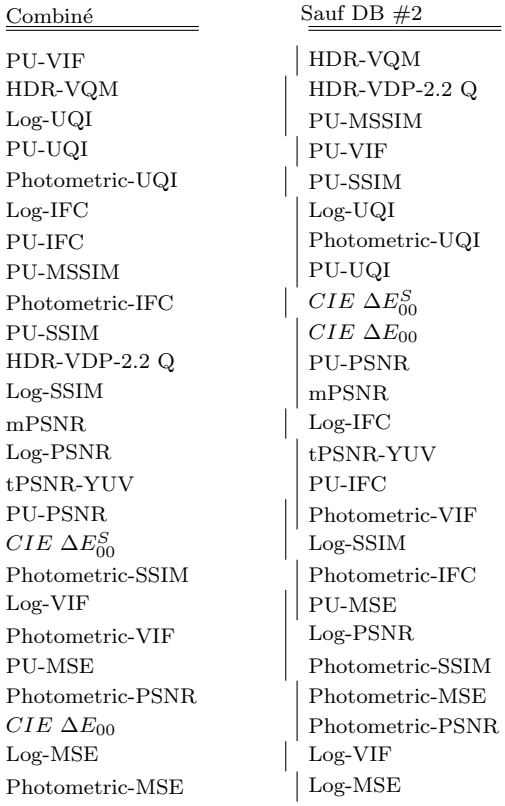


Figure B.12 – Résultats de l'analyse statistique pour l'analyse de la discriminabilité, selon la procédure décrite dans Krasula et al. [KFLCK16]. Les barres signifient l'équivalence statistique entre les métriques de qualité si elles ont la même barre alignée avec deux métriques de qualité. On peut dire que parmi les PU-UQI, Log-UQI et Photometric-UQI, il n'y a pas de différence statistiquement significative. Attendu qu'il existe une différence statistiquement significative entre le HDR-VQM et toutes les autres mesures considérées.

B.6 La relation entre MOS et les comparaisons par paires

Dans le chapitre précédent, l'évaluation des mesures objectives de qualité FR HDR montre que l'alignement des ensembles de données subjectives est délicat et non direct. Cette observation nous amène à nous poser les questions suivantes : Comment trouver une méthode plus robuste pour aligner des bases de données de qualité ? Qu'est-ce qu'une bonne mesure à utiliser pour comparer la qualité de deux stimuli différents ? Pour répondre à ces questions, nous devons d'abord comprendre ce qui cause la variance de la qualité perçue, et nous devrions être en mesure de réduire la variance.

En comparant quatre méthodologies différentes, Mantiuk et al. [MTM12] a constaté que la méthodologie des comparaisons par paires (PC - Pairwise Comparisons) à choix forcé était la plus précise et la plus efficace en termes de temps. L'expérience PC et les résultats du rééchelonnement de PC sont beaucoup moins influencés par les facteurs humains de par leur nature. Ainsi, il peut être utilisé comme une échelle "universelle" à laquelle vous pouvez aligner vos ensembles de données.

Dans le chapitre 5, nous essayons de comprendre la relation entre les méthodes de notation directe –c.-à-d. MOS– et de classement –c.-à-d. le rééchelonnement de PC– et nous comparons les valeurs MOS et les résultats du rééchelonnement des données expérimentales de PC. Nous étudions également l'effet de l'ajout de comparaisons de contenu croisé, en montrant que cela permet non seulement d'unifier l'échelle de qualité à travers le contenu, mais aussi d'améliorer significativement la précision des scores de qualité de rééchelonnement.

B.6.1 Rééchelonnement des données de comparaisons par paires

Les résultats d'une expérience de comparaison par paires peuvent être rassemblés dans une matrice des préférences, également connue sous le nom de matrice de comparaison. Ses éléments contiennent le décompte du nombre de fois qu'une condition est votée aussi bien que l'autre. Ces matrices de préférences peuvent être utilisées pour trouver un score de qualité pour chaque condition en utilisant l'une de plusieurs méthodes *rééchelonnement* [BT52, Thu27, LDSE11, TG11]. Dans ce travail, nous utilisons *pwcmp*, un logiciel open source pour mettre à l'échelle les résultats des comparaisons par paires [POM17]. Comme décrit précédemment, ce logiciel estime les scores de qualité à l'aide d'une méthode bayésienne, qui utilise un estimateur de probabilité maximale pour maximiser la probabilité que les données recueillies expliquent les scores de qualité sous les hypothèses du scénario V de Thurstone.

Les résultats des comparaisons par paires sont généralement rééchelonnés en *Just-Noticeable-Difference (JND)* unités [Eng00, SF01]. Deux stimuli sont espacés de 1 JND si 50% des observateurs peuvent voir la différence entre eux. Pour faire référence à la différence de qualité d'image par rapport à une référence de qualité parfaite, nous décrivons cette mesure de qualité comme *Just-Objectionable-Differences (JODs)* [AVS+17] plutôt que JNDs.

B.6.2 La relation entre MOS et les comparaisons par paires

Détails des expériences subjectives

Afin de comparer les valeurs MOS aux résultats du rééchelonnement de PC, deux tests subjectifs ont été effectués. Dans ces tests subjectifs, nous avons utilisé l'ensemble de données de qualité vidéo HDR créé dans la section 3.2, qui consiste en 60 vidéos HDR compressées. Les expériences ont été menées dans une pièce silencieuse et sombre, conformément aux Recommandations de l'ITU [ITU12b, ITU98]. Un écran SIM2 HDR47 calibré avec une résolution de 1920×1080 pixel a été utilisé dans son mode de rendu intégré natif. La distance des sujets par rapport à l'écran était fixée à trois hauteurs d'affichage [ITU98]. Les expériences réalisées partagent un ensemble de paramètres communs en plus de ceux de la salle de test. Les stimuli ont été présentés en paires avec une représentation côte à côte.

La première expérience menée était une expérience de comparaisons par paires avec un design incomplet. Dans cette expérience, une paire de vidéos avec deux débits binaires consécutifs provenant du même espace couleur ou avec le même débit binaire provenant de deux espaces couleur différents a été comparée, comme cela a été fait dans la section B.4.2. Afin d'analyser les résultats des comparaisons par paires et de comprendre si ces résultats du rééchelonnement sont comparables aux scores de qualité, une deuxième expérience a été menée selon la méthodologie de double stimulus impairment scale (DSIS). Dans cette deuxième expérience, la méthodologie DSIS Variante I avec une présentation côte à côte a été utilisée, comme dans [HRE16]. Toutes les vidéos déformées ont été comparées avec la référence non déformée.

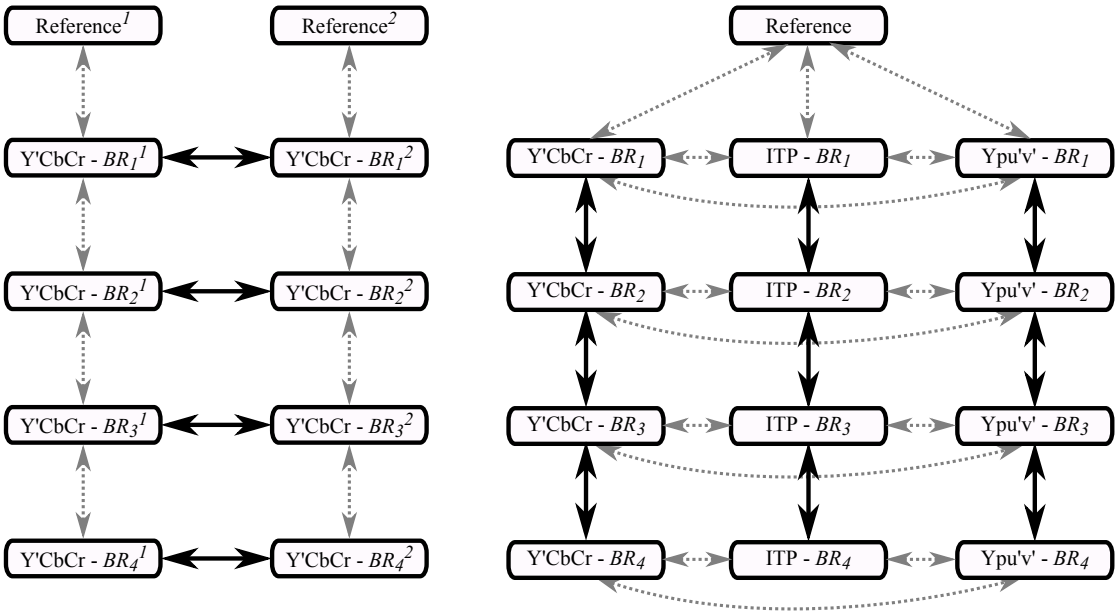
Table B.3 – Linéarité de la relation entre MOS et JOD

Séquence	PCC	SROCC
Balloon	0.9936	0.9835
Bistro	0.9824	0.9890
Hurdles	0.9864	0.9670
Market	0.9696	0.9615
Starting	0.9897	0.9835
Tous les contenus	0.9337	0.9420

Comparaison des MOS et comparaisons par paires

Les matrices de préférence des expériences PC ont été trouvées, et les scores JOD ont été estimés à l'aide du logiciel *pwcmp*. Pour l'expérience DSIS, les valeurs MOS ont été calculées en prenant la moyenne des scores d'opinion. Les intervalles de confiance (CI), par contre, ont été calculés à l'aide du bootstrapping afin de les comparer aux CI des scores JOD. Les scores JOD résultants (notés $JOD_{Standard}$ pour indiquer que la méthodologie des comparaisons par paires standard a été utilisée) ont été tracés par rapport aux valeurs MOS. Les résultats montrent qu'il existe une forte relation entre les valeurs MOS et les scores JOD, et les scores JOD et les valeurs MOS montrent un comportement presque linéaire pour tous les contenus. Cette relation a également été vérifiée avec les calculs PCC et SROCC. Rapporté dans le tableau B.3, les valeurs PCC et SROCC montrent que la relation est presque parfaitement linéaire pour chaque séquence vidéo.

Nous avons remarqué que les valeurs MOS ont des CI proches de l'uniforme ; cependant, les CI des valeurs JOD augmentent au fur et à mesure que les valeurs absolues JOD elles-mêmes augmentent. Cela a été causé par l'accumulation des erreurs d'estimation, qui résulte de la comparaison de paires consécutives.



(a) Expérience supplémentaire de contenu croisé

(b) Expérience supplémentaire sur le même contenu

Figure B.13 – Designs d'expérience pour deux expériences supplémentaires. Les paires supplémentaires sélectionnées sont affichées avec des flèches noires, où Reference^i est la référence (original) pour le contenu vidéo i, BR_j^i est le contenu vidéo i compressé avec le j-ième bitrate ($j = 1$ est le bitrate le plus élevé). Les paires illustrées par des flèches grises en pointillés sont les paires comparées pour les comparaisons par paires standard, comme décrit dans la section 5.2.1. Pour éviter l'encombrement, les comparaisons entre les espaces colorimétriques ne sont pas affichées dans la sous-figure (a).

B.6.3 Extension de PC: Comparaisons inter-contenus

Expérience de comparaison par paires inter-contenus

Les expériences de comparaison par paires sont conçues pour comparer des conditions provenant du même contenu, de sorte que les "pommes" sont comparées aux "pommes". Étant donné que seule la condition de référence est ancrée et que la qualité de toutes les autres conditions est estimée à partir des relations de paires, l'erreur d'estimation est accumulée en s'éloignant du point d'ancrage. Au lieu de cela, comparer des "pommes" à des "oranges" peut introduire de nouvelles informations, améliorer le rééchelonnement et réduire la variance de contenu croisé.

En plus des expériences subjectives décrites dans la section précédente, deux autres expériences ont été menées afin d'analyser et de comprendre les effets des comparaisons croisées par paires. Toutes les variables, à l'exception de la paire de stimuli sélectionnée, sont restées les mêmes. Nous avons été motivés à mener une telle expérience de comparaison de contenu croisé après avoir observé que de telles comparaisons sont effectuées indirectement dans la méthodologie DSIS. Lorsque les téléspectateurs évaluent les séquences, ils jugent la qualité par rapport à toutes les autres séquences qu'ils ont vues, ainsi que les séquences

présentant un contenu différent.

Pour que les expériences supplémentaires soient courtes, nous avons jumelé des vidéos avec des contenus différents et un même débit binaire, comme le montre la figure B.13.(a) en utilisant les comparaisons montrées avec des flèches noires pleines. Les résultats obtenus ont été combinés avec les résultats de l'expérience de comparaison par paires de l'étalon (même contenu) (illustrés par des flèches grises dans la figure B.13) et rééchelonné à nouveau en utilisant le même logiciel *pwcmp*. Les résultats sont présentés dans la section correspondante ci-dessous. Les nouveaux scores JOD obtenus à partir de la combinaison de PC standard et de l'expérience de contenu croisé sont appelés $JOD_{CrossContent}$.

Afin d'établir une comparaison équitable en ce qui concerne le nombre total de comparaisons, une expérience supplémentaire sur le même contenu a également été menée. Afin de maintenir le nombre de paires supplémentaires dans une fourchette similaire, le test a consisté en un total de 90 paires (y compris les versions en miroir). Pour cela, nous avons sélectionné des paires avec des débits binaires consécutifs et des espaces de couleur identiques, comme indiqué par des flèches noires pleines dans la figure B.13.(b). Il s'agit essentiellement d'observations supplémentaires de certaines paires du test PC standard décrit dans la section précédente. Ces paires ont été comparées par 15 personnes (8 hommes et 7 femmes) avec une moyenne d'âge de 29 ans. Ces paires supplémentaires de même contenu ont de nouveau été combinées avec les résultats de l'expérience PC standard et mises à l'échelle avec le logiciel *pwcmp*. Les scores JOD obtenus à partir de la combinaison d'un PC standard et d'une expérience supplémentaire de même contenu sont appelés $JOD_{SameContent}$.

Impact des comparaisons inter-contenus

Les valeurs JOD que nous utilisons ont été trouvées en utilisant trois ensembles différents de données PC. Comme décrit dans la section précédente, $JOD_{Standard}$ a été trouvé en utilisant les données acquises dans l'expérience PC avec le même contenu. $JOD_{SameContent}$ a été trouvé en utilisant la combinaison des résultats de l'expérience PC standard et des résultats de l'expérience de même contenu, et $JOD_{CrossContent}$ a été trouvé en utilisant la combinaison des résultats de l'expérience PC standard et des résultats de l'expérience de contenu croisé.

Le comportement linéaire observé entre les valeurs $JOD_{Standard}$ et MOS s'applique également aux cas $JOD_{SameContent}$ et $JOD_{CrossContent}$. De plus, l'introduction de paires de contenus croisés augmente la corrélation et la linéarité de la relation entre JOD et MOS.

On calcul la pente de la droite la mieux adaptée à chaque contenu, en utilisant les tracés de MOS vs JOD. Afin de trouver l'effet de l'addition de paires de contenus croisés, la variance de ces pentes a été également calculée. La variance des pentes dans le cas de $JOD_{SameContent}$ était de $2,7972$ et dans le cas de $JOD_{CrossContent}$ était de $0,6445$. Cette réduction significative de la variance des pentes implique que le meilleur ajustement linéaire

Contenu		$CI_{Standard}$	$CI_{SameContent}$	$CI_{CrossContent}$	$Ratio_{CC/SC}$
Balloon	BR_1	1.23	1.23	1.53	1.25
	BR_2	2.21	1.68	1.86	1.11
	BR_3	3.03	2.84	2.48	0.87
	BR_4	3.93	3.36	2.56	0.76
Bistro	BR_1	1.60	1.70	1.25	0.73
	BR_2	2.12	1.91	1.46	0.76
	BR_3	2.92	2.49	2.00	0.81
	BR_4	3.34	2.91	2.26	0.78
Hurdles	BR_1	1.45	1.50	1.12	0.75
	BR_2	2.31	1.90	1.55	0.82
	BR_3	3.12	2.36	2.46	1.04
	BR_4	3.43	2.96	2.62	0.89
Market	BR_1	2.12	2.35	0.85	0.36
	BR_2	3.05	2.80	1.63	0.58
	BR_3	4.32	3.18	2.57	0.81
	BR_4	4.73	3.28	2.94	0.90
Starting	BR_1	3.52	3.50	1.29	0.37
	BR_2	4.45	4.06	1.47	0.36
	BR_3	5.61	4.76	1.97	0.41
	BR_4	6.04	5.11	2.29	0.45

Table B.4 – Intervalle de confiance moyen des vidéos avec différents débits binaires (BR_1 est le plus élevé) pour les expériences considérées. La dernière colonne est le rapport du CI des données PC combinées avec des paires de contenu croisé supplémentaires ($CI_{CrossContent}$, CI de $JOD_{CrossContent}$) au CI des données PC combinées avec des paires de contenu identique supplémentaires ($CI_{SameContent}$, CI de $JOD_{SameContent}$). Les CI de l'expérience PC standard ($CI_{Standard}$, CI de $JOD_{Standard}$) sont également rapportés par souci d'exhaustivité.

pour chaque contenu est beaucoup plus proche et qu'il y a moins de variance entre les différents contenus.

Une autre métrique, Std_{p2l}, a été calculée pour les points tracés dans chaque sous-figure présentée. Elle est calculée comme suit :

$$Std_{p2l} = \sqrt{mean(d(P,l)^2)} \tag{B.7}$$

où $d(P,l)$ est la distance perpendiculaire du point P à la ligne l. Std_{p2l} a été calculée pour le meilleur ajustement linéaire du cas 'Tous ensemble'. Le meilleur ajustement linéaire correspond à la ligne violette en pointillés dans les sous-figures (a) à (e). Il est clair que l'ajout de paires de contenu croisé diminue la variance des pentes de la ligne la mieux ajustée à chaque contenu et Std_{p2l} également, rapprochant ainsi les scores JOD sur une échelle de qualité commune.

Afin d'analyser la variation de CI, les valeurs moyennes de CI sont rapportées dans le tableau B.4. Comme le CI ne change pas beaucoup par rapport à l'espace couleur, la

moyenne des valeurs CI a été calculée pour le même débit binaire. La dernière colonne du table B.4 montre que les CIs diminuent pour presque tous les cas jusqu'à 60%, surtout à des débits binaires plus élevés où l'erreur de mise à l'échelle s'accumulerait plutôt dans le PC standard. Avec les comparaisons de contenu croisé, la taille de CI devient plus uniforme entre les différents niveaux de qualité.

Tous les résultats indiquent que le rééchelonnement des données des comparaisons par paires donne des scores JOD fortement corrélés aux valeurs MOS acquises dans l'expérience DSIS. L'introduction de paires de contenu croisé rend JOD plus uniforme et réduit les intervalles de confiance.

B.7 Conclusions et travaux futurs

Dans cette thèse, nous avons abordé certaines des limites et des défis de l'évaluation de la qualité dans le contexte de l'image et de la vidéo à haute gamme dynamique. Plus précisément, le but de cette thèse était d'étudier les nouvelles conditions de la technologie d'affichage HDR et de fournir un aperçu de l'évaluation et de l'analyse de la qualité vidéo HDR. Pour ce faire, nous avons examiné trois aspects de l'évaluation de la qualité HDR.

Tout d'abord, nous avons analysé les paramètres affectant l'évaluation subjective et objective de la qualité HDR afin de comprendre l'influence des nouvelles conditions introduites par la technologie HDR, et à cette fin, nous avons développé un algorithme de rendu d'image HDR. Dans cette partie, nous nous sommes concentrés sur les effets du rendu de l'affichage (liés à la luminosité et au contraste de l'affichage) et de la couleur sur l'évaluation de la qualité HDR.

Deuxièmement, à partir de nos observations, nous avons évalué les méthodes objectives d'évaluation de la qualité des images HDR à l'aide d'un ensemble de données de 690 images créé en alignant les valeurs MOS de différentes bases de données, et nous avons proposé une nouvelle méthode d'analyse de la discriminabilité basée sur la classification pour l'évaluation de la performance métrique objective.

Troisièmement, nous avons comparé les résultats des comparaisons par paires avec les valeurs MOS, dans l'intention de trouver une représentation commune pour aligner les ensembles de données de qualité et d'éliminer de l'étape d'alignement jusqu'alors nécessaire. De plus, nous avons proposé d'inclure des comparaisons de contenu croisé à la méthodologie des comparaisons par paires afin de réduire la variance de contenu croisé et les intervalles de confiance des résultats du rééchelonnement de PC.

La commercialisation rapide de la technologie HDR/WCG et l'augmentation du volume de contenu HDR ouvrent de nouvelles perspectives pour la recherche future. Nous pensons que certains aspects de la technologie HDR/WCG nécessitent des recherches plus approfondies, et nous décrivons un certain nombre d'extensions possibles de cette thèse.

Bien que la méthode de rendu proposée soit capable d'estimer avec précision la luminance émise, elle ne peut pas estimer la chrominance émise pour le moment. Son rendement inclut

l'information de couleur, mais l'information de couleur est acquise simplement en divisant l'image de couleur HDR à la valeur de rétro-éclairage. Ainsi, la méthode de rendu peut être étendue à l'estimation de la chrominance.

Dans le cas de la compression, les distorsions structurelles créées par les différences de canal de luminance sont dominantes par rapport à la perception humaine de la qualité de l'image et de la vidéo HDR. Cependant, dans les cas autres que la compression, les changements de couleur peuvent encore influencer la qualité perceptuelle. Les artefacts de couleur peuvent être créés pour plusieurs raisons telles que les conversions d'espace ou de gamme de couleurs et les conversions EOTF. Par conséquent, une gamme plus large de distorsions de couleur peut être étudiée pour comprendre les effets des artefacts de couleur dans le sens de HDR/WCG.

L'évaluation des paramètres de qualité d'image HDR dans le chapitre 4 a d'importants résultats et conclusions, dont certains peuvent être étendus pour le cas de la vidéo. Dans cette thèse, nous n'avons pas pu procéder à une telle évaluation pour la vidéo HDR en raison de la rareté des bases de données de qualité vidéo HDR accessibles au public au moment de cette étude. Par conséquent, une évaluation similaire peut être faite pour le cas de la vidéo HDR, afin de prendre en compte les caractéristiques temporelles.

Comme nous l'avons vu précédemment, nous pensons que JOD peut être un score de qualité subjectif universel grâce à la facilité et à la robustesse de son calcul. Bien que les premiers résultats soient prometteurs, cette affirmation doit être validée par un plus grand nombre de données avec des distorsions et des niveaux de qualité variés. Pour autant qu'elles soient validées, les valeurs JOD peuvent être utilisées pour aligner différentes bases de données, ce qui peut améliorer à la fois l'évaluation et l'élaboration de mesures objectives de la qualité HDR.

Bibliography

[AA04] M. Aggarwal and N. Ahuja, "Split aperture imaging for high dynamic range". *International Journal of Computer Vision*, vol. 58 (7), 2004. *Cited in Sec.* 1.3.1

[ABA05] A. Aksay, C. Bilen, and G. B. Akar, "Subjective evaluation of effects of spectral and spatial redundancy reduction on stereo images", in *13th European Signal Processing Conference (EUSIPCO)*, IEEE, 2005. *Cited in Sec.* 1.1

[ABDD+14] M. Azimi, A. Banitalebi-Dehkordi, Y. Dong, M. T. Pourazad, and P. Nasiopoulos, "Evaluating the performance of existing full-reference quality metrics on high dynamic range (HDR) video content", in *16th International Conference on Multimedia Signal Processing (ICMSP)*, 2014. *Cited in Sec.* 2.1.3, 2.1.4

[AFR+07] A. O. Akyüz, R. Fleming, B. E. Riecke, E. Reinhard, and H. H. Bülthoff, "Do HDR displays support LDR content?: A psychophysical evaluation". *ACM Transactions on Graphics*, vol. 26 (3), Jul 2007. *Cited in Sec.* 1.4.1, 2

[AMMS08] T. O. Aydin, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Dynamic range independent image quality assessment". *ACM Transactions on Graphics*, vol. 27 (3), p. 69, Aug 2008. *Cited in Sec.* 1.3.2, 1.4.2, 2, 2.2, B.3.2

[AMR+15] A. Artusi, R. K. Mantiuk, T. Richter, P. Hanhart, P. Korshunov, M. Agostinelli, A. Ten, and T. Ebrahimi, "Overview and evaluation of the JPEG XT HDR image compression standard". *Journal of Real-Time Image Processing*, 2015. *Cited in Sec.* 1.3.2, 1.4.1, 1.4.2, 4.1

[AMR+16] A. Artusi, R. K. Mantiuk, T. Richter, P. Korshunov, P. Hanhart, E. T., and M. Agostinelli, "JPEG XT: A compression standard for HDR and WCG images [Standards in a nutshell]". *IEEE Signal Processing Magazine*, vol. 33 (2), pp. 118–124, Mar 2016. *Cited in Sec.* 1.3.2

[AMS08] T. O. Aydin, R. Mantiuk, and H.-P. Seidel, "Extending quality metrics to full luminance range images", in *Electronic Imaging, Human Vision and Electronic Imaging XIII*, pp. 68060B–1–68060B–10, International Society for Optics and Photonics, Mar 2008. *Cited in Sec.* (document), 1.4.1, 1.4.2, 2, 2.1.2, 2.2, 2.2.2, 4, 4.1, 4.3.1, B.1, B.3.2

[ASAU12] Y. Aksoy, O. Sener, A. Alatan, and K. Ugur, "Interactive 2D-3D image conversion for mobile devices", in *IEEE International Conference on Image Processing (ICIP)*, IEEE, 2012. *Cited in Sec.* 1.1

[Ash02] M. Ashikhmin, "A tone mapping algorithm for high contrast images", in *Proceedings of the 13th Eurographics workshop on Rendering*, pp. 145–156, Eurographics Association, 2002. *Cited in Sec.* 1.3.3, 4.1

[Atk12] R. Atkins, *Advanced Methods for Controlling Dual Modulation Display Systems*, Master's thesis, The University Of British Columbia, 2012. *Cited in Sec.* 1.3.3

[AVS+17] V. K. Adhikarla, M. Vinkler, D. Sumin, R. K. Mantiuk, K. Myszkowski, H.-P. Seidel, and P. Didyk, "Towards a quality metric for dense light fields", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. *Cited in Sec.* 5.1, B.6.1

[BADC11] F. Banterle, A. Artusi, K. Debattista, and A. Chalmers, *Advanced High Dynamic Range Imaging: Theory and Practice*, AK Peters (CRC Press), Natick, MA, USA, 2011. *Cited in Sec.* (document), 1.3.3, 2.1.4, B.1

[Bar99] P. G. Barten, *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*, vol. 72, SPIE Press, 1999. *Cited in Sec.* 4.1

[Bar09] M. Barkowsky, *Subjective and Objective Video Quality Measurement in Low-Bitrate Multimedia Scenarios*, Verl. Dr. Hut, München, 2009. *Cited in Sec.* 1.2.3, 4.3.3, B.5.3

[BCMD17a] C. Bist, R. Cozot, G. Madec, and X. Ducloux, "QoE-based brightness control for HDR displays", in *9th International Conference on Quality of Multimedia Experience (QoMEX)*, 2017. *Cited in Sec.* 1.4.1

[BCMD17b] ———, "Tone expansion using lighting style aesthetics". *Computers & Graphics*, vol. 62, 2017. *Cited in Sec.* 1.1

[BCPLC09] F. Boulos, W. Chen, B. Parrein, and P. Le Callet, "Region-of-interest intra prediction for H.264/AVC error resilience", in *IEEE International Conference on Image Processing (ICIP)*, IEEE, 2009. *Cited in Sec.* 1.1

[BCTB13] R. Boitard, R. Cozot, D. Thoreau, and K. Bouatouch, "Temporal coherency in video tone mapping, a survey", in *HDRi2013 - First International Conference and SME Workshop on HDR imaging*, 2013. *Cited in Sec.* 1.3.3

[BDAPN14] A. Banitalebi-Dehkordi, M. Azimi, M. T. Pourazad, and P. Nasiopoulos, "Compression of high dynamic range video using the HEVC and H.264/AVC standards", in *10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine)*, pp. 8–12, IEEE, 2014. *Cited in Sec.* 2.1.3, 2.1.4

[BFSS17] F. Bossen, D. Flynn, K. Sharman, and K. Sühring, "HM software manual", Tech. Rep., Joint Collaborative Team on Video Coding (JCT-VC), 2017. *Cited in Sec.* 3.1.1, 3.2.1, B.4.1

[BKH03] R. Bogart, F. Kainz, and D. Hess, "The OpenEXR image file format", in *Proceedings of ACM SIGGRAPH, Sketches & Applications*, 2003. *Cited in Sec.* 1.3.1

[BLC+04] M. H. Brill, J. Lubin, P. Costa, S. Wolf, and J. Pearson, "Accuracy and cross-calibration of video quality metrics: New methods from ATIS/T1A1". *Signal Processing: Image Communication*, vol. 19 (2), pp. 101–107, 2004. *Cited in Sec.* (document), 1.2.3, 4.3.3, 4.7, 4.4, B.5.3

[BLCCC09] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images". *EURASIP Journal on Image and Video Processing*, vol. 2008 (1), 2009. *Cited in Sec.* 1.1

[BMN+14] N. Burini, C. Mantel, E. Nadernejad, J. Korhonen, S. Forchhammer, and J. M. Pedersen, "Block-based gradient descent for local backlight dimming and flicker reduction". *Journal of Display Technology*, vol. 10 (1), pp. 71–79, Jan 2014. *Cited in Sec.* 1.3.3, 2.1.2, 2.1.3

[BNK+12] N. Burini, E. Nadernejad, J. Korhonen, S. Forchhammer, and X. Wu, "Image dependent energy-constrained local backlight dimming", in *IEEE International Conference on Image Processing (ICIP)*, IEEE, 2012. *Cited in Sec.* 1.3.3, 2.1.2

[BNK+13] ———, "Modeling power-constrained optimal backlight dimming for color displays". *Journal of Display Technology*, vol. 9 (8), pp. 656–665, Aug 2013. *Cited in Sec.* 1.3.3, 2.1.2

[Bol14] M. Boliek, "Information technology - the JPEG 2000 image coding system: Part 1 (amendment 8)", ISO/IEC IS 15444-1/ ITU-TT.800, 2014. *Cited in Sec.* 1.3.2

[Bor14] T. Borer, "Non-linear opto-electrical transfer functions for high dynamic range television", In: BBC Research & Development White Paper - WHP 283. `http://downloads.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP283.pdf`, 2014. *Cited in Sec.* 1.3.2, B.1

[BOSB10] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution", in *20th international conference on Pattern recognition (ICPR)*, pp. 3121–3124, IEEE, 2010. *Cited in Sec.* 4.3.3

[BPLC+11] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin, "Towards a new quality metric for 3-d synthesized view assessment". *IEEE Journal of Selected Topics in Signal Processing*, vol. 5 (7), pp. 1332–1343, 2011. *Cited in Sec.* 5

[BT52] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons." *Biometrika*, vol. 39 (3/4), pp. 324–345, 1952. *Cited in Sec.* 1.1, 3.2.2, 5.1, B.4.2, B.6.1

[CCLS15] S. Cha, T. Choi, H. Lee, and S. Sull, "An optimized backlight local dimming algorithm for edge-lit LED backlight LCDs". *Journal of Display Technology*, vol. 11 (4), pp. 378–385, 2015. *Cited in Sec.* 1.3.3, 2.1.2

[CCPP15] M. Calemme, M. Cagnazzo, and B. Pesquet-Popescu, "Contour-based depth coding: A subjective quality assessment study", in *IEEE International Symposium on Multimedia (ISM)*, pp. 295–300, IEEE, Dec 2015. *Cited in Sec.* 1.1

[CDLN07] F. Crete, T. Dolmiere, Ladret, and M. Nicolas, "The blur effect: Perception and estimation with a new no-reference perceptual blur metric", in *Electronic Imaging, Human Vision and Electronic Imaging XII*, International Society for Optics and Photonics, 2007. *Cited in Sec.* 1.1

[CFL+15] H. Chang, O. Fried, Y. Liu, S. DiVerdi, and A. Finkelstein, "Palette-based photo recoloring". *ACM Transactions on Graphics*, vol. 34 (4), Jul 2015. *Cited in Sec.* 1.1

[CH07] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images". *IEEE Transactions on Image Processing*, vol. 16 (9), pp. 2284–2298, 2007. *Cited in Sec.* (document)

[CHA+14] A. Chapiro, S. Heinzle, T. O. Aydin, S. Poulakos, M. Zwicker, A. Smolic, and M. Gross, "Optimizing stereo-to-multiview conversion for autostereoscopic displays". *Computer Graphics Forum (CGF)*, vol. 33 (2), pp. 63–72, 2014. *Cited in Sec.* 1.1

[CHS+93] K. Chiu, M. Herf, P. Shirley, S. Swamy, C. Wang, and K. Zimmerman, "Spatially nonuniform scaling functions for high contrast images", in *In Proceedings of Graphics Interface 93*, pp. 245–253, 1993. *Cited in Sec.* 1.3.3

[CIE86] CIE, "Colorimetry", Tech. Rep., International Commission on Illumination, 1986, publication No. 15.2. *Cited in Sec.* 1.2.1

[CIE95] ———, "Industrial color difference evaluation", Tech. Rep., International Commission on Illumination, 1995, publication No. 116. *Cited in Sec.* 1.2.1

[CK09] H. Cho and O.-K. Kwon, "A backlight dimming algorithm for low power and high image quality LCD applications". *IEEE Transactions on Consumer Electronics*, vol. 55 (2), pp. 839–844, 2009. *Cited in Sec.* 1.3.3, 2.1.2

[Dal93] S. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity", in A. B. Watson, ed., *Digital Images and Human Vision*, pp. 179–206, MIT Press, Cambridge, MA, USA, 1993. *Cited in Sec.* 1.4.2

[DD02] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images". *ACM Transactions on Graphics*, vol. 21 (3), pp. 257–266, 2002. *Cited in Sec.* 4.1

[DLCMM16] F. Dufaux, P. Le Callet, R. Mantiuk, and M. Mrak, *High Dynamic Range Video: From Acquisition, to Display and Applications*, Academic Press, 2016. *Cited in Sec.* (document), 4, B.1

[DM97] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs", in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pp. 369–378, ACM, 1997. *Cited in Sec.* 1.3.1, 2.1.4

[DM04] F. Drago and R. Mantiuk, "MPI HDR image gallery", `http://resources.mpi-inf.mpg.de/hdr/gallery.html`, 2004, accessed: 2015-11-15. *Cited in Sec.* 4.1, B.5.1

[DM08] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs", in *ACM SIGGRAPH 2008 classes*, ACM, 2008. *Cited in Sec.* 4.1, B.5.1

[DMAC03] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive logarithmic mapping for displaying high contrast scenes". *Computer Graphics Forum*, vol. 22 (3), pp. 419–426, 2003. *Cited in Sec.* 1.3.3

[Dol17] Dolby, "Dolby vision: An introduction to dolby vision", `https://www.dolby.com/us/en/technologies/dolby-vision/dolby-vision-white-paper.pdf`, 2017. *Cited in Sec.* 1.3.2

[DS12] F. De Simone, *Selected Contributions on Multimedia Quality Evaluation*, Ph.D. thesis, École Polytechnique Fédérale de Lausanne, 2012. *Cited in Sec.* 1.1, 4.1, 4.2, 4.3.2, 4.4, B.5.3

[DSNT+09] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel", in *1st International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 204–209, IEEE, 2009. *Cited in Sec.* 1.1, 5

[DSTN+10] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A H.264/AVC video database for the evaluation of quality metrics", in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2010. *Cited in Sec.* 1.1

[DVKG+00] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model". *IEEE Transactions on Image Processing*, vol. 9 (4), pp. 636–650, 2000. *Cited in Sec.* (document)

[EBU03] EBU, "SAMVIQ - subjective assessment methodology for video quality", Tech. Rep., European Broadcasting Union, 2003, BPN 056. *Cited in Sec.* 1.1

[EMP13] M. T. EMPA, "Empa HDR image database", `http://empamedia.ethz.ch/hdrdatabase/index.php`, 2013, accessed: 2015-11-15. *Cited in Sec.* 4.1, B.5.1

[EMU17] G. Eilertsen, R. K. Mantiuk, and J. Unger, "A comparative review of tone-mapping algorithms for high dynamic range video". *Computer Graphics Forum*, vol. 36 (2), pp. 565–592, 2017. *Cited in Sec.* 1.3.3

[Eng00] P. G. Engeldrum, *Psychometric Scaling: A Toolkit for Imaging Systems Development*, Imcotek Pr, 2000. *Cited in Sec.* 5.1, B.6.1

[Fai07] M. D. Fairchild, "The HDR photographic survey", in *Color and Imaging Conference*, pp. 233–238, Society for Imaging Science and Technology, 2007. *Cited in Sec.* 2.1.2, 2.1.4, 4.1, B.5.1

[Fai13] ———, *Color Appearance Models*, John Wiley & Sons, 2013. *Cited in Sec.* (document), 3, 4, B.1, B.4

[FBC09] B. P. F. Boulos, W. Chen and P. L. Callet, "IRCCyN IVC SD RoI database", Available: `http://ivc.univ-nantes.fr/en/databases/SD_RoI/`, 2009, [Online]]. *Cited in Sec.* 1.1

[FGE+14] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling, and H. Brendel, "Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays", in *IS&T/SPIE Electronic Imaging, Digital Photography X*, 2014. *Cited in Sec.* (document), 1.3.1, 3.1.1, 3.2, 4.1

[FKM00] T. Funamoto, T. Kobayashi, and T. Murao, "High-picture-quality technique for LCD television: LCD-AI", in *Proceeding of International Display Workshop*, 2000. *Cited in Sec.* 1.3.3, 2.1.2

[FKZ+17] K. Feyiz, F. Kamisli, E. Zerman, G. Valenzise, A. Koz, and F. Dufaux, "Statistical analysis and directional coding of layer-based HDR image coding residue", in *IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2017. *Cited in Sec.* 1.3.2

[FPSG96] J. A. Ferwerda, S. N. Pattanaik, P. Shirley, and D. P. Greenberg, "A model of visual adaptation for realistic image synthesis", in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pp. 249–258, ACM, 1996. *Cited in Sec.* 1.3.3

[GDSE10] L. Goldmann, F. De Simone, and T. Ebrahimi, "Impact of acquisition distortion on the quality of stereoscopic images", in *Proceedings of the International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2010. *Cited in Sec.* 5

[GHMN10] J. Gu, Y. Hitomi, T. Mitsunaga, and S. Nayar, "Coded rolling shutter photography: Flexible space-time sampling", in *IEEE International Conference on Image Processing (ICIP)*, IEEE, 2010. *Cited in Sec.* 1.3.1

[GKTT13] M. Granados, K. I. Kim, J. Tompkin, and C. Theobalt, "Automatic noise modeling for ghost-free HDR reconstruction". *ACM Transactions on Graphics*, vol. 32 (6), pp. 201:1–201:10, Nov 2013. *Cited in Sec.* 1.3.1

[GN03] M. Grossberg and S. Nayar, "What is the space of camera response functions?", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2003. *Cited in Sec.* 1.3.1

[GRG+16] D. Gommelet, A. Roumy, C. Guillemot, M. Ropert, and J. LeTanou, "Rate-distortion optimization of a tone mapping with SDR quality constraint for backward-compatible high dynamic range compression", in *IEEE International Conference on Image Processing (ICIP)*, IEEE, 2016. *Cited in Sec.* 1.3.2

[GS16] O. Gallo and P. Sen, "Stack-based algorithms for HDR capture and reconstruction", in *High Dynamic Range Video: From Acquisition, to Display and Applications*, Chap. 3, Academic Press, 2016. *Cited in Sec.* 1.3.1

[GSI10] M. Goudarzi, L. Sun, and E. Ifeachor, "Audiovisual quality estimation for video calls in wireless applications", in *Global Telecommunications Conference (GLOBECOM)*, IEEE, 2010. *Cited in Sec.* 5

[GT11] J.-U. Garbas and H. Thoma, "Temporally coherent luminance-to-luma mapping for high dynamic range video coding with H.264/AVC", in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 829–832, IEEE, 2011. *Cited in Sec.* 1.3.2, 3

[HBK+14] P. Hanhart, M. V. Bernardo, P. Korshunov, M. Pereira, A. M. G. Pinheiro, and T. Ebrahimi, "HDR image compression: A new challenge for objective quality metrics", in *6th International Conference on Quality of Multimedia Experience (QoMEX)*, 2014. *Cited in Sec.* 1.4.1

[HBP+15] P. Hanhart, M. V. Bernardo, M. Pereira, A. M. Pinheiro, and T. Ebrahimi, "Benchmarking of objective quality metrics for HDR image quality assessment". *EURASIP Journal on Image and Video Processing*, vol. 2015 (1), pp. 1–18, 2015. *Cited in Sec.* 1.2.3, 1.4.1, 1.4.2, 4, 4.4, B.5

[HDVD17] V. Hulusic, K. Debattista, G. Valenzise, and F. Dufaux, "A model of perceived dynamic range for HDR images". *Signal Processing: Image Communication*, vol. 51, pp. 26–39, 2017. *Cited in Sec.* 2.3

[HKE14a] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Crowdsourcing evaluation of high dynamic range image compression", in *SPIE Optical Engineering+ Applications, Applications of Digital Image Processing XXXVII*, pp. 92170D–1–92170D–12, International Society for Optics and Photonics, 2014. *Cited in Sec.* 1.4.1

[HKE14b] ———, "Subjective evaluation of higher dynamic range video", in *SPIE Optical Engineering+ Applications, Applications of Digital Image Processing XXXVII*, International Society for Optics and Photonics, 2014. *Cited in Sec.* 1.4.1

[HKE+15] P. Hanhart, P. Korshunov, T. Ebrahimi, Y. Thomas, and H. Hoffmann, "Subjective quality evaluation of high dynamic range video and display for future TV". *SMPTE Motion Imaging Journal*, vol. 124 (4), 2015. *Cited in Sec.* 2

[HL87] R. V. Hogg and J. Ledolter, *Engineering Statistics*, Macmillan Pub Co, 1987. *Cited in Sec.* 4.3.3, B.5.3

[Hoe07] B. Hoefflinger, "The eye and high-dynamic-range vision", in *High-Dynamic-Range (HDR) Vision: Microelectronics, Image Processing, Computer Graphics*, Chap. 1, Springer-Verlag Berlin Heidelberg, 2007. *Cited in Sec.* 1.3

[HRDSE12] P. Hanhart, M. Rerabek, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation of the upcoming HEVC video compression standard", in *SPIE Optical Engineering+ Applications, Applications of Digital Image Processing XXXV*, pp. 84990V–1–84990V–13, International Society for Optics and Photonics, 2012. *Cited in Sec.* 1.1

[HŘE15] P. Hanhart, M. Řeřábek, and T. Ebrahimi, "Towards high dynamic range extensions of HEVC: Subjective evaluation of potential coding technologies", in *SPIE Optical Engineering+ Applications, Applications of Digital Image Processing XXXVIII*, pp. 95990G–1–95990G–12, International Society for Optics and Photonics, 2015. *Cited in Sec.* 1.2.3, 1.4.1, 1.4.2, 4.3.3, B.5.3

[HRE16] P. Hanhart, M. Rerabek, and T. Ebrahimi, "Subjective and objective evaluation of HDR video coding technologies", in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016. *Cited in Sec.* (document), 1.2.3, 1.4.1, 1.4.2, 4, 4.4, 5, 5.2.1, B.1, B.6.2

[HTG08] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment". *Electronics Letters*, vol. 44 (13), pp. 800–801, June 2008. *Cited in Sec.* 2.2.2

[HVP+16] V. Hulusic, G. Valenzise, E. Provenzi, K. Debattista, and F. Dufaux, "Perceived dynamic range of HDR images", in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2016. *Cited in Sec.* 2.3

[ITU98] ITU-R, "Subjective assessment methods for image quality in high-definition television"", ITU-R Recommendation BT.710-4, 1998. *Cited in Sec.* 3.1.1, 5.2.1, B.6.2

[ITU04a] ———, "Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference", ITU-R Recommendation BT.1683, Jun 2004. *Cited in Sec.* 1.2.3

[ITU04b] ITU-T, "Method for specifying accuracy and cross-calibration of video quality metrics (VQM)", ITU-T Recommendation J.149, Mar 2004. *Cited in Sec.* 1.2.3, 4.3.3, B.5.3

[ITU04c] ———, "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference", ITU-T Recommendation J.144, Mar 2004. *Cited in Sec.* 1.2.3

[ITU04d] ———, "Tutorial - objective perceptual assessment of video quality: Full reference television", International Telecommunication Union, Geneva, Switzerland, 2004. *Cited in Sec.* 1.2.3

[ITU08] ———, "Subjective video quality assessment methods for multimedia applications", ITU-T Recommendation P.910, Apr 2008. *Cited in Sec.* 1.1, 2.1.4, 3.1.1, 5, B.3.1

[ITU11] ITU-R, "Reference electro-optical transfer function for flat panel displays used in HDTV studio production", ITU-R Recommendation BT.1886, Aug 2011. *Cited in Sec.* (document), 1.3.2, 2.2.2, 3, B.1

[ITU12a] ———, "General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays", ITU-R Recommendation BT.2022, Aug 2012. *Cited in Sec.* 2.2.1

[ITU12b] ———, "Methodology for the subjective assessment of the quality of television pictures", ITU-R Recommendation BT.500-13, Jan 2012. *Cited in Sec.* 1.1, 1.2.3, 2.2.1, 4.1, 4.3.3, 5, 5.2.1, B.5.3, B.6.2

[ITU12c] ITU-T, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models", ITU-T Recommendation P.1401, Jul 2012. *Cited in Sec.* (document), 1.1, 1.2.3, 2.2.1, 4.3.2, 4.5, 4.6, B.5.3, B.9, B.10

[ITU15a] ITU-R, "Parameter values for the HDTV standards for the studio and for international programme exchange", ITU-R Recommendation BT.709-6, 2015. *Cited in Sec.* 1.3.1, 2.1.2, 3, B.3.1, B.4

[ITU15b] ———, "Parameter values for ultra-high definition television systems for production and international programme exchange", ITU-R Recommendation BT.2020, Oct 2015. *Cited in Sec.* 1.3.1, 1.3.2

[ITU17a] ———, "High dynamic range television for production and international programme exchange", ITU-R Recommendation BT.2930, Oct 2017. *Cited in Sec.* 1.3.2

[ITU17b] ———, "Image parameter values for high dynamic range television for use in production and international programme exchange", ITU-R Recommendation BT.2100, Jun 2017. *Cited in Sec.* (document), 1.3.2, 1.2

[JF03] G. M. Johnson and M. D. Fairchild, "Rendering HDR images", in *Color Imaging Conference*, 2003. *Cited in Sec.* 1.4.1

[JLH+16] L. Jin, J. Y. Lin, S. Hu, H. Wang, P. Wang, I. Katsavounidis, A. Aaron, and C.-C. J. Kuo, "Statistical study on perceived JPEG image quality via MCL-JCI dataset construction and analysis", in *IS&T/SPIE Electronic Imaging, Image Quality and System Performance XIII*, International Society for Optics and Photonics, 2016. *Cited in Sec.* 1.1

[KAR16] E. A. Khan, A. O. Akyuz, and E. Reinhard, "Ghost removal in high dynamic range images", in *IEEE International Conference on Image Processing (ICIP)*, IEEE, 2016. *Cited in Sec.* 1.3.1

[Kay98] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection theory*, Prentice Hall PTR, 1998. *Cited in Sec.* 4.3.3, B.5.3

[KD13] A. Koz and F. Dufaux, "Optimized tone mapping with LDR image quality constraint for backward-compatible high dynamic range image and video coding", in *IEEE International Conference on Image Processing (ICIP)*, IEEE, 2013. *Cited in Sec.* 1.3.2

[KE13] P. Korshunov and T. Ebrahimi, "Context-dependent JPEG backward-compatible high-dynamic range image compression". *Optical Engineering*, vol. 52 (10), pp. 102006–1–102006–15, Oct 2013. *Cited in Sec.* 1.3.2

[KFLCK16] L. Krasula, K. Fliegel, P. Le Callet, and M. Klíma, "On the accuracy of objective image and video quality models: New methodology for performance evaluation", in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2016. *Cited in Sec.* (document), 1.2.3, 4.3.3, 4.9, 4.4, B.5.3, B.12

[KHR+15] 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 *7th International Workshop on Quality of Multimedia Experience (QoMEX)*, 2015. *Cited in Sec.* 1.4.1, 2.1, 4.1, 5, B.1

[KHV+17] D. Kane, V. Hulusic, G. Valenzise, E. Zerman, A. Grimaldi, and M. Bertalmío, "Subjects only prefer to view a linear image when the dynamic range of the displayed image matches that of the original scene", in *European Conference on Visual Perception (ECVP)*, 2017. *Cited in Sec.* 2.3

[KJF07] J. Kuang, G. M. Johnson, and M. D. Fairchild, "iCAM06: A refined image appearance model for HDR image rendering". *Journal of Visual Communication and Image Representation*, vol. 18 (5), pp. 406–414, 2007, special issue on High Dynamic Range Imaging. *Cited in Sec.* 4.1

[KMBF13] J. Korhonen, C. Mantel, N. Burini, and S. Forchhammer, "Modeling the color image and video quality on liquid crystal displays with backlight dimming", in *Visual Communications and Image Processing (VCIP)*, IEEE, 2013. *Cited in Sec.* 2, 2.1.2

[KMFH05] C. C. Koh, S. K. Mitra, J. M. Foley, and I. E. J. Heynderickx, "Annoyance of individual artifacts in MPEG-2 compressed video and their relation to overall annoyance", in *Electronic Imaging, Human Vision and Electronic Imaging X*, International Society for Optics and Photonics, 2005. *Cited in Sec.* 1.1

[KNFLC17] L. Krasula, M. Narwaria, K. Fliegel, and P. Le Callet, "Preference of experience in image tone-mapping: Dataset and framework for objective measures comparison". *IEEE Journal of Selected Topics in Signal Processing*, vol. 11 (1), pp. 64–74, Feb 2017. *Cited in Sec.* 1.4.1

[KO05] P. Korshunov and W. T. Ooi, "Critical video quality for distributed automated video surveillance", in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, 2005. *Cited in Sec.* 1.1

[KP86] M. K. Kundu and S. K. Pal, "Thresholding for edge detection using human psychovisual phenomena". *Pattern Recognition Letters*, vol. 4 (6), pp. 433–441, 1986. *Cited in Sec.* 1.4.1, 4

[KRTT12] C. Kiser, E. Reinhard, M. Tocci, and N. Tocci, "Real time automated tone mapping system for HDR video", in *IEEE International Conference on Image Processing (ICIP)*, IEEE, 2012. *Cited in Sec.* 1.3.2

[Lar98a] G. W. Larson, "LogLuv encoding for full-gamut, high-dynamic range images". *Journal of Graphics Tools*, vol. 3 (1), pp. 15–31, 1998. *Cited in Sec.* 1.3.1, 1.3.2, 3, B.4

[Lar98b] ———, "Overcoming gamu and dynamic range limitations in digital images", in *Color Imaging Conference*, 1998. *Cited in Sec.* 1.3.1

[LC10] 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 (1), pp. 011006–1–011006–21, 2010. *Cited in Sec.* 1.1

[LCA05] P. Le Callet and F. Autrusseau, "Subjective quality assessment IRCCyN/IVC database", 2005, http://www.irccyn.ec-nantes.fr/ivcdb/. *Cited in Sec.* 5

[LCMP13] P. Le Callet, S. Möller, and A. Perkis, "Qualinet white paper on definitions of quality of experience", European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Mar 2013, Lausanne, Switzerland, Version 1.2. *Cited in Sec.* 1.1

[LCR01] M. R. Luo, G. Cui, and B. Rigg, "The development of the CIE 2000 colour-difference formula: CIEDE2000". *Color Research & Application*, vol. 26 (5), pp. 340–350, 2001. *Cited in Sec.* 1.2.1, 3.2.2, 4.3.1, B.4.2

[LCTS05] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, "Evaluation of tone mapping operators using a high dynamic range display". *ACM Transactions on Graphics*, vol. 24 (3), pp. 640–648, Jul 2005. *Cited in Sec.* 1.3.2, 1.4.1

[LDSE11] J.-S. Lee, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation via paired comparison: Application to scalable video coding". *IEEE Transactions on Multimedia*, vol. 13 (5), pp. 882–893, 2011. *Cited in Sec.* 1.1, 3.2.2, 5.1, B.4.2, B.6.1

[LFH15] A. Luthra, E. Francois, and W. Husak, "Call for Evidence (CfE) for HDR and WCG Video Coding", ISO/IEC JTC1/SC29/WG11 MPEG2014/N15083, 2015. *Cited in Sec.* (document), 1.3.2, 1.2, 3, 4, 4.1, B.1

[LGYS04] S. Lin, J. Gu, S. Yamazaki, and H.-Y. Shum, "Radiometric calibration from a single image", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2004. *Cited in Sec.* 1.3.1

[LHL+08] F. C. Lin, Y. P. Huang, L. Y. Liao, C. Y. Liao, H. P. D. Shieh, T. M. Wang, and S. C. Yeh, "Dynamic backlight gamma on high dynamic range LCD TVs". *Journal of Display Technology*, vol. 4 (2), pp. 139–146, Jun 2008. *Cited in Sec.* 1.3.3

[LJH+15] J. Y. Lin, L. Jin, S. Hu, I. Katsavounidis, Z. Li, A. Aaron, and C.-C. J. Kuo, "Experimental design and analysis of JND test on coded image/video", in *SPIE Optical Engineering+ Applications, Applications of Digital Image Processing XXXVIII*, pp. 95990Z–1–95990Z–11, International Society for Optics and Photonics, 2015. *Cited in Sec.* 1.1

[LJK11] W. Lin and C. C. Jay Kuo, "Perceptual visual quality metrics: A survey". *Journal of Visual Communication and Image Representation*, vol. 22 (4), pp. 297–312, May 2011. *Cited in Sec.* 1.2

[LK05] C. Lee and C.-S. Kim, "Rate-distortion optimized compression of high dynamic range videos", in *16th European Signal Processing Conference (EUSIPCO)*, IEEE, 2005. *Cited in Sec.* 1.3.2

[LK07] ———, "Gradient domain tone mapping of high dynamic range videos", in *IEEE International Conference on Image Processing (ICIP)*, IEEE, 2007. *Cited in Sec.* 1.3.2

[LK08] ———, "Rate-distortion optimized compression of high dynamic range videos", in *16th European Signal Processing Conference (EUSIPCO)*, IEEE, 2008. *Cited in Sec.* 1.3.2

[LLF13] S. Lasserre, F. LeLéannec, and E. Francois, "Description of HDR sequences proposed by technicolor". *ISO/IEC JTC1/SC29/WG11 JCTVC-P0228, IEEE, San Jose, USA*, 2013. *Cited in Sec.* 2.1.4, 4.1

[LPY+16] T. Lu, F. Pu, P. Yin, T. Chen, W. Husak, J. Pytlarz, R. Atkins, J. Fröhlich, and G. Su, "ITP colour space and its compression performance for high dynamic range and wide colour gamut video distribution". *ZTE Communications*, Feb 2016. *Cited in Sec.* 1.3.2, 3, B.4

[LRP97] G. W. Larson, H. Rushmeier, and C. Piatko, "A visibility matching tone reproduction operator for high dynamic range scenes". *IEEE Transactions on Visualization and Computer Graphics*, vol. 3 (4), pp. 291–306, Oct 1997. *Cited in Sec.* 1.3.3

[LSH+17] Y. Liu, N. Sidaty, W. Hamidouche, O. Déforges, G. Valenzise, and E. Zerman, "An adaptive perceptual quantization method for HDR video coding", in *IEEE International Conference on Image Processing (ICIP)*, IEEE, 2017. *Cited in Sec.* 1.3.2, 2.1

[LT08] C.-C. Lai and C.-C. Tsai, "Backlight power reduction and image contrast enhancement using adaptive dimming for global backlight applications". *IEEE Transactions on Consumer Electronics*, vol. 54 (2), pp. 669–674, 2008. *Cited in Sec.* 1.3.3, 2.1.2

[MB11] A. K. Moorthy and A. C. Bovik, "Visual quality assessment algorithms: What does the future hold?" *Multimedia Tools and Applications*, vol. 51 (2), pp. 675–696, 2011. *Cited in Sec.* 1.1, 1.2.2

[MBK+13] C. Mantel, N. Burini, J. Korhonen, E. Nadernejad, and S. Forchhammer, "Quality assessment of images displayed on LCD screen with local backlight dimming", in *5th International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 48–49, 2013. *Cited in Sec.* 2

[MBK+15] C. Mantel, S. Bech, J. Korhonen, S. Forchhammer, and J. M. Pedersen, "Modeling the subjective quality of highly contrasted videos displayed on LCD with local backlight dimming". *IEEE Transactions on Image Processing*, vol. 24 (2), pp. 573–582, 2015. *Cited in Sec.* 1.3.3

[MDBR+16] R. Mukherjee, K. Debattista, T. Bashford-Rogers, P. Vangorp, R. Mantiuk, M. Bessa, B. Waterfield, and A. Chalmers, "Objective and subjective evaluation of high dynamic range video compression". *Signal Processing: Image Communication*, vol. 47, pp. 426–437, 2016. *Cited in Sec.* 1.3.2, 1.4.1

[MDK08] R. Mantiuk, S. Daly, and L. Kerofsky, "Display adaptive tone mapping". *ACM Transactions on Graphics*, vol. 27 (3), Aug 2008. *Cited in Sec.* (document), 1.3.3, 2.1.2, 2.2(a), 3.2, 4.1, B.3.1, B.1(a)

[MDMS05] R. Mantiuk, S. J. Daly, K. Myszkowski, and H.-P. Seidel, "Predicting visible differences in high dynamic range images: Model and its calibration", in *Electronic Imaging, Human Vision and Electronic Imaging X*, International Society for Optics and Photonics, 2005. *Cited in Sec.* 1.4.2, 2, 2.2, 2.2.2, B.3.2

[MDWE02] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric", in *IEEE International Conference on Image Processing (ICIP)*, IEEE, 2002. *Cited in Sec.* 1.1

[MEMS06] R. Mantiuk, A. Efremov, K. Myszkowski, and H.-P. Seidel, "Backward compatible high dynamic range MPEG video compression". *ACM Transactions on Graphics (TOG)*, vol. 25 (3), pp. 713–723, 2006. *Cited in Sec.* 1.3.2

[MK05] R. Muijs and I. Kirenko, "A no-reference blocking artifact measure for adaptive video processing", in *13th European Signal Processing Conference (EUSIPCO)*, IEEE, 2005. *Cited in Sec.* 1.1

[MKF+15] C. Mantel, J. Korhonen, S. Forchhammer, J. Pedersen, and S. Bech, "Subjective quality of videos displayed with local backlight dimming at different peak white and ambient light levels", in *7th International Workshop on Quality of Multimedia Experience (QoMEX)*, IEEE, 2015. *Cited in Sec.* 2

[MKMS04] R. Mantiuk, G. Krawczyk, K. Myszkowski, and H.-P. Seidel, "Perception-motivated high dynamic range video encoding". *ACM Transactions on Graphics (TOG)*, vol. 23 (3), pp. 733–741, 2004. *Cited in Sec.* 1.3.2, 3

[MKMS07] R. Mantiuk, G. Krawczyk, R. Mantiuk, and H.-P. Seidel, "High dynamic range imaging pipeline: Perception-motivated representation of visual content", in *Electronic Imaging, Human Vision and Electronic Imaging XII*, pp. 649212–1–649212–12, International Society for Optics and Photonics, 2007. *Cited in Sec.* 1.3.1

[MKRH11] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions". *ACM Transactions on Graphics*, vol. 30 (4), 2011. *Cited in Sec.* 1.4.2, 4, B.5

[MMM+11] Z. Mai, H. Mansour, R. Mantiuk, P. Nasiopoulos, R. Ward, and W. Heidrich, "Optimizing a tone curve for backward-compatible high dynamic range image and video compression". *IEEE Transactions on Image Processing*, vol. 20 (6), pp. 1558–1571, June 2011. *Cited in Sec.* 1.3.2, 1.3.3, 4.1, B.5.1

[MMNW13] Z. Mai, H. Mansour, P. Nasiopoulos, and R. K. Ward, "Visually favorable tone-mapping with high compression performance in bit-depth scalable video coding". *IEEE Transactions on Multimedia*, vol. 15 (7), pp. 1503–1518, 2013. *Cited in Sec.* 1.3.2

[MMS06] R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "A perceptual framework for contrast processing of high dynamic range images". *ACM Transactions on Applied Perception*, vol. 3 (3), pp. 286–308, 2006. *Cited in Sec.* 1.4.1, 4.1

[MN99] T. Mitsunaga and S. Nayar, "Radiometric self calibration", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 1999. *Cited in Sec.* 1.3.1

[MND12] S. Miller, M. Nezamabadi, and S. Daly, "Perceptual signal coding for more efficient usage of bit codes", in *SMPTE 2012 Annual Technical Conference & Exhibition*, Society of Motion Picture and Television Engineers, 2012. *Cited in Sec.* 1.3.2, 1.4.2, 3, 4.1, 4.3.1, B.1, B.5.1

[MP95] S. Mann and R. Picard, "Being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures", in *Proceedings of the Society for Imaging Science and Technology (IS&T)*, The Society for Imaging Science and Technology, 1995. *Cited in Sec.* 1.3.1

[MSL+16] S. Mahmalat, N. Stefanoski, D. Luginbühl, T. O. Aydin, and A. Smolic, "Luminance independent chromaticity preprocessing for HDR video coding", in *IEEE International Conference on Image Processing (ICIP)*, pp. 1389–1393, IEEE, 2016. *Cited in Sec.* 2.1

[MT10] A. Motra and H. Thoma, "An adaptive logluv transform for high dynamic range video compression", in *IEEE International Conference on Image Processing (ICIP)*, IEEE, 2010. *Cited in Sec.* 1.3.2

[MTM12] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment". *Computer Graphics Forum*, vol. 31 (8), pp. 2478–2491, 2012. *Cited in Sec.* 1.1, 3.2.1, 5, B.4.2, B.6

[Nam10] H. Nam, "A color compensation algorithm to avoid color distortion in active dimming liquid crystal displays". *IEEE Transactions on Consumer Electronics*, vol. 56 (4), pp. 2569–2576, 2010. *Cited in Sec.* 2.1.2, A.2

[Nam11] ———, "Low power active dimming liquid crystal display with high resolution backlight". *Electronics Letters*, vol. 47 (9), pp. 538–540, Apr 2011. *Cited in Sec.* 1.3.3

[NDSLC15] M. Narwaria, M. P. Da Silva, and P. Le Callet, "HDR-VQM: An objective quality measure for high dynamic range video". *Signal Processing: Image Communication*, vol. 35, pp. 46–60, 2015. *Cited in Sec.* (document), 1.4.2, 2, 2.2, 2.2.2, 4, 4.3.1, B.3.2, B.5, B.5.3

[NDSLC16a] ———, "Dual modulation for LED-backlit HDR displays", in *High Dynamic Range Video: From Acquisition, to Display and Applications*, Chap. 14, Academic Press, 2016. *Cited in Sec.* 1.3.3, 2.1, B.3.1

[NdSLC+16b] M. Narwaria, M. P. da Silva, P. Le Callet, G. Valenzise, F. De Simone, and F. Dufaux, "Quality of experience and HDR: Concepts and how to measure it", in *High Dynamic Range Video: From Acquisition, to Display and Applications*, Chap. 16, Academic Press, 2016. *Cited in Sec.* 4, B.5

[NDSLCP12] M. Narwaria, M. P. Da Silva, P. Le Callet, and R. Pépion, "Effect of tone mapping operators on visual attention deployment", in *SPIE Optical Engineering+ Applications, Applications of Digital Image Processing XXXV*, pp. 84990G–1–84990G–15, International Society for Optics and Photonics, 2012. *Cited in Sec.* 4, B.5

[NDSLCP13] M. Narwaria, M. P. Da Silva, P. Le Callet, and R. Pepion, "Tone mapping-based high-dynamic-range image compression: Study of optimization criterion and perceptual quality". *Optical Engineering*, vol. 52 (10), pp. 102008–1–102008–15, 2013. *Cited in Sec.* 1.4.1, 2.1, 4, 4.1, B.1, B.5

[NDSLCP14a] M. Narwaria, M. P. Da Silva, P. Le Callet, and R. Pépion, "Impact of tone mapping in high dynamic range image compression", in *8th International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2014. *Cited in Sec.* 1.4.1, 2.1, 4.1, 4.2, B.1, B.5.2

[NDSLCP14b] M. Narwaria, M. P. Da Silva, P. Le Callet, and R. Pepion, "Tone mapping based HDR compression: Does it affect visual experience?" *Signal Processing: Image Communication*, vol. 29 (2), pp. 257–273, 2014. *Cited in Sec.* 1.4.1, 4

[NLCV+16] M. Narwaria, P. Le Callet, G. Valenzise, F. De Simone, F. Dufaux, and R. Mantiuk, "HDR image and video quality prediction", in *High Dynamic Range Video: From Acquisition, to Display and Applications*, Chap. 17, Academic Press, 2016. *Cited in Sec.* 4

[NLM+12] M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel, and L.-T. Chia, "Fourier transform-based scalable image quality measure". *IEEE Transactions on Image Processing*, vol. 21 (8), pp. 3364–3377, 2012. *Cited in Sec.* 4.1

[NM00] S. Nayar and T. Mitsunaga, "High dynamic range imaging: Spatially varying pixel exposures", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2000. *Cited in Sec.* 1.3.1

[NMBF13] E. Nadernejad, C. Mantel, N. Burini, and S. Forchhammer, "Flicker reduction in LED-LCDs with local backlight", in *IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 312–316, IEEE, 2013. *Cited in Sec.* 2.1.3

[NMDSLC15] M. Narwaria, R. K. Mantiuk, M. P. Da Silva, and P. Le Callet, "HDR-VDP-2.2: A calibrated method for objective quality prediction of high-dynamic range and standard images". *Journal of Electronic Imaging*, vol. 24 (1), pp. 010501–1–010501–3, 2015. *Cited in Sec.* (document), 1.4.2, 2, 2.2, 2.2.2, 3.2.2, 4, 4.3.1, B.3.2, B.4.2, B.5.3

[NPDSLCP14] M. Narwaria, M. Perreira Da Silva, P. Le Callet, and R. Pepion, "Single exposure vs tone mapped high dynamic range images: A study based on quality of experience", in *22nd European Signal Processing Conference (EUSIPCO)*, IEEE, 2014. *Cited in Sec.* 1.4.1

[NVH16] M. Nuutinen, T. Virtanen, and J. Häkkinen, "Performance measure of image and video quality assessment algorithms: Subjective root-mean-square error". *Journal of Electronic Imaging*, vol. 25 (2), pp. 023012–1–023012–13, 2016. *Cited in Sec.* 1.2.3, 4.3.3, B.5.3

[OCHZ09] S. Ouni, M. Chambah, M. Herbin, and E. Zagrouba, "SCID: full reference spatial color image quality metric", in *IS&T/SPIE Electronic Imaging, Image Quality and System Performance VI*, SPIE, 2009. *Cited in Sec.* 3.2.2, 3.3

[OJKP16] B. Ortiz-Jaramillo, A. Kumcu, and W. Philips, "Evaluating color difference measures in images", in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2016. *Cited in Sec.* 1.2.1, 3.2.2, 3.3, 4.3.2, B.5.3

[OLVD16] C. Ozcinar, P. Lauga, G. Valenzise, and F. Dufaux, "HDR video coding based on a temporally constrained tone mapping operator", in *Digital Media Industry Academic Forum (DMIAF)*, 2016. *Cited in Sec.* 1.3.2

[OZW10] Y.-F. Ou, Y. Zhou, and Y. Wang, "Perceptual quality of video with frame rate variation: A subjective study", in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 2446–2449, IEEE, 2010. *Cited in Sec.* 1.1, 5

[Pau02] D. Paul, "Image-based lighting". *IEEE Computer Graphics and Applications*, vol. 22 (2), pp. 26–34, 2002. *Cited in Sec.* 1.3.1

[PEB+11] Y. Pitrey, U. Engelke, M. Barkowsky, R. Pépion, and P. Le Callet, "Aligning subjective tests using a low cost common set", in *Euro ITV*, 2011. *Cited in Sec.* 4.1

[PFFG98] S. N. Pattanaik, J. A. Ferwerda, M. D. Fairchild, and D. P. Greenberg, "A multiscale model of adaptation and spatial vision for realistic image display", in *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pp. 287–298, ACM, 1998. *Cited in Sec.* 1.3.3

[pfs15] pfstools, "pfstools HDR image gallery", `http://pfstools.sourceforge.net/hdr_gallery.html`, 2015, accessed: 2015-11-15. *Cited in Sec.* 4.1, B.5.1

[PIL+13] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, *et al.*, "Color image database TID2013: Peculiarities and preliminary results", in *4th European Workshop on Visual Information Processing (EUVIP)*, pp. 106–111, IEEE, 2013. *Cited in Sec.* 1.1

[PJI+15] 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. *Cited in Sec.* (document), 1.1, 1.2.3, 5, 5.1, B.1

[PLZ+09] 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 (4), pp. 30–45, 2009. *Cited in Sec.* 1.1, 5, 5.1

[POM17] M. Perez-Ortiz and R. K. Mantiuk, "A practical guide and software for analysing pairwise comparison experiments", 2017. *Cited in Sec.* 5.1, B.6.1

[PPLC08a] S. Péchard, R. Pépion, and P. Le Callet, "IRCCyN IVC 1080i database", Available: `http://ivc.univ-nantes.fr/en/databases/1080i_Videos/`, 2008, [Online]]. *Cited in Sec.* 1.1

[PPLC08b] ———, "Suitable methodology in subjective video quality assessment: A resolution dependent paradigm", in *International Workshop on Image Media Quality and its Applications (IMQA)*, 2008. *Cited in Sec.* 1.1, 5, 5.2.1

[PS10] M. Pinson and F. Speranza, "Report on the validation of video quality models for high definition video content", Tech. Rep., Video Quality Experts Group (VQEG), Jun 2010. *Cited in Sec.* 1.1

[PW03a] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies", in *Visual Communications and Image Processing (VCIP)*, pp. 573–582, International Society for Optics and Photonics, 2003. *Cited in Sec.* 1.1

[PW03b] ———, "An objective method for combining multiple subjective data sets", in *Visual Communications and Image Processing (VCIP)*, pp. 583–592, International Society for Optics and Photonics, 2003. *Cited in Sec.* 4, 4.2, 5, B.5.2

[PW04] ———, "A new standardized method for objectively measuring video quality". *IEEE Transactions on Broadcasting*, vol. 50 (3), pp. 312–322, Sep 2004. *Cited in Sec.* 1.2.2

[PW08] M. Pinson and S. Wolf, "Techniques for evaluating objective video quality models using overlapping subjective data sets", Tech. Rep., US Department of Commerce, National Telecommunications and Information Administration, 2008, nTIA Technical Report TR-09-457. *Cited in Sec.* 1.2.3, 4.3.3, B.5.3

[qua17a] "'quality', Cambridge Dictionary Online", `https://dictionary.cambridge.org/dictionary/english/quality`, 2017, accessed: 2017-11-27. *Cited in Sec.* 1.1

[qua17b] "'quality', Merriam-Webster Online", `https://www.merriam-webster.com/dictionary/quality`, 2017, accessed: 2017-11-27. *Cited in Sec.* 1.1

[RBS03] M. A. Robertson, S. Borman, and R. L. Stevenson, "Estimation-theoretic approach to dynamic range enhancement using multiple exposures". *Journal of Electronic Imaging*, vol. 12 (2), pp. 219–228, 2003. *Cited in Sec.* 1.3.1

[ŘHKE15] M. Řeřábek, P. Hanhart, P. Korshunov, and T. Ebrahimi, "Subjective and objective evaluation of HDR video compression", in *9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2015. *Cited in Sec.* 1.4.1, 1.4.2

[RHLM09] A. G. Rempel, W. Heidrich, H. Li, and R. Mantiuk, "Video viewing preferences for HDR displays under varying ambient illumination", in *Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization (APGV)*, pp. 45–52, ACM, 2009. *Cited in Sec.* 1.4.1, 2

[Ric13] T. Richter, "On the standardization of the JPEG XT image compression", in *Picture Coding Symposium (PCS)*, pp. 37–40, IEEE, Dec 2013. *Cited in Sec.* (document), 1.3.2, 4, 4.1, B.1

[Ric16] ———, "High dynamic range imaging with JPEG XT", in *High Dynamic Range Video: From Acquisition, to Display and Applications*, Chap. 12, Academic Press, 2016. *Cited in Sec.* 1.3.2

[RLC+00] A. M. Rohaly, J. Libert, P. Corriveau, A. Webster, *et al.*, "Final report from the video quality experts group on the validation of objective models of video quality assessment", http://www.vqeg.org/, Mar 2000. *Cited in Sec.* 4.3.2, B.5.3

[RLCW00] A. M. Rohaly, J. Libert, P. Corriveau, and A. Webster, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment", Tech. Rep., Video Quality Experts Group (VQEG), Apr 2000. *Cited in Sec.* 1.1

[RPLCH10] D. M. Rouse, R. Pépion, P. Le Callet, and S. S. Hemami, "Tradeoffs in subjective testing methods for image and video quality assessment", in *IS&T/SPIE Electronic Imaging, Human Vision and Electronic Imaging XV*, pp. 75270F–1–75270F–11, International Society for Optics and Photonics, 2010. *Cited in Sec.* 1.1

[RSSF02] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images". *ACM Transactions on Graphics*, vol. 21 (3), pp. 267–276, Jul 2002. *Cited in Sec.* 1.3.2, 1.3.3, 1.4.1, 4.1

[RSYD05] R. Ramanath, W. E. Snyder, Y. Yoo, and M. S. Drew, "Color image processing pipeline". *IEEE Signal Processing Magazine*, vol. 22 (1), pp. 34–43, Mar 2005. *Cited in Sec.* 1.3.1

[SB06] H. R. Sheikh and A. C. Bovik, "Image information and visual quality". *IEEE Transactions on Image Processing*, vol. 15 (2), pp. 430–444, 2006. *Cited in Sec.* (document), 1.2.1, 2.2.2, 4, 4.3.1, B.5, B.5.3

[SB10] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos". *IEEE Transactions on Image Processing*, vol. 19 (2), pp. 335–350, 2010. *Cited in Sec.* 1.2.2

[SBDV05] H. R. Sheikh, A. C. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics". *IEEE Transactions on Image Processing*, vol. 14 (12), pp. 2117–2128, 2005. *Cited in Sec.* (document), 1.2.1, 2.2.2, 4.3.1, B.5.3

[SF01] D. A. Silverstein and J. E. Farrell, "Efficient method for paried comparison". *Journal of Electronic Imaging*, vol. 10 (2), 2001. *Cited in Sec.* 5.1, B.6.1

[SHS+04] H. Seetzen, W. Heidrich, W. Stuerzlinger, G. Ward, L. Whitehead, M. Trentacoste, A. Ghosh, and A. Vorozcovs, "High dynamic range display systems". *ACM Transactions on Graphics*, vol. 23 (3), pp. 760–768, Aug 2004. *Cited in Sec.* 1.3.3, 2.1, 2.1.2, B.3.1

[SIM14] SIM2, "http://www.sim2.com/HDR/", `http://www.sim2.com/HDR/`, Jun 2014. *Cited in Sec.* 2.1, 2.1.1, 2.2, B.3.1

[SJPK+10] D. Strohmeier, S. Jumisko-Pyykkö, K. Kunze, G. Tech, D. Buğdayci, and M. O. Bici, "Results of quality attributes of coding, transmission and their combinations", Tech. Rep., Mobile 3DTV, Jan 2010. *Cited in Sec.* 1.1

[SLY+06] H. Seetzen, H. Li, L. Ye, W. Heidrich, L. Whitehead, and G. Ward, "25.3: Observations of luminance, contrast and amplitude resolution of displays". *SID Symposium Digest of Technical Papers*, vol. 37 (1), pp. 1229–1233, 2006. *Cited in Sec.* 1.4.1

[SMP14] SMPTE, "High dynamic range electro-optical transfer function of mastering reference displays", SMPTE ST 2084, 2014. *Cited in Sec.* 1.3.2, 1.4.2, 3, 3.1.1, 4, 4.1, 4.3.1, B.1, B.4.1, B.5.1

[SOHW12] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard". *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22 (12), pp. 1649–1668, 2012. *Cited in Sec.* 1.3.2, 3, 3.1.1, 3.2.1, B.4.1

[SSB06] 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 (11), pp. 3440–3451, Nov 2006. *Cited in Sec.* (document), 1.1, 5, B.1

[SSBC10a] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video". *IEEE Transactions on Image Processing*, vol. 19 (6), pp. 1427–1441, 2010. *Cited in Sec.* (document), 1.1, 1.2.3, 5, B.1

[SSBC10b] ———, "A subjective study to evaluate video quality assessment algorithms", in *IS&T/SPIE Electronic Imaging, Human Vision and Electronic Imaging XV*, pp. 75270H–1–75270H–10, International Society for Optics and Photonics, 2010. *Cited in Sec.* 1.1, 5

[STZ+07] S. Srinivasan, C. Tu, Z. Zhou, D. Ray, S. Regunathan, and G. Sullivan, "An introduction to the HDPhoto technical design", In: JPEG Document WG1N4183, 2007. *Cited in Sec.* 1.3.2

[SW03] S. E. Susstrunk and S. Winkler, "Color image quality on the internet", in *Electronic Imaging, Internet Imaging V*, SPIE, 2003. *Cited in Sec.* 3.2.2, 3.3

[TF15] D. Touze and E. Francois, "Description of new version of HDR class A and A' sequences", ISO/IEC JTC1/SC29/WG11 MPEG2014/M35477, 2015. *Cited in Sec.* (document), 3.1.1, 3.2

[TG11] K. Tsukida and M. R. Gupta, "How to analyze paired comparison data", Tech. Rep., Department of Electrical Engineering, University of Washington, 2011, uWEE Technical Report Number UWEETR-2011-0004. *Cited in Sec.* 3.2.2, 5.1, B.4.2, B.6.1

[THOT10] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video", in *2nd International Workshop on Quality of Multimedia Experience (QoMEX)*, IEEE, Jun 2010. *Cited in Sec.* 1.1

[Thu27] L. L. Thurstone, "A law of comparative judgement". *Psychological Review*, vol. 34 (4), pp. 273–286, 1927. *Cited in Sec.* 1.1, 1.4.1, 3.2.2, 5.1, B.4.2, B.6.1

[THW+07] M. Trentacoste, W. Heidrich, L. Whitehead, H. Seetzen, and G. Ward, "Photometric image processing for high dynamic range displays". *Journal of Visual Communication and Image Representation*, vol. 18 (5), pp. 439–451, 2007, special issue on High Dynamic Range Imaging. *Cited in Sec.* 1.3.3

[TKTS11] M. D. Tocci, C. Kiser, N. Tocci, and P. Sen, "A versatile HDR video production system". *ACM Transactions on Graphics*, vol. 30 (4), pp. 41:1–41:10, Jul 2011. *Cited in Sec.* 1.3.1

[TLSS09] A. M. Tourapis, A. Leontaris, K. Suhring, and G. Sullivan, "H.264/14496-10 AVC reference software manual", Tech. Rep. Doc. JVT-AE010, Joint Video Team (JVT), 2009. *Cited in Sec.* 1.3.2, 3

[TM07] A. Tomaszewska and R. Mantiuk, "Image registration for multi-exposure high dynamic range image acquisition", in *WSCG*, 2007. *Cited in Sec.* 1.3.1

[TR93] J. Tumblin and H. Rushmeier, "Tone reproduction for realistic images". *IEEE Computer Graphics and Applications*, vol. 13 (6), pp. 42–48, 1993. *Cited in Sec.* 1.3.3

[TS15] A. M. Tourapis and D. Singer, "HDRTools: Software updates", ISO/IEC JTC1/SC29/WG11 MPEG2015/M35471, IEEE, Ed., Geneva, Switzerland, 2015. *Cited in Sec.* 4.3.1, B.5.3

[UHK16] J. Unger, S. Hajisharif, and J. Kronander, "Unified reconstruction of raw HDR video data", in *High Dynamic Range Video: From Acquisition, to Display and Applications*, Chap. 2, Academic Press, 2016. *Cited in Sec.* 1.3.1

[VCL+11] P. Vangorp, G. Chaurasia, P.-Y. Laffont, R. W. Fleming, and G. Drettakis, "Perception of visual artifacts in image-based rendering of façades", in *Proceedings of the 22nd Eurographics workshop on Rendering*, pp. 1241–1250, Eurographics Association, 2011. *Cited in Sec.* 1.1

[VDSLD14] G. Valenzise, F. De Simone, P. Lauga, and F. Dufaux, "Performance evaluation of objective quality metrics for HDR image compression", in *SPIE Optical Engineering+ Applications, Applications of Digital Image Processing XXXVII*, pp. 92170C–1–92170C–10, International Society for Optics and Photonics, 2014. *Cited in Sec.* (document), 1.3.2, 1.4.1, 1.4.2, 2, 2.1.4, 2.2, 2.2.1, 2.11, 2.2.2, 2.12, 2.3, 4, 4.1, 4.3.1, 4.4, B.3.2, B.1, B.5, B.5.1

[VLD07] P. Vangorp, J. Laurijssen, and P. Dutré, "The influence of shape on the perception of material reflectance". *ACM Transactions on Graphics*, vol. 26 (3), Jul 2007. *Cited in Sec.* 1.1

[VQE00] VQEG, "VQEG FR-TV Phase I database", `https://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-i/frtv-phase-i.aspx`, 2000, [Online]]. *Cited in Sec.* 1.1

[War91] G. Ward, "Real pixels", in J. Arvo, ed., *Graphic Gems II*, Academic Press, 1991. *Cited in Sec.* 1.3.1

[War94] G. J. Ward, "The RADIANCE lighting simulation and rendering system", in *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pp. 459–472, ACM, 1994. *Cited in Sec.* 1.3.1

[War03] G. Ward, "Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures". *Journal of Graphics Tools*, vol. 8 (2), pp. 17–30, 2003. *Cited in Sec.* 1.3.1

[WB02] Z. Wang and A. C. Bovik, "A universal image quality index". *IEEE Signal Processing Letters*, vol. 9 (3), pp. 81–84, 2002. *Cited in Sec.* (document), 1.2.1, 4.3.1, B.5.3

[WB06] ———, *Modern Image Quality Assessment*, Morgan & Claypool, 2006. *Cited in Sec.* 1.2

[WB09] ———, "Mean squared error: Love it or leave it? a new look at signal fidelity measures". *IEEE Signal Processing Magazine*, vol. 26 (1), pp. 98–117, Feb 2009. *Cited in Sec.* 1.2.1

[WBSS04] 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 (4), pp. 600–612, 2004. *Cited in Sec.* (document), 1.1, 1.2.1, 2.2.2, 4, 4.3.1, B.5, B.5.3

[Win05] S. Winkler, *Digital Video Quality: Vision models and metrics*, John Wiley & Sons, Chichester, West Sussex, England Hoboken, NJ, 2005. *Cited in Sec.* 1.2

[Win12] ———, "Analysis of public image and video databases for quality assessment". *IEEE Journal of Selected Topics in Signal Processing*, vol. 6 (6), pp. 616–625, Oct 2012. *Cited in Sec.* 1.1

[WM08] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics". *IEEE Transactions on Broadcasting*, vol. 54 (3), pp. 660–668, Jun 2008. *Cited in Sec.* 1.2.2

[WS04] G. Ward and M. Simmons, "Subband encoding of high dynamic range imagery", in *Proceedings of the 1st Symposium on Applied Perception in Graphics and Visualization (APGV)*, pp. 83–90, ACM, 2004. *Cited in Sec.* 1.3.2

[WS06] ———, "JPEG-HDR: A backwards-compatible, high dynamic range extension to JPEG", in *ACM SIGGRAPH 2006 Courses*, ACM, New York, NY, USA, 2006. *Cited in Sec.* 1.3.2, 1.4.1, 4.1, B.1, B.5.1

[WSB03] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment", in *37th Asilomar Conference on Signals, Systems Computers*, vol. 2, pp. 1398–1402, IEEE, 2003. *Cited in Sec.* (document), 1.2.1, 2.2.2, 4, 4.3.1, B.5, B.5.3

[XPCH17] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017. *Cited in Sec.* 1.1

[YJK16] S. Yu, C. Jung, and P. Ke, "Adaptive PQ: Adaptive perceptual quantizer for HEVC main 10 profile-based HDR video coding", in *Visual Communications and Image Processing (VCIP)*, IEEE, 2016. *Cited in Sec.* 1.3.2

[YMMS06] A. Yoshida, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Analysis of reproducing real-world appearance on displays of varying dynamic range". *Computer Graphics Forum*, vol. 25 (3), pp. 415–426, Sep 2006. *Cited in Sec.* 1.4.1

[ZBW11] H. Zimmer, A. Bruhn, and J. Weickert, "Freehand HDR imaging of moving scenes with simultaneous resolution enhancement". *Computer Graphics Forum*, vol. 30 (2), pp. 405–414, 2011. *Cited in Sec.* 1.3.1

[ZHV+17] E. Zerman, V. Hulusic, G. Valenzise, R. Mantiuk, and F. Dufaux, "Effect of color space on high dynamic range video compression performance", in *9th International Workshop on Quality of Multimedia Experience (QoMEX)*, IEEE, Jun 2017. *Cited in Sec.* 2.1, 3.3

[ZHV+18] ———, "The relation between mos and pairwise comparisons and the importance of cross-content comparisons", in *IS&T/SPIE Electronic Imaging, Human Vision and Electronic Imaging XXII*, International Society for Optics and Photonics, 2018. *Cited in Sec.* 5.4

[Zou07] G. Y. Zou, "Toward using confidence intervals to compare correlations". *Psychological Methods*, vol. 12 (4), pp. 399–413, Dec 2007. *Cited in Sec.* 2.2.2

[ZVD16] E. Zerman, G. Valenzise, and F. Dufaux, "A dual modulation algorithm for accurate reproduction of high dynamic range video", in *IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, IEEE, Jul 2016. *Cited in Sec.* 2.1, 2.1.2, 2.3, 4.3.1, B.5.3

[ZVD17] ———, "An extensive performance evaluation of full-reference HDR image quality metrics". *Quality and User Experience*, 2017. *Cited in Sec.* 3.2.2, 4.4

[ZVDS+15] E. Zerman, G. Valenzise, F. De Simone, F. Banterle, and F. Dufaux, "Effects of display rendering on HDR image quality assessment", in *SPIE Optical Engineering+ Applications, Applications of Digital Image Processing XXXVIII*, International Society for Optics and Photonics, 2015. *Cited in Sec.* 2.1.2, 2.3

[ZW97] X. Zhang and B. A. Wandell, "A spatial extension of CIELAB for digital color-image reproduction". *Journal of the Society for Information Display*, vol. 5 (1), pp. 61–63, 1997. *Cited in Sec.* 1.2.1, 4.3.1, B.5.3