Context-based Quality of Experience in Immersive Multimedia

Thèse N° 7272

Présentée le 18 janvier 2019

à La Faculté des sciences et techniques de l'ingénieur
Groupe Ebrahimi
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

ANNE-FLORE NICOLE MARIE PERRIN

Acceptée sur proposition du jury

Prof. P. Frossard, président du jury
Prof. T. Ebrahimi, directeur de thèse
Prof. F. Pereira, rapporteur
Dr M. Rerabek, rapporteur
Prof. S. Süsstrunk, rapporteuse

2019

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

This is not the end. It is not even the beginning of the end.
But it is perhaps, the end of the beginning.
Winston Churchill

Acknowledgments

I would like first to thank my thesis advisor Prof. Touradj Ebrahimi of MMSPG at EPFL. This position was a lifetime opportunity that I am grateful for. The provided environment of research was ideal, and the comments I received were always to the point.

I would also like to thank the experts who were involved in the thesis jury, Pascal Frossard, Sabine Süsstrunk, Fernando Pereira, and Martin Rerabek. Thank you for the fruitful and encouraging feedbacks. Without your dedicated participation and input, the thesis could not have been successfully corrected.

I had the immense pleasure of having numerous and various partners for most of my contributions.

- I would like to thank all down to Dolby Laboratories that have been involved in our partnership. I especially thank David Brooks for his exhaustive and precise help in defining our common studies, Sherif Gallab for the wonderful encounter and the exciting exchanges on High Dynamic Range (HDR) software and hardware, and Walt Husak to make this collaboration happen.

- From B<>com, I had the pleasure to collaborate with Cambodge Bist. He has been one of the nicest and brightest workmates I have worked in pair with. I am also addressing very special gratitude to Eric Cozot, who has been a former Professor of mine. He always had a huge positive and determinant influence in my life. I am gratefully indebted to him for this.

- When interacting with Deutsche Telekom (DT), I enjoyed the exchange of knowledge between Saman Zatootaghaj, Steven Schmidt, and me. I have learnt a lot from our collaboration. I also want to address my gratitude to Sebastien Möller for his guidance and comments all along this work.

- My trip to Norvegian University of Science and Technology (NTNU) was one of the most breathtaking I had undertaken.

  I greatly value all the time spent there with Samir Mahmalat. I cannot find words that express all the experiences we have shared, both professionally and personally. I enjoyed all the in-depth and various discussions we had. I am delighted to count him nowadays among my friends.

  I want to thank Sebastian Arndt for our fruitful and passionate debate about physiological signals. My thanks are also addressed to Wendy-Ann Mansilla, who has cleared her busy schedule to organize us a workshop on filmmaking

and to guide us during our stay. I direct gratitude to Andrew Perkis, who made sure we tasted what it's like to be at the Norwegian technical institute and to feel as welcome as possible.

- I would like to mention that some works were the results of shared efforts with students from the EPFL, in the context of teaching duties. I am glad I have met and worked with both Floriane Gilliéron and Bertrand Champenois, whose dedication and enthusiasm were stimulating.

- My work is a part of the QoE-Net project. I want to thank all the parties involved in this project. I do not count all engaging and captivating moments we shared. I have learnt a lot from all supervisors and enjoyed every interaction I had the chance to have with them. Thank you Maria, Luigi, Bernard, Sebastien, Alexander, Andrew, Tibor, and Lingfen.

  I obviously cannot forget to mention the ESRs, who I mostly call friends nowadays. They have been a strong technical and personal support all along this thesis. I am actually sad that this project is now over as I will have fewer opportunities to interact with them. My sincerest thanks to Samir, Werner, Avsar, Nabajeet, Arslan, Saman, Elisavet, and Peter.

I had a great pleasure, and I can even say honor, to have shared my lab with beautiful people.

First, Martin, Philippe, Eleni, and Lin, with whom I learnt all about subjective assessment and test design. They also bore my poor level of English at my arrival. A special mention is directed to Martin, who has been following me closely during my first works. He kept questioning me so that I define further or structure my thoughts. I could rely on him anytime and appreciated it much.

Last but not least, Irene, Evangelos, Evgeniy, Pinar with whom I share countless marvelous memories. We faced best and hardest times together. They have always been there whenever I ran into trouble spot or had a question about my research or writing. They steered me in the right direction whenever I had doubts and needed it. They also represented a strong moral and emotional support. I will never be able to thank them enough for their presence and engagement.

I also always found unconditional supports from former colleagues or new friends met during meetings and conferences. I particularly want to name Patrick Gioia and Jean-Marc Vesin.

I would also like to acknowledge my mum, Pinar, Irene, Jean-Marc, and Samir as the second (third, fourth, etc.) readers of this thesis. Their generosity and involvement moved me and will never be forgotten.

From a more personal perspective, I am here mentioning people who have supported me all along the way. First, friends from the corridor and broadly the EPFL: Ashkan, Beril, Damien, Helena, Hermina, Jean-marc, Leila, Marina, Meloivia,

Sasan, Sibille, Tolis and Vijay. Second, long-term friends, namely Yona, Thais, Nana. "Les Phéromones sont làààààà!"

Finally, I must express my very profound gratitude to my family, especially my parents and my siblings, for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. They have always been supportive and encouraging, or even pushy when I had hard times to do it myself. I will never be able to write how profound is my gratitude and how happy I am to have them in my life.

0.1 Abstract

Since 1895, when the Lumière brothers came out with the projected cinematograph, motion pictures and more generally multimedia contents have considerably improved and are still experiencing a very dynamic development. Advances involved improving various modalities that influence the Human Visual System (HVS) or the human perception. Indeed, migrations concerning audio, resolution, color, brightness, depth representations and frame rate have been performed or are being implemented.

A high-quality multimedia content targets at least an acceptable level of consumer's satisfaction. It assumes to be a combination of sensory stimuli whose goal is to entertain, to educate or to inform, and can convey an artistic intent. Measuring the end-user satisfaction is a challenge as it is a complex multi-dimensional concept.

To understand, measure and predict user satisfaction, researchers focused on the perceived quality of multimedia contents. Quality is the expression of an individual comparison and judgment process. It is based on inherent characteristics of a stimulus. It turns out that network-related impairments, major distortions in content delivery, are not considered when evaluating contents' inherent features. The quality measure considering the entire service pipeline (including delivery) is referred to as Quality of Service (QoS). A more user-centric quality concept has been created under the consideration that not only technical quality and delivery artifacts influence an experience. Accordingly, Quality of Experience (QoE) represents the "user's degree of delight or annoyance for an application or service in regards to his prior knowledge, expectations and current state".

This work addresses the design of new methodologies for QoE subjective evaluation. The research includes identification, definition, and investigation on influence factors of QoE. Possibility to predict QoE using physiological signals gathered during subjective evaluations is also considered.

This thesis is positioned as a framework for the assessment of emerging technologies. Its primary focus is on High Dynamic Range (HDR), and Wide Color Gamut (WCG) representations, for these emerging technologies provide faithful and realistic content; immersive technologies, such as 360° imaging and Virtual Reality (VR) gaming; as well as the combination of typical multimedia representations (e.g. Standard Definition (SD), High Definition (HD) and 4K contents along with no audio, stereo and surround audio).

From conducted experiments on HDR and WCG, important recommendations came out for these technologies' deployment regarding representation and compression strategies while indicating future directions for QoE assessment.

We designed a new subjective methodology for 360° HDR contents. It investigates which dimensions should be included in immersive and realistic contents evaluation while considering representation constraints (e.g., no spatial pair comparison possible for 360° contents).

Expectations and novelty effect evaluations, two technologies-related aspects that influence QoE-centric subjective assessment, have been implemented in the context of VR gaming.

Finally, attempts to predict the Sense of Presence (SoP), a critical QoE-related notion when analyzing physiological signals (brain activity (Electroencephalography (EEG)), heart activity (Electrocardiography (ECG)) and respiration) have been performed for various multimedia consumption scenarios (in terms of quality, resolution and sound systems as well as typical use cases).

0.2 Résumé

Les technologies multimédia connaissent un développement considérable depuis l'invention du cinématographe par les frères Lumière en 1895. Plusieurs axes d'amélioration ont été envisagés, en commençant par l'influence du système visuel humain (HVS) sur la perception humaine. En effet, des progrès majeurs ont été fait ou sont en cours dans les domaines de l'audio, de la résolution, de la couleur, de la luminosité et du contraste, de la fluidité du mouvement et pour finir en matière d'immersion.

On dit d'un contenu multimédia de haute qualité qu'il doit créer chez le spectateur un sentiment de satisfaction et plaisir. En effet, une telle stimulation multisensorielle a pour but de divertir, d'informer ou d'éduquer.

Pour comprendre, évaluer et prédire la satisfaction d'un téléspectateur, les chercheurs se sont penchés sur l'étude de la qualité perçue des contenus multimédia. Pourtant, il semblerait que les attentes du spectateur et ses expériences passées ont une grande influence sur son jugement et devraient donc être intégrées dans l'évaluation de sa satisfaction. Suite à cette observation, une mesure de satisfaction, plus centrée sur l'utilisateur, du service multimédia a été créée: la qualité d'expérience (QoE). Elle mesure le degré de satisfaction ou de mécontentement procuré par un service ou une application, en tenant compte des attentes, des expériences passées et du contexte de l'utilisateur.

Une évaluation subjective est le moyen le plus précis pour quantifier la satisfaction d'un téléspectateur. Il faut bien sûr s'assurer de la validité de la méthodologie suivie lors de l'expérience. Les résultats collectés durant une expérience sont d'une grande valeur comme ils permettent une analyse approfondie d'un effet.

Les travaux entrepris lors de cette thèse examinent de nouvelles méthodologies de test subjectifs pour l'évaluation de la QoE. Sont inclus l'identification, la définition, l'étude et la prédiction des facteurs influençant la QoE. Le contexte d'évaluation se concentre sur des technologies immersives et émergentes. Les technologies étudiées incluent les représentations HDR et WCG, certaines technologies immersives ainsi que la combinaison de représentations traditionnelles.

Suite aux études menées sur les représentations HDR et WCG, sont formulées des suggestions concernant la représentation et la compression de ces flux tout en indiquant quels sont les développements futurs nécessaires pour l'évaluation de la QoE. Une nouvelle méthodologie de tests subjectifs pour les contenus omnidirectionels en HDR a été conçue en identifiant puis évaluant les facteurs impactant le réalisme et l'immersion procurées par ces expériences multimédia. Les contraintes

dues à la visualisation de ces contenus, comme l'impossibilité d'une comparaison spatiale de ces contenus, ont été surmontées. Après avoir identifié deux importants facteurs d'impact sur la QoE, l'effet de nouveauté et les attentes du téléspectateur, ces derniers ont été étudiés dans le cadre de jeux vidéo en réalité virtuelle. Enfin, il a été entrepris de prédire la sensation de présence (SoP), une importante notion proche de la QoE, grâce à l'analyse de signaux physiologiques (activité cérébrale via l'encéphalographie (EEG), l'activité cardiaque via l'électrocardiographie (ECG) et la respiration). Différents scenarios ont été envisagés lors de cette étude, de la combinaison de différents niveaux de qualité, résolution et son à la considération de cas d'utilisation plus réalistes.

Keywords: Quality of Experience (QoE), Subjective evaluations, High Dynamic Range (HDR), Omnidirectional (360°), Virtual Reality (VR), Physiological signals, Electroencephalography (EEG), Electrocardiography (ECG), Expectations, Compression.

Contents

List of Figures

List of Tables

List of Acronyms

ACR Absolute Category Rating

AI Artificial Intelligence

AMC Advanced Media Coding

ANOVA Analysis of Variance

ANS Autonomic Nervous System

AQ perceived Audio Quality

AR Augmented Reality

ATSC Advanced Television Systems Committee

AV1 AOMedia Video 1

BBC British Broadcasting Corporation

BCI Brain-Computer Interface

BL Base Layer

bpm beats per minute

BVP Blood Volume Pressure

CC Codec Configuration

CCM Compound Content Management

CCR Comparison Category Rating

CfE Call for Evidence

CI Confidence Interval

CIE Commission Internationale de L'Eclairage

CMYK Cyan Magenta Yellow and Key color model

CNS Central Nervous System

CL Constant Luminance

CRT Cathode Ray Tube

CSP Common Spatial Pattern

DCF Design rule for Camera File

DCM Dynamic Causal Modeling

DCR Degradation Category Rating

DEAP Database for Emotion Analysis using Physiological signals

DLBC Dual-Layer Backward-Compatible

DMCVT Dynamic Metadata for Color Volume Transform

DMOS Differential Mean Opinion Score

DR Dynamic Range

DSLR Digital Single-Lens Reflex

DSCQS Double-Stimulus Continuous Quality-Scale Method

DSCS Double Stimulus Comparison Scale

DSIS Double Stimulus Impairment Scale

DT Deutsche Telekom

DVB Digital Video Broadcasting

DVD Designed Viewing Distance

E-P $expectation - perception$

EBU European Broadcasting Union

ECG Electrocardiography

EDA Electrodermal Activity

EEG Electroencephalography

EETF Electrical-Electrical Transfer Function

EHRI Extremely High Resolution Imagery

EL Enhancement Layer

EMG Electromyography

EOG Electrooculography

EOTF Electro-Optical Transfer Function

EPFL Ecole Polytechnique Fédérale de Lausanne

ERP Event-Related Potentials

ETSI European Telecommunications Standards Institute

EV Exposure Value

Exif Exchangeable image file format

fMRI functional Magnetic Resonance Imaging

fNIRS functional Near-Infrared Spectroscopy

FN False Negative

FOV Field Of View

FP False Positive

fps frames per second

FS Familiarization Session

GEQ Game Experience Questionnaire

GES Geodesic EEG System

GPS Global Positioning System

GQ Game Questionnaire

GS Group Specification

GSN Geodesic Sensor Net

GSR Galvanic Skin Response

H.264 Advanced Video Coding, also referred to as MPEG-4 AVC or H264

HCI Human-Computer Interaction

HD High Definition

HDR High Dynamic Range

HEVC High Efficiency Video Coding

HF High Frequency

HLG Hybrid Log-Gamma

HMD Head Mounted Display

HR Heart Rate

HRV Heart Rate Variability

HSI Hue-Saturation-Intensity

HSL Hue-Saturation-Lightness

HSV Hue-Saturation-Value

HVS Human Visual System

IaG Intra-Group Analysis

IA Interest in Audio content

IBL Image Based Lighting

ICA Independent Component Analyses

IeG Inter-Group verification

IEQ Immersive Experience Questionnaire

IL Immersiveness Level

IP Internet Protocol

IPC Independent Principal Components

IPQ Igroup Presence Questionnaire

IRT Institute of Research and Technology

ITU International Telecommunication Union

ISO International Standards Organization

IV Interest in Video content

JPEG Joint Photographic Experts Group

JND Just-Noticeable Difference

kNN k-Nearest Neighbor

LCD Liquid Crystal Display

LF Low Frequency

LMS Long-Medium-Short response

LUT Look-Up Table

Mbps Megabits per second

MMR Multi-Mapper Resolution

MMSPG Multimedia Signal Processing Group

MOS Mean Opinion Score

MP Megapixel

MPEG Moving Picture Experts Group

MR Mixed Reality

MRI Magnetic Resonance Imaging

NCL NonConstant Luminance

NHK Nippon Hōsō Kyōkai

NTNU Norvegian University of Science and Technology

NTSC National Television System Committee

OB Overall Bitrate

OETF Opto-Electronic Transfer Function

OOTF Opto-Optic Transfer Function

OPQ Open Profiling of Quality

OQ perceived Overall Quality

PAL Phase Alternating Line

PANAS Positive and Negative Affect Schedule

PC Pair Comparison

PCA Principal Component Analysis

PeEQ Pre-Experiment Questionnaire

PeFQ Pre-Familiarization Questionnaire

PENS Player Experience of Need Satisfaction

PET Positron Emission Tomography

PNA Personal Navigation Assistant

PNS Peripheral Nervous System

PoEQ Post-Experiment Questionnaire

ppd pixels per degree

PQ Perceptual Quantizer

PrQ Presence Questionnaire

PSD Power Spectral Density

PVD Preferred Viewing Distance

QEEMS International Young Researcher Summit on Quality of Experience in Emerging Multimedia Services

QHD Quad High Definition

QoE Quality of Experience

QoP Quality of Perception

QoS Quality of Service

QP Quality Parameter

QUL Quality and Usability Lab

RBF Radial Basis Function

Rec. Recommendation

RExt Range Extension

RGB Red-Green-Blue color space

SA Surrounding Awareness

SAMVIQ Subjective Assessment of Multimedia Video Quality

SD Standard Definition

SDR Standard Dynamic Range

SDSCE Simultaneous Double Stimulus for Continuous Evaluation

SbS Side-by-Side

SC Sanity Check

SECAM Sequential Colour with Memory

SERVQUAL multiple-item scale for measuring consumers perceptions of service quality

SI Spatial perceptual Information

SLBC Single-Layer Backward-Compatible

SLNBC Single-Layer Non-Backward-Compatible

SMPTE Society of Motion Picture & Television Engineers

SNS Somatic Nervous System]

SoP Sense of Presence

SS Single Stimulus

SSCQE Single Stimulus Continuous Quality Evaluation

std standard deviation

SVM Support Vector Machine

TF Transfer Function

TI Temporal perceptual Information

TN True Negative

TP True Positive

TS Technical Specification

TV Television

TMO Tone Mapping Operator

UEQ User Experience Questionnaire

UHD Ultra High Definition

UX User Experience

VE Virtual Environment

VoD Video on Demand

VQEG Video Quality Experts Group

VR Virtual Reality

WCG Wide Color Gamut

ZHAW Zurich University of Applied Science

4K Ultra High Definition

Chapter 1

Introduction

Contents

Scientists rely on observations since centuries to develop their knowledge. Through history, tremendous examples illustrate the acceptability to confirm a theoretical assumption empirically or to form a model based on experiments.

Researchers mimic the learning process that any human goes through when learning. Human acquisition of knowledge, also referred to as cognition, roots from perceptual experiences [Chadha 2010]. Since birth, a baby sees parents and relatives walking around. When strong enough, a baby tries to reproduce what is seen and attempts to walk. Improvements of the baby rely upon perceptual experiences: observing and seeing other humans or feeling, for instance, pain when not making the right move or when falling. Perceptual experiences participate in the formation of knowledge necessary to get to walk.

In the Cambridge dictionary, there are two ways to define experience. First, experience is "(the process of getting) knowledge or skill from doing, seeing, or feeling things". Experience is also "something that happens to you that affects how you feel". The COST Action IC1003 (Qualinet) consortium combined both views when expressing experience as "an individual's stream of perception and interpretation of one or multiple events." [Brunnström 2013]. Experience is both the act of going through an event and processing this event in cognitive constructions.

Cognition arises within an individual when encountering an event that provokes reactions in oneself. It is a passive mental reaction to experiences, neither an instrument in the act of knowing nor a mere sensory awareness [Chadha 2010]. The formation of knowledge, memory, judgment, evaluation, decision making and problem-solving constitute what is referred to as cognition. From cognition are derived expectations, personality, and the evaluation of both context and system

performance. In that sense, past and present experiences determine the perception of future experiences.

Cognitive functions are individual-specific. However, a group of individuals can share constructions. For instance, a French person is more likely to have gone through similar experiences as another French than those experienced by a Pakistanis. Consequently, it is reasonable to assume that people sharing features developed joint constructions and beliefs and will react similarly to the same event [Möller 2014]. Likewise, customers with comparable characteristics show more similarities in their cognitive constructions, and so in their evaluations of experience. This fact exhibits the importance of understanding the profile of the targeted audience when providing a service.

A user profile comprises the set of characteristics representing the consumer [Brunnström 2013, Möller 2014]. This set of human influence factors includes physiological features such as visual acuity, being color blinded or right-handed, as well as cultural and socio-demographic profiles, the present emotional state or the expertise of the user regarding a technology, to name but a few.

As stated earlier, the user profile influences how an event is perceived and provides meaningful information to understand, and ultimately predict, the subjective evaluations of perceptual experiences. Predicting the perceived quality of an experience is a milestone when developing multimedia technologies, products, and services. Indeed, to develop and improve services to reach high quality, usability, efficiency, and enjoyability, one needs the ability to evaluate the experience provided by the system. Thanks to evaluations, service providers can improve their product and ultimately reach larger market shares.

An evaluation can be subjective, through experiments run on a representative sample of the population. Statistic tests are applied to the gathered subjective measures to conduct a comprehensive analysis of what is evaluated. Although being the most reliable way to appraise a system, subjective evaluations are time, workforce, resources and money consuming. To tackle this issue, it is current to develop objective metrics. These have the same purpose as subjective evaluations while being mathematical models. Inherently, objective metrics are an approximation of the measure and are thus less reliable than subjective analyses. Let us quote Blythe and Monk in "Funology 2: from usability to enjoyment" [Blythe 2018] who stated the following about objective models: "They cannot and will not replace details inquiries into situations, practices, and experience. But they systemize, spurn debates about what belongs and what not."

Multimedia services extensively rely on subjective evaluations. These experiments are considered as the source of quality evaluation as, in addition to provide analyses of a system, subjective tests are also used to validate objective models. We stress here the necessity to put efforts in developing accurate, reliable and valid subjective evaluations methodologies.

The initiative to include user profile in service quality evaluations has been undertaken in the multimedia field for about two decades, and even more actively since 2011. At this point, the multimedia ecosystem guaranteed great experiences to TV

consumers. Faithful High Definition (HD) streams are delivered while taking advantage of advanced high-efficiency compression tools (e.g., Advanced Video Coding, also referred to as MPEG-4 AVC or H264, VP9 and later on High Efficiency Video Coding (HEVC) and AOMedia Video 1 (AV1)), adaptive streaming and satisfactory network management. Considering the maturity of multimedia technologies, it is therefore not reasonable anymore to assume that system performance (quality) or the efficiency of the entire delivery pipeline (Quality of Service (QoS)) are accurate predictors of broadcasting service quality.

An accurate predictor of high service quality is the end-user satisfaction concerning the provided perceptual experience. Though, it is more difficult for a service provider to aim for customers' satisfaction (e.g., more satisfying or enjoyable experience) than targeting system and delivery efficiency. Indeed, service performance is evaluated on features which are precise, measurable and tangible (e.g., delay, packet loss, true colors, amount of blur). Satisfaction models need to be adapted towards the specificities of the tasks. Besides, user satisfaction is highly subjective and dependent on the customer's needs, expectations and former experiences.

The Qualinet consortium has been created to address this concern and to take the first steps towards future metrics for perceptual quality prediction. Qualinet initiated the transition from usual quality and QoS assessments to QoE evaluation. More reliable and accurate than its predecessors, QoE puts the user at the center of multimedia experiences appraisal. To predict the subjective service quality, QoE keeps considering systems efficiency and the reliability of delivery while it extends concerns about the user profile and the context of use. QoE encompasses evaluations of the source signal, context, and system influence factors as well as human factors. In other words, QoE covers the dichotomy of quality; on one hand, the system and contextual factors which are expected to influence perception, and on the other hand, the resulting perception of the user, regarding quality features.

The major contribution of the consortium is the definition of QoE in 2013, accepted by the entire research community of this field, and standardized by the International Telecommunication Union (ITU) in 2016:"**Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/ or enjoyment of the application or service in the light of the user's personality and current state.** [Brunnström 2013]."

Besides, the Qualinet community established a strong background addressing concepts and theories related to QoE, highlighting the types of application for which QoE assessment is relevant and also digging deeper in several applications to initiate the use of QoE for information and communication systems and services [Möller 2014].

In details, researchers inspired from the developments in Human-Computer Interaction (HCI), mainly related to the concept of user experience, to form the basis of QoE. Both notions share numerous characteristics. However, QoE relates to the

emotions and needs of users while user experience is still assessing experiences from the perspective of service providers. With QoE, service providers aim to go beyond the usability of systems. After defining precisely QoE, long discussions exhibited the classification of influence factors in human, technical system, and context of use features. It has been agreed that QoE results from the perception of a multi-dimensional event. As such, new empirical and psychophysical methods are needed to extract significant dimensions. Meanwhile, analyses have been conducted to relate physiological signals to (1) quality degradation, (2) emotions evoked by different types of media, (3) cognitive and memory-related processes. Additionally, efforts have been deployed to encourage progress in evaluations for interactive multimedia.

The applications envisioned were broad and diverse. They encompass (1) audio and video systems (speech communication, text-to-speech systems, audio-visual broadcasting such as conferencing and telemeeting, audio coding and transmission, spatial audio services), (2) new interaction modalities (e.g., haptics controls, wearable technologies), (3) complex quality prediction models (e.g. to map QoS to QoE based on reality and enjoyment skews, Quality of Perception (QoP) [Ghinea 2008] or Strohmeier's Open Profiling of Quality (OPQ) [Strohmeier 2010]), (4) identification of perceptual quality features as well as technical influence factors (e.g., for 3D video services), (5) use of crowdsourcing in QoE evaluation (the whys and wherefores), (6) web browsing, (7) Gaming QoE (taxonomy), (8) mulsemedia (role of sensory experience beyond audio-visual and how to specify and evaluate sensory effects such as force feedback, background light, wind and odor), (9) research paradigms, goals and questions, as standardized methods are still largely missing.

The last point is the starting point of our reflexion. There is a need for standardized methodologies to assess QoE. The present state of the art (widely accepted definition of the concept, several works on influence factors and objective models, on numerous applications) forms a firm basis on which we can rely and act on for establishing recommendations.

With the transition from conventional multimedia communication to various enriched and immersive technologies (such as 3D, Ultra High Definition (4K), HDR, 360°), elegant and complex approaches are necessary to develop reliable, effective and more importantly mutual metrics predicting subjective quality. Metrics and measures must be, as highlighted in the above, context-dependent. In our research, we focused on the use of different multimedia technologies, which define our context of use. The idea behind this decision is that each technology relates to specific expectations, habits of consumption and user behavior. Therefore, system, context, and human influence factors need then to be redefined on a case-by-case basis.

Proposing subjective methodologies, held in common by most multimedia technologies, which offer accurate ways to extract and evaluate influence factors is a challenge we addressed. It answers the need of the entire community to move forward in the understanding and prediction of QoE, to ultimately provide consumers

with satisfying multimedia services. Indeed, services providers crave for accurate and efficient methodologies on which they can rely on to evaluate their solutions.

A small anecdote: at the beginning of my position, I took part in a meeting for the H2020 project (No 688619) ImmersiaTV. One of the attendees directed a question to me: "Anne-Flore, as a QoE expert, could you tell us what are the practices or protocols that are used to evaluate QoE". I found myself speechless and wondering what to answer. There are so many different ways to evaluate a system, especially in terms of QoE, that there is no methodology being at once clear, "simple" and suiting all contexts. However, this is precisely what was missing in the QoE ecosystem although being much required. It accurately illustrates the aim of my work.

This thesis reports the activities performed to investigate the context-based QoE in immersive multimedia within the framework of the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Action QoE-NET, grant agreement No 643072.

The contributions of the performed works are detailed below. Then, a related-work chapter introduces the technical background necessary to a non-expert reader. It also presents actions taken by the community regarding multimedia technologies and subjective assessment. Then, we report comprehensively our activities aiming at recommendations for context-based QoE in immersive multimedia. Chapters are organized following the same order as the section below. Lastly, we provide a concise summary of the outcomes of our research, sorted by immersive-technologies- and QoE-related findings. Besides, clear key takeaways and future perspectives are stressed at the end of the conclusion.

1.1 Contributions

The contributions of this thesis are categorized into four parts.

The first part supports the current transition from SDR imagery to HDR and Wide Color Gamut (WCG). The new format represents more accurately luminances and chrominance information. The resulting experience is more realistic and more immersive as more faithful to real stimuli. Two works were conducted for the development of HDR and WCG recommendations and growth on the market. We conducted two subjective quality assessments on HDR and WCG representations and delivery. From our conclusions, we draw guidelines towards advanced HDR imaging, and we pointed out two strategies to evaluate QoE.

Then, we explored the said strategies to conduct QoE-driven evaluations. On one hand, we tackled the extraction of systems and context influence factors. The use case addressed was the evaluation of HDR 360 ° images.

On the other hand, we investigated user-based features closely related to technical and contextual factors. Expectations and novelty effects impact every technology, and fulfills the previous constraint. We took advantage of VR gaming, that is a relatively new technology on the market, to study a broad diversity of user profiles,

especially concerning expectations.

Lastly, we investigated whether the use of physiological signals in subjective evaluations may lead to accurate predictors of the Sense of Presence (SoP). We argue that SoP is one of the most significant high-level concepts of QoE for immersive multimedia. Observing and predicting SoP through physiological signals is a major step toward QoE prediction and measure.

In the following, we report the motivation and details about the four parts mentioned above.

1.1.1 Support the transition from SDR to HDR

As presented in Sections 2.1.3 and 2.1.4, HDR and WCG are promising realistic scene representations. However, faithfulness has a cost which results, in this context, in need for memory space to embed information. Content representation and compression are the current main lines of investigation to optimize the use of memory resources for HDR imaging.

Dolby is an actor in the development of HDR imaging. The teams of the firm have proposed compelling and comprehensive solutions tackling numerous issues from representation to rendering (e.g., Dolby Vision ^{TM}, ICtCp color space, and the Perceptual Quantizer (PQ) Transfer Function (TF)). We collaborated to offer fair comparisons of their solutions with classical systems.

We evaluated two solutions proposed to improve the HDR format. These solutions consisted of a new color space, designed particularly for wide ranges of luminance and color, and an Electrical-Electrical Transfer Function (EETF) which optimizes the code words distribution of a stream before its compression.

This fruitful partnership was extended to a second work. Dolby Vision ^{TM} is an end-to-end HDR and WCG framework which implements state-of-the-art compression solutions. We seized this opportunity to study which compression scheme is the most compliant with broadcasters constraints (backward-compatibility of codecs) and multimedia service requirements (high-quality SDR and HDR delivery).

While important recommendations for the HDR ecosystem result from this partnership, we sharpened our strategies for context-based QoE evaluations. Besides, we became fully aware that influence factors depend mainly on the immersive technologies used.

1.1.2 Contextual- and system-based influence factors for HDR 360 ° imaging

Based on the expertise gained during the two previous quality assessment experiments, we identified the need to inquire for more details to further understand QoE. In this chapter, we focused on influence factors related to the context and technical features of the system.

HDR 360 ° imaging combines the functionalities of two interesting and immersive technologies. There is an need to understand how both technologies strengths

and weaknesses mingle. Moreover, it is worth knowing if the combination of both promising technologies yields to even more valuable perceptual experiences.

The study included the capture of media stimuli, their conversion to compliant rendering systems, the creation of a pair-comparison subjective evaluation methodology for omnidirectional imaging, the design of the experiment, the implementation of a testbed and the analysis of the subsequent results.

We highlighted the importance of technical features of the system but also considerations about aesthetics or interactions in subjective evaluations of QoE.

1.1.3 User-centric influence factors for VR gaming

After investigating context and system influence factors, we meant to explore human-related factors. We conducted an exploratory research on the inclusion of the novelty effect in QoE evaluations of emerging technologies. This work hinted to study user expectations. These are worth exploring, as, in addition to be part of the main influence factors of QoE, they are strongly related to the novelty effect.

Therefore, we prospected further the evaluation of user expectations as human influence factors. A cross-platform comparison was conducted to observe whether mobile, computer and VR Head Mounted Display (HMD) gaming platforms induce different levels of QoE while playing the same game. We assumed that different expectations were formed towards each gaming platform. We studied user expectations before and after the test to also relate QoE to the evolution of expectations. Additionally, the balance between unimodal and multimodal evaluations was considered in this context. That is, there is a trade-off to reach between evaluating one dimension of an experience and appraising too many influence factors.

1.1.4 Prediction of QoE: physiological signals

Lastly, what stroke me particularly is that subjective evaluations rely on explicit subjects ratings. It means that various biases (e.g., experimental, cognitive) are impacting the collected scores.

Meanwhile, more and more attention is given to physiological analyses for emotional, perceptual and cognitive research. We wondered if biosignals could help for QoE understanding and prediction.

We seized this opportunity of considering more implicit evaluation methods. Though more demanding in time and signal processing, physiological signals indicate body reactions directly related to experienced stimuli. Besides, those signals embed temporal information. As QoE varies in time, and individual score only evaluates a stimulus in overall, physiological signals are highly promising.

We focused our work on the prediction of SoP. This notion is a high-level dimension of QoE, especially for immersive multimedia. Also, SoP embraces the three aspects mentioned above, namely emotional, perceptual and cognitive reactions.

Two evaluations verified if the features extracted from Electroencephalography (EEG), Electrocardiography (ECG) and respiration signals could lead to the pre-

diction of the level of SoP. In the first study, SoP is assumed to be related to immersion settings such as quality, resolution and sound systems. The second investigation focused on typical multimedia scenarios, in which SoP is linked to the level of immersion realized by the rendering device.

1.1.5 Conclusion

From our contributions, numerous recommendations for the evaluation of future systems are provided. Primary efforts have been dedicated to the extraction of influence factors and their assessment. More sophisticated subjective evaluations have been considered. Several datasets have been created and shared publicly to encourage and facilitate the research on areas covered during the thesis. Finally, personal insights and future perspective are specified.

Chapter 2

Related work

Contents

The forebear to today's multimedia technologies is the projected cinematograph, invented by the Lumière brothers in France, more than a century ago. Since then, constant efforts of broadcasters and digital media scientists as well as tremendous progress in software, hardware and media representations have led to current multimedia (immersive) technologies.

Multimedia technologies combine sensory stimuli such as visual, audio or haptic signals. They provide a powerful mean of communication and expression, mainly intending to entertain, to educate or to inform, and can convey an artistic intent. When qualified as immersive, multimedia technologies aim to deliver an illusion of

reality which includes the viewer within. They are regarded as more desirable means to deliver satisfying experiences.

Multimedia experiences are a common part of everyday life, from videos watched on Youtube, to advertisements seen on a website or TV programs watched at home. For instance, there is a high demand and consumption for online video content: the proportion of consumer Internet traffic allocated to Internet Protocol (IP) video traffic represented 73 % in 2016, and is expected to reach 82 % by 2021 [1]. Also, the ten-fold increase of Video on Demand (VoD) services such as Netflix, Amazon Prime Video, CBS, HBO, Hulu, Vudu, who compete with early actors like Youtube or British Broadcasting Corporation (BBC) stresses the need for quality multimedia delivery services.

Those multimedia services must keep their consumers satisfied. This means they have two challenges. The first one is to secure and pursue the growth of realistic and immersive technologies. Service providers focus on achieving competitive advantage through new or better experiences. The second consists to accurately measure or predict the level of satisfaction of their consumers. Measuring the level of consumer's satisfaction is essential to achieve the first challenge. Yet the concept of satisfaction is complex: it is a multi-dimensional concept.

In the past decades, the transition from quality to QoE measures indicated the intent to develop more user-centric services. However, no agreement nor recommendation has been reached on how to evaluate QoE. Significant investigations are still seriously needed to accurately evaluate consumers' satisfaction through this measure.

Before introducing our activities addressing the need of satisfaction measurement, this chapter provides the background knowledge necessary to the reading of this thesis. This chapter also introduces progresses achieved in the multimedia field in terms of media technologies and evaluation methodologies.

2.1 Evolution of multimedia technologies

Multimedia experiences take advantage of several perceptual senses, mainly vision and sound. Enhancing such experiences means one wants to enrich the information brought to individuals. For instance, sound is represented on an increasing number of channels, which ultimately recreates volumetric audio sources. Visual information are expanded spatially (resolution) and temporally (frame rate). Also, better reproduction of colors and brightness have been envisioned. Finally, the immersive aspect of media experiences has been explored in terms of depth cues (stereoscopy, light field or point clouds) as well as full coverage of the field of view (360° and VR). The evolutions of audio, resolution, color, brightness and immersiveness representations are presented in the following.

[1] Cisco Visual Networking Index: Forecast and Methodology, 2016-2021, https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html

2.1.1 Audio

Less than a century ago, appeared both the first movie with sound (1926), and the first movie with dialog (1927). Since the beginning of TV with sound, significant advances in technology led to high quality surround audio, and move towards a spatial representation of sound.

Differences between widely distributed sound systems (mono, stereo and surround), developed through the years, relate to the number of independent channels providing audio information. The more channels there are, the more realistic will be the audio experience. By realistic we mean that more ambient and spaciousness effects are offered, which recreate a more lifelike stimulus. In monophonic systems (mono), all audio signals are mixed in a single channel. This system is highly interesting in speech systems as providing all information to all users with the same sound level. Stereophonic systems, or stereo, have two independent audio channels, enabling the spatial localization of sounds. However, such systems should be carefully used as signal interferences and cancellation may appear with more than one signal [Hulsebos 2002]. Moreover, reaching a full and uniform coverage of the room space is usually not possible due to obstacles and sound waves reflections. Comments about signals interference and cancellation also apply to surround systems. Usually designed with five or seven independent channels, they cover quite efficiently the room space. The widely known 5.1 surround setup includes three front speakers (left, center and right), two back speakers (left and right) and a low-frequency effect channel (which is reflected as 0.1 channel in the naming convention). Figure 2.1 shows an example of surround system settings. For more information on audio systems and their settings, the reader may refer to the Recommendation (Rec.) ITU-R BS.775 [ITU 2012a].

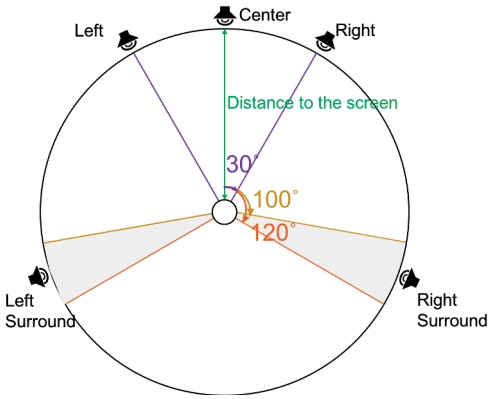


Figure 2.1: Reference arrangement for surround sound system.

Since two decades, efforts have been dedicated to spatial/3D audio, with multichannel systems [Boone 1995, Rabenstein 2006] or virtual sound [Gardner 1999, Herre 2015]. This emerging audio setup aims to fully immerse users in experiences, which is one of the factors that increase users' QoE.

2.1.2 Resolution

The resolution of multimedia content may be expressed through various measures, e.g., pixel resolution (number of total pixels in the image), angular resolution (pixels per minute of angle), spatial resolution (pixels per inch) and, temporal resolution (frames per second (fps)). We focus here on pixel resolution of display and consider only square pixels.

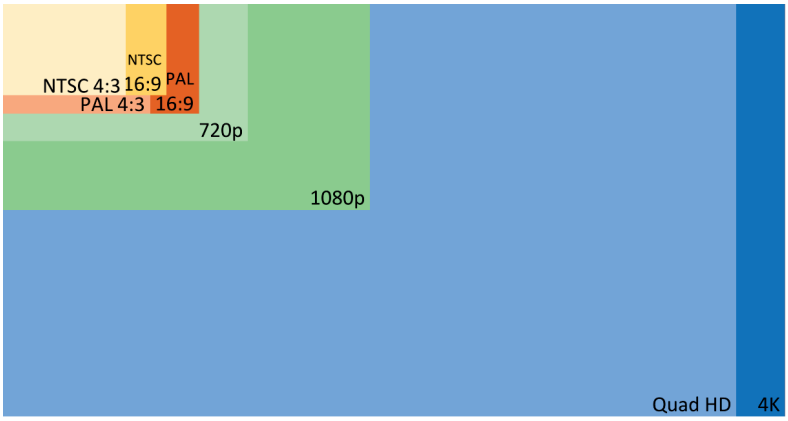


Figure 2.2: Common display resolutions

Pixel resolution, also named resolution, is defined as a set of two positive integer numbers, which indicate the number of pixel column (width) and the number of pixel row (height). It thus indicates the size of the active display area while providing the overall number of pixels realizing the image.

An important aspect of this measure is the aspect ratio (ratio between image width and height), which indicates the rectangle shape of the active area on display. Most used aspect ratios are 4:3 and 16:9 [ITU 2011b]. Defined as the wide-screen aspect ratio, 16:9 is used for Extremely High Resolution Imagery (EHRI) [ITU 2008a] to match the aspect ratio of HD contents.

The increase of pixel resolution led to contents which carry a higher amount of information (e.g., texture details and clear edges). Viewers considerably benefit from it as such information generates a stronger sense of reality and presence due to high-fidelity representations [ITU 2015c]. For instance, for the same size of display, Ultra High Definition (UHD) brings four times more information to the viewer than an HD content. This capability enhances the content towards a more faithful and realistic representation. If the pixel size remains the same, UHD covers a larger field of view than HD. This fact increases viewer's SoP and immersion, which are capital for providing high QoE to a user.

Standard Definition (SD) has been defined by numerous standardization committees such as National Television System Committee (NTSC), Phase Alternating Line (PAL), and Sequential Colour with Memory (SECAM). This explains there is not one pixel resolution related to SD, but several of them, depending on the content characteristics (e.g., frame rate, number of lines, interlaced or progressive

scanning method and aspect ratio). Such variability in digital images and videos formats prevents worldwide broadcasting and broadly used applications. This issue has been tackled when defining highest definition resolution.

Reference document	Resolution		Aspect ratio	*height* × *width*
Rec. ITU-R BT.601, BT.1358	SD	NTSC 480i (525 line)	4:3 16:9	640 × 480 854 × 480
		PAL/SECAM 576i (625 line)	4:3 16:9	768 × 576 1024 × 576
Rec. ITU-R BT.709	HD	720p 1080p	16:9 16:9	1280 × 720 1920 × 1080
Rec. ITU-R BT.1201, BT.1769, BT.2020, BT.2246	UHD	Quad HD 4K (Digital Cinema)	16:9 17:9	3840 × 2160 4096 × 2160
Rec. BT.2020	8K		16:9	7680 × 4320

Table 2.1: Resolution and aspect ratio of digital imagery formats

Table 2.1 introduces various (pixel) resolutions that are mainly encountered nowadays. It also shows where to find properties and definition of these formats, when standardized by the ITU. Figure 2.2 illustrates these most common display resolutions. When the aspect ratio is not indicated, it is 16:9, except for 4K.

Nowadays, the race for increasing display pixel resolution has several purposes:

- The level of details conveyed by a higher number of pixel gives more faithful and realistic contents. This characteristic is essential for viewers' satisfaction. However, one should keep in mind that such amount of data to deliver to the user raises several concerns, among which is the compression scheme to use to match bandwidth limitations. If HEVC and other codecs can encode and decode UHD contents efficiently, this may not be the case for future resolution formats, such as 16K.

- The increase in field of view leads to an increase in peoples' SoP and immersion. It is capital for emerging technologies to provide such experiences to the users.

- Higher resolution formats are needed for numerous emerging technologies. For instance, VR HMDs suffer from a clear lack of resolution. Numerous consumers complained about visible pixels or color aliasing. Indeed, such near-eye displays do not have a sufficiently high definition yet: the Oculus Rift DK1 and the HTC Vive displays provide 7 and 11 pixels per degree (ppd), respectively. However, the perceptual resolution of the Human Visual System (HVS) is 60 ppd, corresponding to an arc-minute angle per pixel [ITU 2015b].

2.1.3 Color

Color perception is based on the human sensory response to visible light waves (wavelengths within the region of 380 nm to 830 nm [Reinhard 2010]). It relies on photo-receptors contained in the human eye retina, meaning about 6.5 million cones and 100 million rods [Pratt 2007]. Rods are highly sensitive to light stimulation through photons. However, rods only render a reduced visual acuity. Cones are responsible for color perception as they implement the Long-Medium-Short response (LMS) response of the HVS. Each of the three different types of cone cells is particularly sensitive to specific light wavelengths: long (L, red), medium (M, green), or short (S, blue). Additionally, those photo-receptors are responsible for visual acuity [Ledda 2004]. This explains that rich and immersive technologies have the purpose of showing accurate and vivid colors.

Before proceeding with color spaces specifications, several important terms must be defined: intensity (I), brightness (Br), luminance (Y), lightness (L*), hue (H) and saturation (S) [Plataniotis 2013, Tyagi 2018], that are often confused in the literature. First, light is considered in radiometry as radiant energy, expressed in Joules. Intensity measures light energy that radiates from or incident on a surface. It is expressed in watts per square meter. Brightness is attributed to the visual sensation that an area emits more or less light. Since brightness relates to the complex process of perception, a quantifiable measure has been defined by the Commission Internationale de L'Eclairage (CIE): the luminance (cd/m^2 or nits), which is the radiant power weighted by a spectral sensitivity function that is characteristic of human vision. Human vision has a nonlinear perceptual response to luminance. The resulting response is called lightness (L*). The nonlinearity model is assumed to be roughly logarithmic. The human interpretation of colors relies on the luminance (luma), hue and saturation components. Hue and saturation together describe the chrominance (chroma). Hue is a color attribute, associated with the dominant color perceived by an observer. Saturation is the extent of white light that is mixed with pure color. The pure spectrum colors, indicated by hues, are saturated and carry no white light.

Based on the fact that three types of cones are responsible for color perception, three components are necessary and sufficient to accurately model colors, under weighting function constraints [Tyagi 2018]. Inspired from the HVS or from colorimetry considerations, color models are sorted in four categories [Plataniotis 2013].

1. Colorimetric color models that are based on physical measurements of spectral reflectance (e.g., CIE XYZ tristimulus values color space).

2. Psychophysical color models, which mimic human interpretation of color (e.g., Hue-Saturation-Lightness (HSL), Hue-Saturation-Value (HSV), and Hue-Saturation-Intensity (HSI)).

3. Physiologically inspired color models, which define the three primaries colors related to the three types of cones in human retina (e.g., Red-Green-Blue color space (RGB)).

4. Opponent-color models that are based on empirically studies on human perception and implement pairwise opponents primary colors such as yellow-blue and red-green color pairs (e.g., IPT [Ebner 1998b, Ebner 1998a], and YUV color space family).

Every color space is designed to target specific applications. HSL color space family is especially efficient for digital image processing (e.g., edge detection, segmentation, skin detection or background subtraction), mainly due to the separability of chromatic and achromatic values and the efficiency of hue-based segmentation. YUV family of color space, composed of one luminance and two chrominance channels, are especially recommended for compression operation as they enable chroma subsampling. RGB-like color spaces are in line with color models used by hardware devices such as cameras, displays or color image scanners.

A comparison, in terms of specification, pro and cons, between common color space in digital imaging (e.g., Munsell, RGB, CIE XYZ, CIE Lab, HSV, YUV, and YCbCr) are presented in [Ibraheem 2012]. For a complete description of color spaces and their application to color image processing, we recommend [Poynton 1995, Pitas 2000, Pratt 2007, Plataniotis 2013, Tyagi 2018] to the reader.

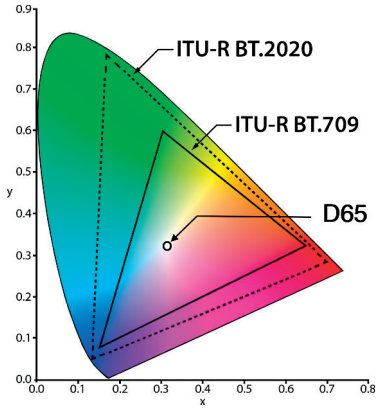


Figure 2.3: Representation of BT.709 and BT.2020 gamuts on the CIE xy chromacity diagram

CIE XYZ tristimulus values model can represent the entire spectrum of visible color; however, this fact is not true for all color spaces. CIE xy chromacity diagram (1931) is a representation of the visible color spectrum. The portion of visible light represented by a color space, named color gamut, can be displayed on the chromacity diagram as the triangle formed by the three primaries of a color space. Figure 2.3 presents the CIE xy chromacity diagram. An important indication on this scheme is also the color space white point, which is reached when three color channels are contributing equally. When the display white point is unavailable, the illuminant D_{65} is often applied as reference light. It is defined as natural daylight light source with a color temperature of 6500 Kelvin [Reinhard 2010].

Within the last decades, major developments have been seen in color representations, especially with the definition of WCG. With the extension of hardware

capabilities and the shift of digital imaging representation from 1 byte onto 10, 12 or even more bits per pixel per channel, new richer color representations are enabled. As Figure 2.3 shows, when compared to legacy colorimetry standard Rec. ITU-R BT.709 [ITU 2015e], the WCG standard Rec. ITU-R BT.2020 [ITU 2015a] is represented by a larger gamut. This depicts the expansion of the color palette ultimately leading to viewers' enhanced visual experience.

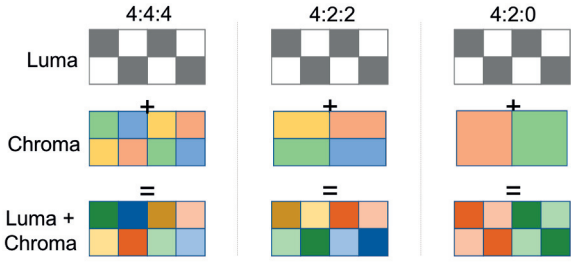


Figure 2.4: Illustration of the resolution of luma and chroma channels depending on subsampling format.

Compression solutions implement chroma subsampling on account of the human perception, which is more sensitive to luminance than chrominance. It is recalled that there are about 15 times more rods than cones in the eye retina. To reduce the delivery of unnecessary information, compression operators subsample chroma data. This explains the interest of color space with one luminance and two chrominance components. In terms of notations, 4:4:4 refers to a high fidelity content with no chroma subsampling, while chroma information has been reduced by a factor of 2 and 4 for 4:2:2 and 4:2:0 formats, respectively [Poynton 2002]. The resolution of luma and chroma channels are illustrated in Figure 2.4. If there is a correlation between luma and chroma information, any error made during the subsampling may propagate into the luma component and degrade the image. An example of such behavior is explained in [Labs 2016a], which stresses the importance of uncorrelated luma and chroma channels.

We focus briefly on the YCbCr color space, the most used color space from the YUV family in the compression field. Rec. ITU-R BT.601 [ITU 2011b] defined the YCbCr color space in 1982. YCbCr is made of a NonConstant Luminance (NCL) channel Y and blue- and red-difference chroma components Cb and Cr, respectively. Over the years, several variants of YCbCr transcended its limitations. For instance, it is common to apply a nonlinear filter (e.g., gamma correction, introduced in the next section) to remap luminance channel information. This operation helps to reduce the perceptual errors of quantization. When applying the filtering on the luma channel only, we obtain the Y'CbCr color space, when also applying it on chroma channels, we get Y'Cb'Cr'. A linearization may be employed on luma, leading to the Constant Luminance (CL) Y'CbCr (or Yc'CbcCrc) which tackles the problem of interrelation between luminance and chrominance channels [Mahmalat 2016].

2.1.4 Brightness

Condition	Illumination (cd/m^2)
Starlight	10^{-3}
Moonlight	10^{-1}
Indoor lightning	10^{2}
Sunlight	10^{5}
Peak luminance of Cathode Ray Tube (CRT) monitors	10^{2}
Peak luminance of HDR monitors	10^{3}
Range of human vision	Five order of magnitude simultaneously

Table 2.2: Ambient luminance levels for some common lighting environments [Reinhard 2010]

The proportion of rods, when compared to cones, shows the importance of brightness in human vision. Although the HVS covers enormous dynamic ranges, from daylight ($10^{8}cd/m^2$) to night ($10^{-6}cd/m^2$) luminances, it can only perceive a limited range simultaneously. Usually subdivided in two regions, there are cone-mediated photopic (from 10 to $10^{8}cd/m^2$) and rods-mediated scotopic (from 10 to $10^{-6}cd/m^2$) ranges [Ledda 2004]. To cope with large luminance ranges, our visual system has a nonlinear response to incoming signals. This process reduces the perceived dynamic range and can take several minutes [Ledda 2004]. This enables human sight to accommodate to about five orders of magnitude simultaneously [Reinhard 2010]. To indicate to the reader the order of magnitude of luminance in real-life, Table 2.2 presents the light intensity under several conditions and indicates the peak luminance of CRT and HDR monitors.

Let us define several terms, used recurrently through the entire thesis. A *scene* "is either artificial or real environment that may become the topic of an image" [Reinhard 2010]. The *dynamic* of a scene is the difference between the scene's highest and lowest luminances.

HDR aims at the capture, delivery and rendering of detailed visual information, which perceptually matches a depicted scene. It offers to represent visual information within a large variation of luminance, i.e., from deep blacks (0.005 cd/m^2) to very bright (10 000 cd/m^2) values. This is possible through the use of higher bit-depths to convey the information of pixels and more accurate capture.

HDR new characteristics are beyond the capabilities of existing standards and include lower minimum luminance, higher peak luminance, more extensive contrast range, and improved precision, which minimize quantization errors to below human perception levels. For instance, thanks to the HDR ability to carry extensive ranges of luminances, bright specular highlight and details in blacks are less saturated than in SDR. HDR qualities are such that Reinhard et al. claimed in [Reinhard 2010] that the transition from SDR to HDR is a progress in imagery similar to the shift from black-and-white to color TV. It is thus both exciting and necessary to evaluate,

define or use this technologies presently.

2.1.4.1 Acquisition

Several challenges must be overcome for the acquisition of HDR images. The first concern comes from the nature of the scene. A computer-based scene, generated employing realistic image synthesis and global illumination models, can be created for HDR imagery. However, the capture of natural scenes is more problematic due to the limited dynamic range of imaging sensors. Decades of research tackled this issue [Reinhard 2010, Mantiuk 2015, Unger 2016]. The most common practice is to capture multiple-exposure of a scene (with a single- or multi-sensor SDR systems) and recreate the HDR representation through a bracketing operation [Mertens 2009]. But last years have seen the growth of high-end Digital Single-Lens Reflex (DSLR) cameras specialized for HDR acquisition (e.g., Nikon, Canon, Arri).

2.1.4.2 Rendering

HDR systems must ensure the compatibility of transmitted information along the entire HDR pipeline, from the capture of scene's light to its playback, regardless of the display capabilities and compression approach. TFs define the conversion functions required to realize consistency over the entire HDR pipeline.

Every display has its nonlinearity response to scene light. For instance, the nonlinear relationship between the input voltage (V) and light (L) output in CRT is typically approximated by gamma curves (power functions $L = kV^{\gamma}$). The gamma correction is often applied with a γ value of 2.2 [Reinhard 2010]. Additionally, this correction helped to prevent quantization artifacts in digital systems.

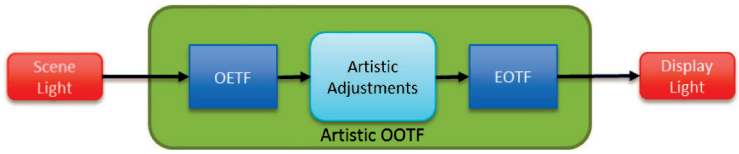
However, gamma curves are unsatisfactory on extended ranges of luminance and color [ITU 2015d]. When applying gamma correction, the distribution of code words becomes uneven, as too many of them are allocated to bright regions, and too few are dedicated to dark areas. This is not a serious issue regarding the limited range of SDR, but it is critical for HDR. Also, the response curves of current and future displays may not exactly follow that of CRT displays. Therefore, it is imperative to have TFs dedicated to large color volume representations.

There are four TFs types, surveyed hereinafter.

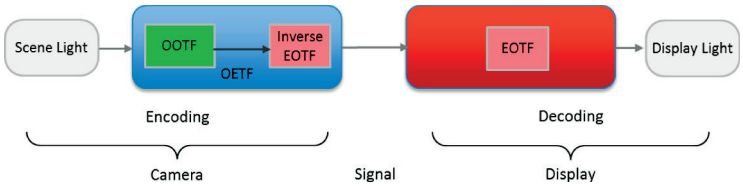
- Opto-Electronic Transfer Functions (OETFs) convert linear scene light into the video signal, typically within a camera.

- Electro-Optical Transfer Functions (EOTFs) convert the video signal into the linear light output of the display.

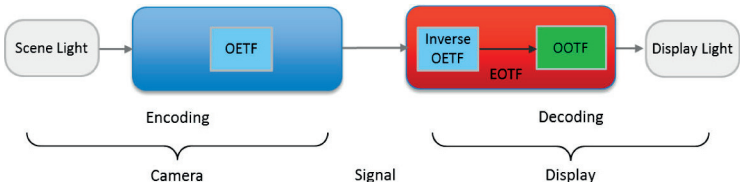
- Opto-Optic Transfer Functions (OOTFs) have the role of applying the 'rendering intent'. OOTFs are typically a succession of OETFs, artistic adjustments, and EOTFs.



(a) Reference Opto-Optic Transfer Function (OOTF)



(b) OOTF in the camera (e.g. PQ)



(c) OOTF in the display (e.g. Hybrid Log-Gamma (HLG))

Figure 2.5: Operating diagrams of OOTF functions [ITU 2015d].

- Electrical-Electrical Transfer Functions (EETFs) are linear light mapping functions. They convert the linear light output of the reference display to match that of the used display.

The Rec. ITU-R BT.2390 [ITU 2015d] describes in details TFs use and functions. Figure 2.5 illustrates the typical combinations of EOTF, OETF, artistic adjustments, and OOTF in HDR pipelines. More precisely, Figure 2.5a illustrates the typical OOTF, while Figures 2.5b and 2.5c show the discrepancies between having the OOTF in the camera or in the display, respectively.

Regarding HDR and WCG imaging, Rec. ITU-R BT.2100 and the Society of Motion Picture & Television Engineers (SMPTE) ST.2084 [SMP 2014a, ITU 2016] have standardized an EOTF, the PQ and an OETF, the HLG, which are briefly introduced below.

PQ

The EOTF PQ [SMP 2014a] is based on the model shown in Figure 2.5b, in which the OOTF is applied in the sensor.

PQ targets the dynamic range of a reference screen covering the entire range from 10 000 cd/m^2 down to less than 0.001 cd/m^2, and is capable of rendering the BT.2020 color gamut. PQ function is inspired from Barten model [Barten 1999] of contrast sensitivity of the human eye, to prevent any visible quantization artifacts using 12-bit coding precision. In Figure 2.6, is presented the Barten ramp [ITU 2015c], logarithmic, gamma, and PQ functions. Below the Barten ramp, step edges are invisible and gradient are smooth. The closest to the Barten ramp a function is, the

more efficient is the representation of luminances in terms of bits. PQ is, therefore, a good compromise, as it encodes highlights and blacks efficiently when compared to logarithmic and gamma curves.

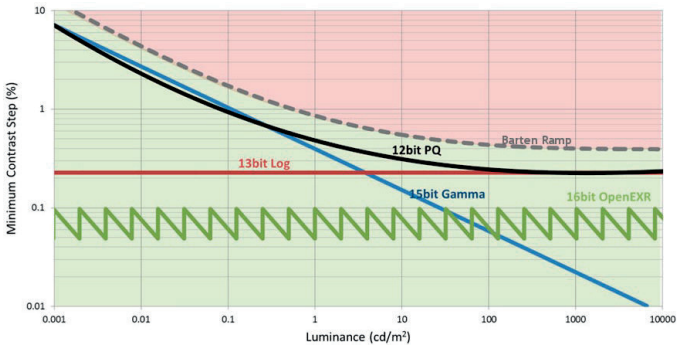


Figure 2.6: Representation of Barten ramp, 15-bit 2.4 gamma, 13-bit logarithmic and 12-bit PQ curves (courtesy of Dolby Laboratories)

The PQ processing follows two steps, which consist in mapping and calibration operations. The mapping is designed to mimic and replicate the human visual contrast sensitivities over a specific range of luminance values. It helps to realize an extensive range of brightness levels for a given bit depth.

The PQ Reference EOTF expression is:

$$\begin{cases} Y = \left(\frac{max[E'^{1/m_2} - c_1, 0]}{c_2 - c_3 E'^{1/m_2}} \right)^{1/m_1}, \\ F_D = EOTF[E'] = 10000Y, \end{cases} \tag{2.1}$$

where
E' denotes a non-linear color value R', G', B' or L', M', S' in PQ space [0,1]
FD is the luminance of a displayed linear component R_D, G_D, B_D or Y_D or I_D, in cd/m^2.
Y denotes the normalized linear color value, in the range [0:1]
m_1, m_2, c_1, c_2 and c_3 are constants defined in [ITU 2016].

PQ has been developed in the laboratories of Dolby and implemented in their end-to-end HDR and WCG framework, namely Dolby Vision ™ [2] [Dolby 2016].

HLG

The OETF HLG implements the model presented in Figure 2.5c, in which the OOTF is on the display side. Authors of [Borer 2016] described and analyzed comprehensively this TF. Also, the ARIB STD-B67 [Borer 2015a] standardized it prior to the ITU. The idea of this function is to combine a logarithmic function, which efficiently represents high luminances, and a gamma curve, which satisfactorily encodes deep darks, as shown in Figure 2.6.

[2] https://www.dolby.com/us/en/brands/dolby-vision.html

The HLG reference OETF expression is:

$$E' = OETF[E] = \begin{cases} \sqrt{E}/2 & \text{if } 0 \leq E \leq 1, \\ a \times ln(E-b) + c & \text{if } 1 < E, \end{cases} \tag{2.2}$$

where
E is the signal for each color component $\{R_S, G_S, B_S\}$ proportional to scene linear light and scaled by camera exposure, normalized to the range [0:12].
E' is the resulting non-linear signal R', G', B' in the range [0:1].
a, b and c are constants defined in [ITU 2016].

The HLG [Borer 2015b] was designed to be as compliant as possible with current systems and largely inspires from previously established television transfer curves. This function has been created during a collaboration between the BBC and Nippon Hōsō Kyōkai (NHK).

Efforts on defining power function for OETF are still going on [Hatchett 2018]. This is explained by the fact that PQ is demanding in terms of computational power when compared to power functions and HLG.

Tone Mapping Operators (TMOs)

Backward compatibility is a strong requirement for broadcasters when shifting from one standard to another. Also, decades of research on HDR were conducted on SDR monitors due to the lack of hardware solutions for HDR display. This justifies efforts made to develop TMOs, which are functions that map HDR luminances to cope with SDR limited dynamic range. These operators are EETFs that map an electrical signal to match the dynamic range of the current (SDR) display.

Many operators have been suggested during the last decades of extensive research on TMOs. Some operators are based on common processing operations of multimedia signals such as scaling, clipping or gamma correction. More advanced content- and HVS-based solutions offer better visual experiences on SDR display. For instance, those operators rely on scene characteristics, such as contrast, sharpness, and color saturation, or HVS properties, which are glare, loss of visual acuity and color perception. TMOs are broadly classified into two categories, local and global operators. Global operators apply the same mapping function to all pixels. Also, they are fast and computationally efficient. Local operators apply unique mapping per pixel, considering a reduced neighborhood around the pixel. Even though more computationally demanding, these operators appear to preserve local contrasts better.

For more information about TMOs, we indicate two thorough reviews for the reader [Devlin 2002, Eilertsen 2017].

2.1.4.3 Compression

HDR and WCG representations convey an extensive amount of information when compared to legacy standards. Therefore, efficient compression algorithms compli-

ant with HDR and WCG are seriously needed. For more clarity in the following descriptions, when referring to HDR, we implicitly take into account WCG.

Compression solutions

Several current compression schemes are available for HDR image and video compression, such as Moving Picture Experts Group (MPEG)-4 [Richardson 2004], Joint Photographic Experts Group (JPEG) 2000 [Skodras 2001], JPEG XR [Dufaux 2009], JPEG XT [Artusi 2016] and HEVC Range Extension (RExt) [ITU 2015f]. These standards form the basis on which are built upon new compression solutions dedicated to HDR.

Several compression solutions approaches have been advanced over the last decades of research. First, single-layer backward-compatible solutions were implemented, using metadata to enhance the SDR stream to HDR. Second, dual-layer backward-compatible approach implements the previous method with an additional layer compensating the error made during the conversion step. Lastly, single-layer non-backward-compatible solutions provide HDR streams and provide metadata to convert streams to SDR. Below, we give a survey of the compression schemes that are particularly relevant to our work.

The review [Myszkowski 2008] presents the first HDR-centered compression solutions. The majority of HDR solutions implement backward-compatible and/or perceptually-based approaches. Among the first HDR compression solutions, [Mantiuk 2006a] proposed a compression scheme which extends MPEG-4 codec. In this scheme, the backward-compatibility is ensured by decomposing HDR stream in two layers. The first one, the Base Layer (BL), conveys an SDR-compliant stream. The second is an Enhancement Layer (EL) which carries decorrelated residuals, allowing to retrieve the HDR representation.

In 2014 was released the HEVC-RExt [Flynn 2016]. This codec supports sample bit depth up to 16 bits and 4:4:4 chroma sampling, which makes it particularly well suited for HDR delivery. Since then, this codec has been the basis for new HDR compression solutions. Among them, Diaz et al. [Diaz 2016] proposed a residual-based (dual-layer) backward-compatible approach which includes a prediction step when compared to [Mantiuk 2006a]. The BL conveys the SDR stream multiplexed with metadata. Those metadata carry prediction information to retrieve an HDR stream. The EL contains the quantized residuals of errors made by the prediction. A really interesting feature of this codec is its ability to drop the EL under dramatic bandwidth constraints. Another HEVC-RExt-based solution, implemented in [Zhang 2016], removes imperceptible information from HDR streams. In that regard, the implementation takes advantage of the lacks of the HVS concerning luminance and contrast as well as spatial and temporal frequency perception. This perceptually-based solution realizes high compression ratio while sustaining a satisfactory perceptual quality.

Recently, momentum was given to single-layer approaches that realize SDR and HDR delivery by means of side dynamic metadata. The first approach is to de-

liver HDR streams multiplexed with metadata, which carry the mapping tuning to reconstruct SDR streams. The second solution favors backward compatibility and sends SDR information in its layer. Hence, metadata convey the mapping settings for HDR retrieval. An example of such strategy was developed by Technicolor and is presented in [Lasserre 2016]. Dolby Vision ™ framework implements both approaches [SMP 2017].

Each approach suits with specific contexts of HDR delivery:

- Single-layer backward-compatible solutions ensure faithful representation of SDR but convey compromised HDR.
- Single-layer backward-compatible approaches provide high-quality HDR stream but deliver SDR with limited accuracy.
- Dual-layer backward-compatible solutions reach a trade-off between SDR and HDR quality

There is a lack of subjective studies which compare the three strategies. Therefore, the selection of a compression strategy for HDR delivery is prevented.

Standardization processes

To facilitate the transition from SDR to HDR, several processes have been standardized these last years. Two types of metadata, static and dynamic, were enabled for HDR delivery. Static metadata were standardized in SMPTE ST.2086 [SMP 2014b]. Also referred to as mastering display metadata, they convey the characteristics of the stream (e.g., color volume, minimum and maximum luminance). Dynamic metadata are content-dependent information. They were standardized in SMPTE ST.2094 [SMP 2017], under the name of Dynamic Metadata for Color Volume Transform (DMCVT), as they provide transformation parameters enabling the adaptation of stream to the target display (concerning dynamic range and color gamut). For instance, one can transmit SDR-compliant streams which can be upgraded in HDR streams thanks to DMCVTs, or the reverse. This explains the development of single-layer codecs, which may retrieve SDR or HDR, depending on the stream representation, by using metadata information.

Simultaneously, the European Telecommunications Standards Institute (ETSI) standardized different compression strategies: the ETSI Group Specification (GS) Compound Content Management (CCM) [ETS 2017] has instituted the specifications of a backward-compatible dual-layer compression solution; the ETSI Technical Specification (TS) 103 433 has introduced a high-performance backward-compatible single-layer system for use in consumer electronics devices.

Standardization processes are so far not favoring any approach for HDR and WCG compression (single/dual-layer, (non)backward-compatibility), but are showing efforts towards a smooth transition from present to future TV.

Regarding delivery, four systems are recognized by the broadcasting community:

Dolby Vision ™, HDR10, BBC/NHK, and Technicolor/Philips [3]. Table 2.3 summarizes the capabilities of each system. It should be noted that HDR10 is part of the Dolby Vision ™ framework. Their differences are introduced in [Chinnock 2016].

Dolby Vision	HDR10
ST-2084 (PQ) EOTF Base layer multiplexed with static metadata Enhanced content layer with dynamic metadata Backward compatible	ST-2084 (PQ) EOTF Non backward compatible

BBC/NHK	Technicolor/Philips
ARIB STD-B67 HLG no metadata mostly backward compatible	ST-2084 (PQ) EOTF Single content with static metadata backward compatibility if additional encoder

Table 2.3: HDR delivery methods

Actually, Dolby Vision ™ is already deployed for cinema [Chinnock 2016]. For instance, it was used for production in the last avenger movie of the Marvel franchise. This indicates that we can expect the arrival of HDR on the market in the near future.

This fact is stressed out by the efforts made by two different and independent organizations (UHD Alliance and Ultra forum). Their mission is to develop and promote UHD, HDR and WCG.

Evaluations of compression solutions

Extensive works have been performed to understand the impact of EOTF on the efficiency of compression and the quality of decoded contents. EOTFs are responsible for the conversion of digital code words in visible light information. A content-based distribution of code words preserves contrast details during the quantization operation, which leads to less visible compression artifacts. Authors of [Baumann 2015] observed that non-linear functions (e.g., PQ) maintain lowest luminance details while more linear functions preserve better bright information. Litwic et al. [Litwic 2016] investigated the impact of the non-linearity of EOTF on compression efficiency in terms of bit rate. They concluded that the impact of the non-linearity of EOTFs is content-dependent, but overall, a bit rate increase for HDR delivery is unnecessary, except for contents which contain a significant proportion of high luminance levels.

A comparison between a single-layer codec HEVC (Main 10 profile [Sullivan 2012]) and a backward-compatible (tone-mapped) SDR solution sending metadata for HDR retrieval has been conducted in [Azimi 2015]. For the same bit rate, single-layer HDR contents reached higher grades than retrieved HDR streams.

[3] https://www.smpte.org/sites/default/files/users/user27446/HDRSMPTEPresentationMarch21%2C2017V2.compressed.pdf

Accordingly, authors strongly recommend the selection of non-backward-compatible solution among single-layer strategies.

In 2015, MPEG issued a Call for Evidence (CfE) for HDR and WCG video coding [Luthra 2015]. They investigated whether currently developed HDR and WCG compression schemes can compete with state-of-the-art standard video coding techniques (HEVC Main 10 Profile). As reported in [Hanhart 2016], results of the CfE demonstrated that several solutions significantly improved the anchor's efficiency. The nature of the proposal was HDR-based, and thus did not rely on the evaluation of the perceptual quality of SDR. At that time, the consideration of backward-compatibility was premature.

In our review of the literature, we have not encountered any work which evaluates both SDR and HDR streams. However, such an evaluation may be of great use for broadcasters. Moreover, HDR delivery is still in its infancy and, to the best of our knowledge, no recommendation regarding bandwidth allocation for HDR delivery has been provided up to now in the literature.

To further explore HDR and WCG, we recommend the reading of three interesting and complete books [Reinhard 2010, Artusi 2011, Mantiuk 2015].

2.1.5 Immersion

Other aspects can influence multimedia experience of human being, to mention but a few, engagement, immersion and SoP. Such high-level aspects of QoE, are further revised in section 2.2.3.1. Their implementation in immersive and emerging technologies is described here.

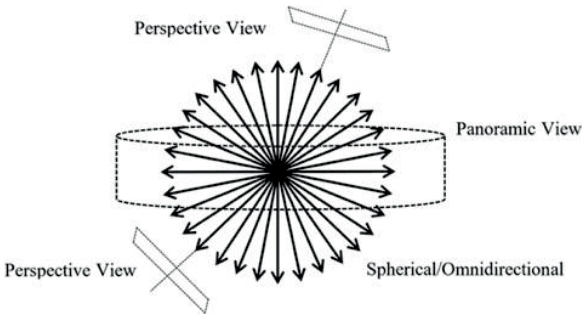


Figure 2.7: Naming convention of omnidirectional contents.

On one hand, immersive technologies got the propensity to widely cover the observer's field of view. This explains the increase in display size and resolution, which was the first step towards this purpose. But 4K covers a visual field still largely inferior to HVS capabilities. This justifies the development of the 360° imaging, also named spherical or omnidirectional format. An omnidirectional content contains the information of a full sphere of views from a center of projection. When the range of pitch values (rotation about the lateral axis) is limited, the content is no longer omnidirectional but panoramic. When considering the information in one direction along with its neighborhood, we obtain an omnidirectional content's perspective

view, also referred to as viewport. The exposed naming convention is illustrated in Figure 2.7. This contents' representation probes the engagement and immersion created when one has access to the information of the entire scene environment.

On the other hand, immersive technologies tend to focus more and more on interactive experiences. This last decade witnessed the growth of HMDs that have been created to enable the viewer to explore 360° content. Additionally, VR, Augmented Reality (AR) and Mixed Reality (MR) are nowadays hot topics with huge momentum, especially for gaming experiences. The viewer encounters a shift from passive to active experiences. Interactions are assumed to induce simultaneously engagement and immersion in users leading to higher SoP and ultimately QoE. Additionally, several improvements are foreseen in future interactive and immersive experiences, such as the integration of haptic or more generally sensory feedbacks. This announces the transition from multimedia to mulsemedia experiences [Ghinea 2014].

Omnidirectional imaging

We focus here on the acquisition, and storage of 360° contents.

Regarding the acquisition, omnidirectional content is captured through various approaches. A large amount of information coming from all directions can be acquired using a set of reflecting surfaces (mirrors) arranged in a specific configuration. However, the combination of back-to-back ultra-wide angle lenses (fish-eye) is more established as ensuring the single viewpoint constraint [Nayar 1997]. However, severe limitations result from systems using a single camera, as, for instance, a high resolution cannot be achieved. Omnidirectional multi-camera systems are tackling the high-resolution constraint. However, such complex system requires a perfect calibration and synchronization of all stacked cameras [Ikeda 2003].

Several cameras Implementing the above-mentioned double fish-eye lenses are already available for consumers on the market, such as the Ricoh theta [4], the Samsung Gear 360 [5], the VIRB® 360 [6], or even smartphones plugin camera like the Insta360TM Nano [7], or the Giroptic iO [8]. Professional cameras, not necessarily available to the public, implement more elegant solutions: even though not covering the full sphere, the array of sensors of the Lytro immerge exploits 6 degrees of freedom through light field acquisition [9]. GoPro Odyssey packs 16 synchronized GoPro sensors [10]. The stack of high-quality Canon M5 and Panasonic GH5 constitute the mooovr [11]. Lastly, the EYE TM Professional VR Camera [12] combines 42 Blackmagic micro cameras along with their Rokinin lenses, distributed along 3 axes. This list is non-exhaustive but is representative of the ecosystem of acquisition systems.

[4] http://www.ricoh-imaging.ch/fr/theta-360.html
[5] http://www.samsung.com/global/galaxy/gear-360/
[6] https://buy.garmin.com/fr-CH/CH/p/562010
[7] https://www.insta360.com/product/insta360-nano/
[8] https://www.giroptic.com/giroptic-io
[9] https://www.lytro.com/immerge
[10] https://shop.gopro.com/EMEA/odyssey/
[11] http://mooovr.com
[12] http://360designs.io/product/eye-vr-camera-full-3-axis-package/

360° spherical information does not match typical images rectangular representations. Therefore, omnidirectional storage implies the use of panoramic projection. Most known mapping functions are the cubic and equirectangular projections. The first one projects the spherical data on a centered cube with unit length sides while the second projects spherical data following constant spacing of latitude. A review of panoramic projections is provided in [Yu 2015b], while researchers are still tackling this issue [Abbas 2017]. The equirectangular projection is especially popular as this projection guarantees the most the backward-probability with current representations.

To deliver 360° contents, usual imaging compressions schemes may be applied to the captured content. For instance, Ricoh Theta cameras store every content after using a legacy JPEG compression. Authors of [Yu 2015b] developed a framework to evaluate the coding efficiency for omnidirectional videos. Their proof of concept relies on the use of the MPEG4 codec. Several works build upon present standards, such as JPEG and HEVC, to come out with new compression schemes [Řeřábek 2016, Zare 2016] or implemented new solutions to match spherical contents' characteristics [Tosic 2009].

VR

The creation process for VR may be similar to 360°, or may be point-cloud-based. Numerous application may be created through VR. The most popular applications being online museum visits [Wojciechowski 2004] and VR gaming [Zyda 2005]. For further information on VR, we suggest the reading of [Sherman 2002].

Visualization for VR and omnidirectional contents

Concerning highly immersive and engaging settings, virtual rooms are recreating a human scale environment. However, such 4- to 6-sided cave automatic virtual environments are not available to most end-users. Computers or any hand-held display (e.g., smartphones or tablets) are far from being ideal visualization displays. Such platforms implement interactions through motion detection. Usually not precise on unrelated to the movement operated in the virtual environment, these interactions prevent a smooth, immersive and engaging exploration of the 360° content when compared to VR headsets.

Most promising displays are HMDs. Developed during the hype of virtual reality gaming and still under development, those systems provide affordable platforms to experience omnidirectional cinema. They implement HVS stereoscopic vision, providing two different views to each eye to provide depth information. The main limitation of these displays is the low resolution offered at such a small distance from the user's eyes.

They can be sorted into two categories, standalone HMDs, which embed the display in the rendering system, and mobile HMDs, using hand-held devices as screen. Google daydream project has developed both types of HMDs [13]. Similarly,

[13] https://vr.google.com/daydream/

Oculus created an embedded variant, the Oculus Rift, and a mobile variant, the Samsung Gear. Sony and HTC developed high-quality standalone HMDs. Numerous cardboards (e.g., Google cardboard, Merge VR) are nowadays available to any consumer. Such affordable and complete systems are far more popular for consumers than standalone HMD [embedded vision alliance 2016].

2.1.6 The use of multimedia technologies in this thesis

In this thesis, we decided to focus on HDR (and WCG) imaging and immersive technologies such as 360° and VR.

HDR is to my mind the most exciting and challenging technology about to arrive on the market. I do not refer here to HDR-able screens that you can presently buy in electronics stores. On the contrary, I mean the ongoing construction of an entire ecosystem for HDR content creation, processing, compression, and delivery. In this prospect, I wanted to dedicate a part of my work in the development of guidelines for HDR usage, and contribute to standardization activities that are in progress.

From another perspective, one can see HDR as an improvement of SDR formats. That is, HDR can be applied to multimedia technologies using SDR representations. With this in mind, I conducted an experiment which is observing if there is an interest to combine HDR with 360° imaging. In this work, we focused on consumer cameras and display systems. The idea was to design experiences as similar as possible to experiences made by any user.

Using 360° opened my mind to immersive technologies. These technologies are the future. For instance, there are numerous fantasies about holography and virtual, augmented and mixed reality. So many application may derive from those technologies than putting efforts in these formats is worth it. For instance, real-time dinner with family while being in different places or leaving fully immersive experiences for various communication tools (e.g., education, entertainment, cultural development).

The most advanced area concerning VR is VR gaming. Main issues regarding subjective assessments, namely content creation and display, have been tackled and are still under development. It is thus possible to run experiments to understands what VR gaming brings to a user. In this light, we have conducted a work which examined differences between different gaming platforms, including VR.

Using new or immersive multimedia in tests introduces complexity: characteristics of the technology influence the experience. When conducting physiological analysis on the perception of a high-level concept of QoE, namely SoP, we decided to consider most current multimedia formats to get free from most technology-related biases. We also ran an experiment exploring reaction to typical scenarios of consumption.

Let us now introduce how evaluations, explorations and more broadly user satisfaction measures are conducted in multimedia appraisals.

2.2 Evolution of users' satisfaction measures

The improvement of technologies relies on measures of consumers' satisfaction. This section presents the evolution of satisfaction measures, from quality to QoE.

2.2.1 Quality

Quality may be regarded under two perspectives. The first one is to come back to the origins of the word, "qualitas", which means a set of inherent characteristics. The standardization committee ITU in Rec. ITU-R E.800 [ITU 2008c] follows this approach when defining quality as "the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs. The characteristics should be observable and/or measurable. When the characteristics are defined, they become parameters and are expressed by metrics". The second is an evaluation of goodness, degree of satisfaction or fulfillment of needs. This approach understands quality from an individual's point of view. In [Möller 2014], authors specified quality as the "judgment of the perceived composition of an entity with respect to its desired composition". It should be detailed that desired composition refers to the set of internal references and expectations against which the perceived composition is being compared.

Quality in multimedia applications refers to a measure. Indeed, the interest in quality is mostly motivated by the need to prove, challenge or benchmark technologies, services or applications. Quality may yield capital pieces of information to fulfill this need. Thus, the second approach is more in line with our research interests. In this thesis, this second definition stands for quality.

Quality appears to be inversely proportional to the amount of distortions in contents, weighted by the importance of impairments. This implies that the amount and the degree of influence of artifacts measure the quality.

A first approach to measure quality hinges on the explicit assessment of multimedia contents, with various levels of impairments, by subjects. Subjective quality measurements remain currently the more accurate way to assess human perception [Keimel 2011]. However, this assessment method is time-, money- and manpower-consuming. Moreover, restrictions induced by the testing environment and test design may influence the results of the experiment and can prevent a comprehensive study of distortions.

A second way to assess quality in multimedia contents is the use of objective metrics. They are mathematical models, which evaluate quality based on measurable parameters. These models are usually psychological (vision-based), or engineering (signal-driven) approaches [Wu 2017]. Moreover, metrics can be classified depending on the input resources needed: full reference (original and distorted content), reduced reference (original content properties and distorted content) and no reference (only distorted content) metrics. In [Chen 2015], an overview and a classification of existing video quality metrics are given.

Regarding the use of objective and subjective assessments, best practices are (1)

to first study an effect through subjective evaluations, and optionally derive mathematic models to encode observed patterns, or (2) construct a theoretical model, based on human vision for example, and verify its results correlation with ground-truth scores, resulting from subjective evaluations.

Objective metrics are highly important for conducting a fast assessment of a system as well as providing a way to compare various services. Subjective evaluations are useful for performing proofs of concept, observing and determining individual's response to the multimedia presentation, and understand further cognitive and perceptual processes in human perception.

Quality is nowadays still used for benchmarks of multimedia services or applications [Hanhart 2016]. However, numerous restrictions in evaluating only quality have been raised. Additional characteristics, unrelated to the content quality only, may influence the perception of an audiovisual experience. Those characteristics are network-related, as the delivery of multimedia content is rarely lossless. This explains the shift envisioned from quality to QoS in multimedia evaluations.

2.2.2 QoS

QoS evaluates objectively the amount of network distortion as well as the properties of the provided content. QoS assessment ties together the measure of network distortions and application properties. Application properties, listed non exhaustively, are resolution (SD, HD, UHD), representation (3D, HDR), frame rate, bit rate, color gamut, color space, video and audio quality, temporal and spatial features. The device presenting the service (TV, desktop computer, laptop, tablet, smartphone, etc.) is also included in the QoS application layer. QoS network layer represents all the features, distortions, and artifacts impacting the provided content. The main distortion created by networks are packet loss, jitter and delay. Limited bandwidth, as well as encoder distortions (blocking effect, blurring, edginess), also reduce the quality of the content. Overall, QoS deals mostly with physical, measurable performance factors of networks and delivery platforms in general.

Reviews [Chen 2015, Alreshoodi 2013, Takahashi 2008] present the categories of objective quality assessment models of services:

- Parametric packet-layer model: using packet headers information to predict the service quality.
- Parametric planning model: using quality designed parameters to predict the service quality.
- Media layer model: analyzing the media signal.
- Bit-stream model: using packet headers and payload information for quality assessment (model between packet layer and media layer model)
- Hybrid model: combining any previous models.

To summarize, QoS is the generalization of quality for an entire telecommunications service. According to Rec. ITU-R E.800 [ITU 2008c], QoS is the "totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service".

After decades of research on QoS, researchers started to wonder whether using QoS to represent users' satisfaction was suitable. The definition of QoS emphasizes that it measures the ability to satisfy the end-user and not the level of satisfaction reached by the latter. Hence, QoS is not powerful enough to fully express all multimedia contents and communication service aspects. This explains the transition from QoS to QoE to assess multimedia services and technologies.

2.2.3 QoE

Multimedia is about sharing experiences with others. Tremendous progresses of multimedia entertainments aimed to realize rich experiences. QoE measures the degree of richness of the experience. The concept of QoE has emerged in multimedia fields mainly with the basic motivation to cover a wider scope and to be more user-centric than QoS, which is focusing on providers perspective. Users' previous experiences, expectations, current state of mind and socio-professional category affects experiences. Including those aspects in QoE better quantifies the user's overall satisfaction with a service [Brunnström 2013]

ITU-T Rec. P.10/G.100 [ITU 2007b] defines QoE as "the overall acceptability of an application or service, as perceived subjectively by the end-user." with those additional notes: "(1) QoE includes the complete end-to-end system effects (client, terminal, network, services infrastructure, etc.). (2) Overall acceptability may be influenced by user expectations and context." However, Möller argued in [Möller 2017] that the term acceptability may not be appropriate for a QoE definition. The author defined the QoE as follows: "QoE is the degree of delight of the user of a service. In the context of communication services, it is influenced by content, network, device, application, user expectations and goals, and context of use." The most accepted definition of QoE is provided by the COST Action IC1003 (Qualinet) white paper [Brunnström 2013]: "QoE is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state". This definition generalizes the previous one. In 2016, the ITU updated the QoE definition to the Qualinet definition in [Amendment 2016] to be in line with most works conducted on QoE.

Another concept which already considers the user is the User Experience (UX). "UX deals with studying, designing and evaluating the experiences that people have through the use of (or encounter with) a system" [Brunnström 2013]. There are numerous similarities between UX and QoE. However, QoE is a broader concept in that it also takes into consideration the content and not just the system evaluated.

Based on the definition in [Amendment 2016, Brunnström 2013], QoE is a multi-dimensional and multi-modal notion which combines three distinct aspects: human,

system and context factors. Table 2.4 presents examples of influencing factors that QoE comprehends.

Factors influencing QoE	Examples
Human factors	Demographic and socio-economic background, physical and mental constitution, user's emotional state.
System factors	Content related (e.g., content type and temporal characteristics), media related (e.g., resolution, compression scheme, sampling rate), network related (e.g., bandwidth, delay, jitter, packet loss, error rate) and device related (e.g., display size, usability, security).
Context factors	Temporal context (e.g., time of the day, duration), economic context (cost), task context (frequency of use) and social context

Table 2.4: Influence factors of QoE [Hewage 2013]

Csikszentmihalyi [Csikszentmihalyi 1999] emphasized that alike quality, depending on the application, QoE may measure how an experience is faithful, truthful, immersive, contextual, engaging, effective, useful, interactive or intuitive. He also identified main challenges to face for the investigation and development of QoE. In this thesis, we will focus on several of them, namely content- and context-dependent quality assessment methods, multi-modal evaluations, presence, immersion and quality appraisal, and virtual reality immersive experiences.

Keimel et al. [Keimel 2011] stressed out that no single measurable quantity can fully represent QoE. Additionally, many aspects of HVS and perceptual processes are not sufficiently understood. This leaves limited chances to model QoE comprehensively presently. Cognitive and underlying perceptual processes need to be further explored, as attributes of immersive and rich technologies should be investigated. This indicates the need for conducting subjective evaluations to discover more about the aspects described above.

For more details about the many dimensions and different applications of QoE, the reader is recommended to refer to the excellent book on advanced concepts, applications and methods of QoE [Möller 2014] as well as the well documented and exhaustive survey [Chen 2015]. Also, the comprehensive review the QoE for HTTP adaptive streaming [Seufert 2015] bring important aspect to our attention. Numerous substantial outcomes may be retained from this study (e.g., the method to identify QoE factors, taking into account technology-, server- and client-based attributes). Also, the Rec. ITU-T G.1031 [ITU 2014c] provides various influence of web-browsing for the context and system aspects of QoE.

2.2.3.1 SoP

Today's TV, especially UHD may be regarded as a spatially limited Virtual Environment (VE). The SoP is one of the most significant quality measures for VEs. As such, SoP is seen as a notable high-level aspect of QoE. The first multimedia service considering SoP as a measure of users' satisfaction is telepresence in Rec. ITU-T F.734 [ITU 2014a]. But more interestingly, Rec. ITU-R BT. 2246 [ITU 2015c] related to UHD TV, integrates a study on the "sense of being there" and "sense of realness" for London Olympics super high-quality vision public viewing operation.

SoP is mainly known as a desired quality metric for immersive environments [Slater 1997]. It has been studied in various multimedia applications, which explains the number of highly used presence questionnaires. However, scientists do not always agree on the SoP definition. Immersion and engagement are also discussed, due to their highly related to SoP concepts. Table 2.5 presents the definitions of SoP, immersion and engagement (also referred to as involvement) in various state-of-the-art questionnaires.

Discrepancies in definitions are such that presence and immersion do not represent the same constructs depending on the used questionnaire [Youngblut 2003]. In [Slater 1997], presence and immersion are technology-based while [Witmer 1998] focus on a psychological perspective of presence.

We find easier to understand the concepts of immersion as being technology-related information, while presence is a psychological response to immersion and engagement. We decided to select the definitions of Igroup Presence Questionnaire (IPQ) for immersion and the one from Presence Questionnaire (PrQ) for presence. Descriptions of engagement are similar enough, but our works mainly set stimuli duration, so we have selected the definition from PrQ.

One may argue that we could have selected all definitions from this questionnaire. This fact is accurate, but we had to consider that if presenting subjects with presence and immersion questions, the discrimination between both concepts would be readily understood as a difference of perspective (technology- or psychological-based). This choice is supported by [Kalawsky 2000].

Regarding the evaluation of these three constructions, considered questionnaires in Table 2.5, contain numerous items which may cause boredom and lack of attention in subjects during subjective evaluations. To avoid such biases, Bouchard et al. [Bouchard 2004] verified that single-item measure of SoP is well understood, reliable and valid. They though pledge for additional questions to investigate presence construction further. [Baños 2004] related SoP with immersion and content-(dis)like. Authors stressed the importance of considering SoP as a multi-component construct due to the lack of agreement on its definition. They identified two categories of variables that impact the SoP, namely the characteristics of media and users.

We recommend the reading of the two works [Lessiter 2001, Witmer 1998] to deepen knowledge of the SoP concept. Various studies in television services, mainly virtual environments, which evaluate SoP are presented in [Freeman 1999].

QoE aspect	IPQ [Schubert 2001]	PrQ[Witmer 1998],	Player Experience of Need Satisfaction (PENS) [Rigby 2007]	Immersive Experience Questionnaire (IEQ) [Jennett 2010]
Presence	Presence is a state of consciousness, the (psychological) sense of being in the virtual environment, seen as a function of immersion.	Subjective experience of “leaving" the intrinsic world and “being present" in a virtual environment. Both involvement and immersion are necessary for experiencing presence.	Being truly "in the experience", not only with regards to engagement, but also in terms of emotional and imaginative response. Physical, emotional and narrative presence type are considered.	Psychological sense of being in a virtual environment.
Immersion	The extent to which the computer displays are capable of delivering an illusion of reality to the senses of a human.	Perceptual feeling of being enveloped by, included in and interacting with an environment that provides a continuous stream of stimuli and experiences. Immersion is a variable of the technology.		Psychological experience of engaging with an audiovisual experience.
Engagement		Psychological state experienced as a consequence of focusing one’s energy and attention on a coherent set of stimuli		An engaged user is one that has invested time, effort and attention in the experience.

Table 2.5: Presence, immersion and engagement definitions

2.3 Subjective evaluations

The interesting idea of *Wisdom of Crowds* is introduced in Aristotle's Politics: a group of individuals, when compared to an expert, will make more accurate decisions when it comes to problem solving, decision making, innovating and predicting. Indeed, by collecting a crowd's individual ratings, one should obtain a probability distribution of scores centered near the real value of the quantity to be estimated.

This philosophical idea is a mathematical and probabilistic phenomenon appearing under four constraints: 1) Diversity: the crowd, also called population sample, is composed of individuals coming from a large variety of sociocultural contexts (age, gender, education, ethnicities or social environment). This constrains the distribution of scores to be representative of the global population. 2) Independence: each person's opinion is expressed freely, under no influence whatsoever. 3) Decentralization: any interaction between individuals is prevented to preclude a joint judgment, lessening the effect of the crowd 4) Aggregation: aggregate all the ratings in a final score. This mathematical model and its constraints define the process implemented by subjective evaluations.

In multimedia subjective evaluations, additional biases must be considered when designing a test. For instance, providing the same stimulation to all individuals of the tested population and preventing test conditions (e.g., visualization material and environment) to bias the experiment results are compulsory.

Standardization committees have gathered guidelines to design any subjective evaluation properly in recommendation documents. ITU provides recommendations for subjective assessment regarding television pictures quality in Rec. ITU-R BT.500 [ITU 2012c], quality for multimedia applications in Rec. ITU-T P.910 [ITU 2008d], video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment in Rec. ITU-T P.913 [ITU 2014d], viewing environment for HD TV program material in Rec. ITU-R BT.2035 [ITU 2013a], parameters values for HD TV standards in Rec. ITU-R BT.709 [ITU 2015e] or viewing conditions for SD and HD TV pictures on flat panel displays in Rec. ITU-R BT.2022 [ITU 2012b], image parameter values for HDR TV for use in production and international program exchange ITU-R BT.2100 [ITU 2016], and many other recommendations. It should be mentioned that ITU-T and ITU-R are relevant for standardization and radiocommunication respectively. Series P stand for "Terminals and subjective and objective assessment methods", BT for "Broadcasting services" and BS for "Broadcasting sound".

This section provides aspects of these recommendations that are most relevant to our context of study. First, viewing conditions are listed, followed by brief descriptions of subjective evaluation methodologies and rating scales. The essential test material selection process is then exposed. Later on, requirements regarding subjects and instructions to observers are presented. Common practices to design a test and to analyze collected subjective scores are provided afterwards. The next paragraphs present the use of physiological signals as a more implicit assessment and considerations about subjective test validity. These are not necessarily standardized

(yet) although they have fundamental importance.

Notable highlights of settings used for evaluations performed during this thesis are also provided. It is highly suggested for someone interested in designing a subjective test to read ITU recommendations. It should be noted that various quantity aspects, such as quality or level of impairments, can be evaluated during multimedia subjective tests. This quantity is defined on a case-by-case basis.

2.3.0.1 Viewing conditions

An overall trend when going to the cinema is to select a place around the very middle of the room: the perfect place is not too far from the screen, not too close either; placed to enjoy the quality of the surround sound system; not affected by the "way out" lights which are too bright compared to the room darkness.

This trend emphasizes the importance of viewing conditions such as distance to the screen, absolute and relative brightness of the environment, the display and the sound importance, especially noise disturbances, during an audiovisual experience. In the context of subjective experiments, most aspects of the environment can have an impact on subjects' ratings.

Two different viewing environments have been defined for evaluation protocols in Rec. ITU-T P.913, Rec. ITU-R BT.500 and BT.2022 [ITU 2014d, ITU 2012c, ITU 2012b]. On one hand, laboratory viewing environments are highly controlled environments providing critical conditions to check systems. On the other hand, home viewing environments are under less controlled conditions and are more compliant with TV-consumers' home settings.

Despite being a non-realistic home environment, a laboratory viewing environment is free from exterior biases. This explains the use of such a controlled environment in all conducted evaluations of this thesis. Its specifications are presented below.

A laboratory test room provides a comfortable non-distracting environment, free from noise pollution. Room illumination should be low (10 lux) with a background chromaticity corresponding to average daylight (D65). The peak luminance of any light source can range from 75 to 250 cd/m^2 and the ratio between the background luminance behind the monitor to the peak luminance has to be set at about 0.15.

Concerning viewing distance, it is essential to differentiate flat panels and mobile devices (e.g., smartphones or tablet). Flat panels have a fixed position. Thus, visualization distance can be set as Designed Viewing Distance (DVD), also called optimal viewing distance which is determined such that the visual angle perceived between two pixels is one arc of minute angle at the viewer's eyes [ITU 2010, ITU 2015c, ITU 2013a, ITU 2015b, ITU 2012b]. Table 2.6 indicates the optimal viewing angle and optimal viewing distance for various display resolutions and reports in which document the recommendation can be found. Regarding mobile displays, the viewing distance is adjusted to the subject's preference, referred to as Preferred Viewing Distance (PVD).

When it comes to environmental settings, tuning of display parameters is of

Reference	Resolution	Optimal horizontal viewing angle	Optimal viewing distance
Rec. ITU-R BT.1543 and Rec. ITU-R BT.1847	1280×720	21°	$4.8H$
Rec. ITU-R BT.709	1920×1080	31°	$3.2H$
Rec. ITU-R BT.1769	3840×2160	58°	$1.6H$

Table 2.6: Optimal horizontal viewing angle and optimal viewing distance in active display area height (H) as a function of the content resolution

key importance. Monitor resolution, contrast, brightness and characteristics have to be set on a case-by-case basis. For instance, there are discrepancies among recommendations about display peak luminance: Rec. ITU-R BT.1886 and BT.2035 [ITU 2007a, ITU 2013a], focusing on production and broadcasting constraints, advocate for a display peak luminance of 100 cd/m^2 whereas Rec. ITU-T P.913 [ITU 2014d], concentrated on the audiovisual quality of all multimedia services, argues for 200 cd/m^2, and Rec. ITU-T P.910 [ITU 2008d] indicates a peak luminance ranging from 100 to 200 cd/m^2. It must be noted that the previous recommendations have not examined new representations such as WCG or HDR. ITU-R BT.2390 [ITU 2015d] defines HDR peak luminance range between 1 000 and 10 000 cd/m^2. Specific calibrations and verifications must be applied to the display. For instance, display device linearity can be confirmed using conventional 16:9 or 4:3 aspect ratio [ITU 2011b, ITU 2005] digital TV reference test patterns [ITU 1994].

Regarding the environments where our evaluations took place, Multimedia Signal Processing Group (MMSPG) test laboratories fulfill ITU laboratory viewing environment settings. The laboratory rooms' walls are homogeneously painted in mid-grey. The temperature and ambient lighting are adjustable. Calibration of display devices has been applied according to Rec. ITU BT.2022 [ITU 2012b] or will be mentioned if done otherwise.

2.3.1 Subjective evaluation methodologies

Subjective test methodology varies depending on the purpose of the evaluation (e.g., qualitative or quantitative study, comparison between systems or observation of a system behavior). In the following, the main subjective evaluation methodologies for (audio)visual contents are described. Each methodology targets a specific evaluation purpose, therefore several aspects are considered:

1. A first aspect is **test material characteristics** (still pictures or videos). The time dimension of videos impacts test methodology selection in several ways, such as content selection (variety of temporal complexities) and considerations of stimuli repetition and duration.

2. A second aspect consists of choosing between two alternatives for the **appraisal procedure**: collecting scores all along the stimulus (continuous assessment) or gathering an overall rating for each stimulus (overall evaluation).

3. **Stimuli presentation** is another aspect. A single-stimulus evaluation is presenting one stimulus at a time and results in the appraisal of only this content's characteristics. A double stimulus evaluation presents two stimuli simultaneously, enabling stimulus comparison. Two kinds of scoring may appear, either one grade is given for the first content when compared to the second, or two independent grades are provided, one for each stimulus.

4. Regardless of stimuli presentation, another aspect is **whether or not reference content is included** in the evaluation. Reference content is usually the source signal of content (raw, uncompressed) and its main characteristic is the absence of defects. In specific cases, selected anchors are set as reference systems, defining the state-of-the-art performance basis. Reference presentation enables the stabilization of subjects' scores, especially when impairments to be measured are hardly noticeable. However, including reference should be avoided when its quality is from fair to bad.

5. References in a test can be **hidden or explicit**. Hidden-reference methodologies provide the experimenter the possibility to detect outliers. Explicit-reference provides a benchmark.

6. Finally, methodologies are usually suggesting a **rating scale** to measure either quality or level of impairments.

Methodologies combine these six aspects defined above. Here are their specifications, that can be found in Rec. ITU-R BT.500, ITU-T P.910 and P.913 [ITU 2012c, ITU 2008d, ITU 2014d].

2.3.1.1 Single Stimulus (SS) and Absolute Category Rating (ACR)

As its name mentions, in SS methodology, a single still picture or video is presented to the observer at a time. The viewer is required to provide his assessment of the entire presentation (overall evaluation). If reference contents are to be included, a hidden-reference fashion should be respected. The rating scale used during SS experiments is an adjectival categorical judgment method. In terms of naming convention, the ACR method refers to a SS evaluation for quality appraisal.

The strength of this methodology is its ability to evaluate a high number of stimuli in a short time period. Nevertheless, SS ratings can act as make-believe, introducing a confusion between the evaluated effect and the content influence upon subjects. Using a hidden-reference methodology removes some content influence from the SS ratings.

2.3.1.2 Comparison Category Rating (CCR), Double Stimulus Comparison Scale (DSCS) and Pair Comparison (PC)

CCR, DSCS and PC represent the same evaluation methodology, presenting two stimuli sequentially. Subjects are asked to rate a first stimulus quality compared to a second stimulus. The presentation order between reference and evaluation content must be balanced (50% of pair comparisons start with the reference stimulus).

The CCR method produces fewer ratings than the ACR in the same time period. Subjects' like or dislike in contents minimally influence CCR ratings. The cognitive load on subjects is high when using CCR. Indeed, evaluating sequential comparisons requires an effort of memory. This leads to an increase of subjects' exhaustion. Ratings can, thus, be unreliable.

2.3.1.3 Double Stimulus Impairment Scale (DSIS) and Degradation Category Rating (DCR)

DSIS and DCR are the same methodologies. They are similar to the CCR methodology. Yet, reference is always presented as the first stimulus. Also, the used rating scale measures the level of impairment. Last but not least, this methodology includes a variant (namely variant II or Side-by-Side (SbS)), enabling two stimuli presented in a side-by-side fashion. DSIS variant II can be implemented by rendering two stimuli on a single display or two different screens. Perfect synchronization and calibration are mandatory in the latter context. To cope better with hardly noticeable impairments, video stimuli repetition can be operated.

Just as CCR, subjects' opinions in content minimally influence DCR ratings. When using variant I or variant II with repetitions, it produces fewer rating than SS in a same period of time. DSIS variant II reduces the cognitive load demanded from subjects. DCR methodology is particularly suitable for color detection or minor impairments, hardly detectable by SS method. DCR variant I ratings may contain a slight bias due to fixed order in stimuli presentation as observers know that the first stimulus is the reference.

2.3.1.4 Continuous evaluations for videos

Simultaneous Double Stimulus for Continuous Evaluation (SDSCE) enables a real-time measure of the levels of discrepancies between two impaired sequences. Double-Stimulus Continuous Quality-Scale Method (DSCQS) measures system quality relative to a reference, and is especially recommended for stereoscopic image coding quality assessment. Single Stimulus Continuous Quality Evaluation (SSCQE) measures service efficiency and provides insights to analyze its behavior. Continuous methodologies are not presented in detail as they were not used in performed works.

2.3.1.5 Subjective Assessment of Multimedia Video Quality (SAMVIQ)

In this methodology, contents are prepared and gathered as sets of variants of the content. Each set is evaluated entirely before assessing the next one. The order

in which one set is presented is random. Within each set, the presentation order and time of visualization of the variants of the content are left to subjects' decision. Subjects have the possibility to view the same variant several times.

Stimuli are rated using a continuous quality scale such as DSCQS method. This methodology can be designed as a non-reference, explicit- or hidden-reference methodology. It can be regarded as a single stimulus continuous method with random access.

2.3.2 Rating scale

The previous section reported that various rating scales can be used when evaluating a multimedia experience. There are two types of rating scales: adjectival categorical judgments and non-adjectival judgments. The second type includes continuous scales (usually appraised using a slider) or numerical scales.

The proper scale selection is part of evaluation methodology design. However, some rating scales are recommended when using specific subjective methodologies. Table 2.7 reports various rating scales for categorical judgments, indicating the assessment type (quality or impairment level) and which subjective methodologies usually apply them. Their description can be found in Rec. ITU-T P.800, P.910 and ITU-R BT.500 [ITU 1996, ITU 2008d, ITU 2012c].

The context of the evaluation as well as the analysis to be performed have an impact on the number of points of the rating scale. The more the points, the more refined the assessment will be, but the more complex will be the statistical analysis. Using a scale with large number of levels is allowed, yet it should be prevented. In practice, mostly 3, 5, 7, 9 and 11 are used. It is also usual to see an assessment ranging from 0 to 100. It should be noted that scales are assumed to be linear. Also, typically, the highest grade of a scale indicates the highest quality or the lowest level of impairment.

When investigating subject agreement level with a statement, Likert scales are widely used. Table 2.7 shows one example of such a scale.

Assessment	Adjectival categorical judgments			Likert scale
Scale	5 Excellent 4 Good 3 Fair 2 Poor 1 bad	5 Imperceptible 4 Perceptible, but not annoying 3 Slightly annoying 2 Annoying 1 Very annoying	3 Much better 2 Better 1 Slightly better 0 Same -1 Slightly worse -2 Worse -3 Much worse	5 Strongly agree 4 Agree 3 Neither agree nor disagree 2 Disagree 1 Strongly disagree
Type Methodology	Quality ACR, SAMIQ	Impairment SS, DCR	Quality and Impairment CCR	Agreement level

Table 2.7: Adjectival categorical judgment and Likert rating scales

2.3.3 Test material selection

The importance of sample population variety has been explained previously. The same approach must be followed for experiment contents: the more variety the contents have, the more accurate the conclusions drawn from the evaluation will be.

In order to select a wide range of different contents, content characteristics, such as frequency, spatial and temporal complexity and dynamic range are taken into consideration. They often must be representative of both natural and computer-generated contents.

Spatial and temporal complexities are described in Rec. ITU-R BT.1788 [ITU 2011a].

Spatial perceptual Information (SI) is the amount of spatial information in the video scene that the viewer perceives. SI is computed as the maximum value of the standard deviations of all the frame gradients. The gradient operator is applied through horizontal and vertical filtering with a Sobel operator.

$$SI = max_{time}\{std_{space}[Sobel(F_n)]\} \tag{2.3}$$

where F_n is the n^{th} frame.

Temporal perceptual Information (TI) evaluates the perceptual temporal complexity (motion) of a video scene. TI measures the maximum standard deviation of pixel-wise difference between each frame of the sequence. The change in pixel values over two successive frames expresses the motion.

$$TI = max_{time}\{std_{space}[F_n(i,j) - F_{n-1}(i,j)]\} \tag{2.4}$$

where $F_n(i,j)$ is the pixel value at coordinates (i,j) of the n^{th} frame.

When using audio cues, a sound energy analysis can be done. The sum of squared signal values measures the energy quantity conveyed by the audio channel.

Regarding HDR contents, three main content characteristics are evaluated: Dynamic Range (DR), luminance distribution histograms and key value.

DR is the highest luminance L_{max} over lowest luminance L_{min} ratio, clipped at 0.01 and 99.9^{th} percentile to make the metric robust against outliers. This is the most popular unit for expressing DR. It is measured in f-stops. In the case of videos, L_{max} and L_{min} are taken from all frames. In images, DR measure is called static luminance range.

$$DR = \log_2(\frac{L_{max}}{L_{min}}) \tag{2.5}$$

In case of dynamic range representation, one can compute the base 10 logarithmic of DR as follows:

$$DR_{10} = \log_{10}(\frac{L_{max}}{L_{min}}) \tag{2.6}$$

A DR histogram is a visual representation of the luminance distribution.

The key value is a popular metric in HDR imaging to indicate the average brightness of a scene [Reinhard 2010]. It is expressed as depicted in the equation below:

$$KeyValue = \frac{\log L_{avg} - \log L_{min}}{\log L_{max} - \log L_{min}} \tag{2.7}$$

with $logL_{avg}$ being the log geometric image luminance mean.

When dealing with videos, their duration has to be defined. Rec. ITU-T P.913, ITU-R BT.1788 and BT.2021 [ITU 2014d, ITU 2011a, ITU 2015b] advise that video stimuli should last between 5 to 20 seconds, and encourages to use 10-second stimuli.

Considering source stimuli number, standard practice is to include four to six different scenes. This is to ensure the selected contents are diverse enough for the results to be generalizable.

After verifying that the content characteristics employ adequate variety, expert vision tests can be performed to select source contents. The selection is performed considering the observable impairments introduced by the systems under test. Depending on the evaluation purpose, test material will change. For instance, for an overall system performance evaluation, the content selected has to be representative of all classes of contents. On the other hand, for identification of the system limitations, content with more specific characteristics will be selected.

2.3.4 Subjects

The reliability and sensitivity of a test procedure, as well as the pursued effect size (quantitative measure of a phenomenon) determine the number of assessors to include in an experiment.

Rec. ITU-R BT.500 [ITU 2012c] states that the number of participants should range from 4 to 40 subjects. Based on statistical theory determining a minimum sample size to be able to observe a statistically significant effect size, at least 15 observers are included in experiments. Fewer participants are necessary in a controlled laboratory environment than in home viewing environments.

Observers are categorized as naive viewers or expert viewers. Naive or non-expert viewers have no expertise in impairments or degradations introduced by the system under test. A population with non-expert viewers is necessary to evaluate the comprehensive behavior of end users.

Before the experiment, subjects should be screened for correct visual acuity (no errors on 20/20 Snellen chart [Snellen 1868]) and color vision (Ishihara charts [Ishihara 1960]). The experimenter should reject any potential participant if these previous requirements are not fulfilled or if anyone is prone to epileptic seizures.

Regarding population characteristics, one should target a well-balanced age and gender distribution. If dealing with more than one population, sampling can be probabilistic (random distribution of subjects in various samples) or non-probabilistic (distribution based on a judgment, such as experience with virtual reality).

2.3.5 Instructions for subjects

Before the experiment, observers should be provided with detailed oral and written description of the purpose of the study, the assessment method, impairment or quality levels likely to occur, grading scale and test process (sessions, breaks, stimuli duration and evaluation).

After the experiment and screening descriptions, subjects take part in a training session mimicking a real test session. While training stimuli are visualized, the experimenter introduces the quality levels explicitly. Usually, three quality or impairment levels are presented in the training: the example with the highest level is displayed first, followed by the example with the lowest level. An example with a moderate level is displayed last.

When the purpose of evaluation is to identify system limitations, the experimenter may put a focus on parts of content that are likely to contain impairment. Test material may be used as training material only within this context. Otherwise, different contents must be selected for the training session.

In order to reduce experiment bias, participants should be reminded that there is no right or wrong answer, what matters is their opinion. At any time before test session, subjects can ask questions that must be answered.

The purpose of the training process is to let subjects get familiar with the assessment procedure, to stabilize their votes over the full range of impairments and to encourage them to ask questions if anything is unclear. It is the time to verify the experiment settings and comfort of subjects.

2.3.6 Test design

Implementations of experiments are carried out following the steps described below:

1. Informed consent:
   Subjects should be informed of their rights and about the experiment when presented with experiment advertisement, subjects consent and information forms. Written documents provided to participants followed the Rec. ITU-T P.913 [ITU 2014d].

   The thesis works were implemented under HREC 006-2015 Ecole Polytechnique Fédérale de Lausanne (EPFL) Human Research Ethics Committee application. This guarantees the rights and welfare of participants in our research involving humans. Subjects should take part in experiments on a voluntary basis. Collected data and subjects' identity are kept confidential. Subjects must be protected from any physical or emotional harm.

2. Pre-screening of subjects:
   As already mentioned, subjects should be screened for correct visual acuity and color vision. Additionally, any person prone to epileptic seizures is not eligible for the tests. No test can be run on under-age subjects without a responsible

person's agreement. The participation of underage subjects should not be favored in the first place.

3. Instructions and training:
   Section 2.3.5 introduced the purpose of training session and its implementation.

4. Test session(s):
   Also named voting sessions, tests sessions are run successively, interleaved with break periods. Their design is described below.

5. Questionnaire and/or interview (optional):
   To supplement information gathered during test sessions, additional information can be collected before and after test sessions. They usually capture sociocultural context of subjects (e.g., age, gender, television watching habits) and their overall feedback.

Depending on the assessment method, number of stimuli (deriving from the number of systems evaluated and test material) and stimuli duration, a test session or an experiment can be too long. To prevent tiredness and lack of attention, the recommended experiment length is limited to one hour and a half (see Rec. ITU-R BT.500 and BT.2021 [ITU 2012c, ITU 2015b]). When test duration is long (more than an hour in overall), frequent breaks and adequate compensation should be offered to participants.

Designing a test requires to carefully select a test methodology that minimizes the number of times a given source stimulus is shown. Traditionally, test sessions last up to 20 minutes, and no test session can last for more than 45 minutes.

Test (and training) sessions repeat this following pattern: stimuli presentation, scoring period. Typically, a mid-gray screen is displayed between two patterns, in a quiet environment. Scoring period duration can be set or left undefined.

Stimuli should be shown in a pseudo-random order. The presentation order must also ensure that two stimuli derived from the same source content will not be presented successively. Similarly, presentation order has to prevent the same impairment from occurring consecutively. To minimize any ordering effects on results, implementing a different stimuli presentation order per subject is better, even though not always practical.

These practices reduce unrelated influences of content and impairments order on grades. They also limit the boredom for subjects and learning effect. Consequently, ratings are more reliable.

Test stimuli partitioning in test sessions should be randomized and balanced in terms of source content, taking into account any effect on subjective ratings such as tiredness and adaptation. The order of sessions themselves should be randomly presented. Some stimuli repetitions in different test sessions can be included to check the coherence of ratings. Also, dummy stimuli can be added at the beginning of each test session to stabilize subjects' ratings. Dummy evaluations are not to be taken into account during score processing.

Test sessions can also be designed in such a way that the experiment implements a within-subjects or between subjects assessment method. In a within-subjects design, subjects are exposed to all independent variables. Thus, each test session contains every impairment variant or influence factor introduced by the system under evaluation. In such a design, a small number of people is needed to run the experiments. The population assesses all factors and results are more sensitive to the changes in these factors. However, subjects will also be more likely to present learning effects and to feel more tired as the experiment will last long. A between-subjects design implements test sessions containing one impairment or influence factor, such that each participant is only exposed to a unique independent variable, creating mutually exclusive participant groups. This design prevents any learning effect, avoids tiredness or boredom in subjects and results in short experiments. However, more participants will be required and randomization of factors can be complex to implement depending on the study context.

2.3.7 Data processing

The subjective evaluation of systems is assessed by integer distributions. Those distributions are not usable if they are not adequately summarized and represented. To draw reliable and accurate conclusions, rating distributions are statistically analyzed.

Before processing any scores, subjects's reliability must be tested by performing an outliers detection. Scores should be tested if they follow a normal distribution (see Rec. ITU-T P.1401 [ITU 2012d]), mainly due to the Analysis of Variance (ANOVA) analysis requirements. Observer screening is standardized in Rec. ITU-R BT.500, Section 2.3 in annexe 2 [ITU 2012c] and consists in using β_2 test, based on kurtosis coefficients. Outliers detection process rejects subjects who have greater than 20% of their scores outside of the 25^{th} and 75^{th} percentiles of the overall scores distribution.

Mean and standard deviation (std) values of content scores gathered through various test conditions should be computed over all subjects. As mentioned at the beginning of this chapter, the mean of the distribution of scores accurately approximates the influences on the systems being tested. Mean scores could be either Mean Opinion Scores (MOSs) or Differential Mean Opinion Scores (DMOSs). The MOS averaging process is usually applied when subjects rated a stimulus in isolation, that means no comparison was included in the assessment method (e.g., ACR, ACR-HR and SAMVIQ). DMOS is considered when comparisons are performed. It computes the difference between reference stimulus grades with impaired content scores (e.g., ACR-HR, DSIS, CCR). The computation of MOS is introduced in Rec. ITU-T P.913 [ITU 2014d] and below.

Let N be the number of assessors having participated in an experiment, J be the number of test conditions, K be the number of contents, and μ_{ijk} be the score reported by observer i under the test condition j for the content k, MOS is computed as follows:

$$\bar{\mu}_{jk} = \frac{1}{N}\sum_{i=1}^{N}\mu_{ijk} \tag{2.8}$$

When presenting MOSs, one should associate them with their Confidence Intervals (CIs). These values show the possible range of errors made by the population sample and are computed by taking into account the standard deviation of the scores and the sample size of the participants. This computation is done under the assumption that scores follow a Student t-distribution.

It is highly recommended (Rec. ITU-T P.1401 [ITU 2012d]) to compute 95% CIs, given by $[\bar{\mu}_{jk} - \delta_{jk}, \bar{\mu}_{jk} + \delta_{jk}]$ where:

$$\delta_{jk} = 1.96\frac{\sigma_{jk}}{\sqrt{N}} \tag{2.9}$$

The standard deviation for each stimulus, σ_{jk}, is given by:

$$\sigma_{jk} = \sqrt{\sum_{i=1}^{N}\frac{(\bar{\mu}_{jk} - \mu_{ijk})^2}{N-1}} \tag{2.10}$$

Then, common practices are to perform a linear Pearson's correlation study to relate influence factors. It should be recalled that correlation does not imply causation. Independent variables are the combinations of influence factors, which are the test conditions and test materials.

Linear Pearson correlation coefficient (LPCC) is defined in Rec. ITU-T P.913 [ITU 2014d] as

$$LPCC(x,y) = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n})(\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n})}} \tag{2.11}$$

where x and y are data arrays and n is the number of data points. A $LPCC$ value of 1 indicates a total positive correlation, 0 shows no relation, and -1 is a complete negative correlation.

In statistics, the effect size is a quantitative measure of the magnitude of a phenomenon. In our context, the effect size of independent variables is observed trough MOSs and CIs as well as correlation results.

To understand if the phenomenon studied has different effect size on two population samples, statistical analyses are necessary. Discriminability [Zakzanis 2001], also named sensitivity, measures the magnitude of difference between two sample means. Therefore, it indicates whether the difference between distributions is statistically significant or not.

$$\bar{d} = \frac{|\mu_{smp1} - \mu_{smp2}|}{\sigma_N} \tag{2.12}$$

The equation 2.12 stands when populations size and the std of their distributions are the same. If $\bar{d}$ is low, it indicates that scores of the two populations come from the same distribution. There is a statistically significant difference between the populations' results if $\bar{d}$ is high.

Statistical analysis can be performed through a Student's t-test [Haynes 2013], which determines whether the difference between the scores means of two populations is significant or insignificant. Assuming that scores exhibit a (Gaussian) normal distribution with an unknown std, this test compares the empirical means of the two populations scores distribution. A two sample Student's t-test statistical hypothesis test is implemented as below:

$$t_N = \frac{|\mu_{smp1} - \mu_{smp2}|}{\frac{\sigma_N}{\sqrt{N}}} \sim \bar{d} \times \sqrt{N} \tag{2.13}$$

We can observe that in Equation 2.13, population size N influences the statistical significance of the difference between empirical means. That is statistical significance can be reached, as long as the test is performed on a sufficiently large population.

Another statistical test is the ANOVA analysis. When compared to Student t-test, ANOVA, also based on the assumption of (Gaussian) normal distribution of scores, is not as strongly biased by sample size and enables the analysis of the differences between the means of several groups. Here are the details of ANOVA analyses.

Let us define variability between independent variables as the between-groups variance, which results from influence factors effect. The variability within independent variable is the within-group variance, which occurs due to discrepancies across subjects. F is the ratio of variability between independent variables over variability within independent variables. If F is large, there is at least one group which presents a statistically significant difference from the other independent variables. If F is small, there is no statistically significant difference between distributions. Figure 2.8 illustrates two scenarios depicted above.

To express when F is large enough to conclude there is a statistically significant difference between populations, a statistical hypothesis test is performed. Null hypothesis H_0 is defined as 'Group distributions are equal' and alternative hypothesis H_1 is defined as 'At least one of the group distribution differs from others.' If the null hypothesis is correct, F tends to be equal to 1. Thus, we compute the probability that a value equal to or greater than the experimental F-value would be observed if the null hypothesis is true. If the probability of observed F-value, i.e., the p-value, is low, the null hypothesis is rejected and the alternative hypothesis is more likely to be true. A significance level of 0.05 (p-value <= 0.05) expresses a risk of 5% to conclude that there is a difference when no actual difference exists. The p-values of independent variables, and combination thereof if applicable, are analyzed to conclude the statistical difference between independent variables.

As the alternative hypothesis does not indicate which group is statistically different from others if a p-value is smaller than 0.05, post-hoc tests, such as a Bonferroni

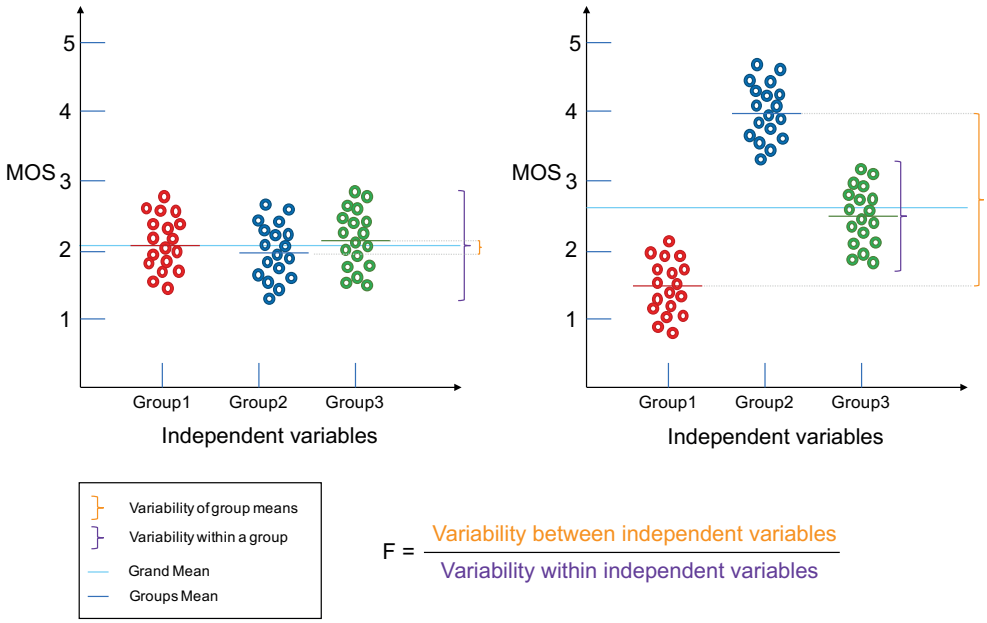


Figure 2.8: ANOVA illustration. On the left, F is small, indicating small variation between independent variables. On the right, F is large, showing that at least one variable follows a different distribution.

multiple comparison test [Armstrong 2014], can be implemented to identify which distribution differs from others.

2.3.8 Physiological signals

During the last decades, subjective evaluations including physiological signals gained momentum [Engelke 2016]. Standardization committees are taking interest in such assessments as ITU and Video Quality Experts Group (VQEG) have established working groups to develop recommendations for physiological assessment methodologies. For instance, P.PHYSIO is a 2018 ITU-T SG12 work program for speech processing[14].

Physiological signals allow the observation of implicit and uncontrollable physical reactions to stimulation. They remove biases such as experiment effect, reduce learning effect and are not strongly impacted by cognitive load. Those facts firmly justify the interest in using physiological signals in subjective evaluations.

Physiological signals originate from two nervous systems, the Central Nervous System (CNS) and the Peripheral Nervous System (PNS).

CNS lies in the brain and spinal cord, sending or receiving messages to control the entire body activities, known to include emotional information. EEG is the primary physiological method to record central nervous system activity. It is a monitoring method that records the electrical activity of the brain. Typically, a

[14] http://www.itu.int/itu-t/workprog/wp_item.aspx?isn=13800

non-invasive set of electrodes are placed on the scalp surface. It measures voltage fluctuations generated by the simultaneous activation of thousands of neurons.

PNS consists of the nerves outside the brain and spinal cord. It is composed of the Somatic Nervous System (SNS), which is responsible for touch, vision, smell, taste, hearing and balance, and the Autonomic Nervous System (ANS) which regulates physiological signals involuntary responses. Various peripheral signals can be recorded. Among them are

- ECG, recording the electrical activity of the heart using electrodes placed on the skin,
- Galvanic Skin Response (GSR), also named Electrodermal Activity (EDA), skin conductance, electrodermal response, etc. GSR records the skin resistance variations that come from sweat glands in the skin. Sweating is a physiological or psychological arousal indication,
- skin temperature,
- respiration,
- Blood Volume Pressure (BVP), an optical non-invasive method measuring variations in blood volume in an arterial extremity.

We stress here that CNS and PNS modalities bring complementary and redundant information [Chanel 2006]. Therefore, any combinations of EEG with peripheral signals provide efficient and robust measures of human responses.

However, recording those signals during subjective tests reveals to be more complicated than expected. Preparation of the equipment for physiological signals (e.g., material setting, subjects preparation, material cleaning), as well as resting periods and base signals (e.g., eye movements, deep breath) recordings increase test duration drastically. Subjects come across different kinds of difficulties. They can experience boredom during the setting phase and discomfort with the equipment along the experiment. All of these matters can dramatically impact their ratings. Despite expensive high-quality material, physiological signals will contain a high noise level.

2.3.9 Subjective test validity

This section introduces the notion of validity of subjective evaluations. It also goes over various considerations that need to be considered when designing a test. Validity measures the degree to which a subjective experiment genuinely evaluates what is meant to be appraised.

As not fulfilling some requirements can invalidate experimental results, it is essential to understand the strengths and weaknesses of the experiment methodology. It helps to make sure to draw comprehensive and accurate conclusions. We can distinguish several subjective test validity types, all contributing to assess the soundness and robustness of subjective evaluation design.

Probably the most important one is referred to as construct validity, showing how well the experiment measures up to its claim. For instance, a test evaluating the SoP should not only investigate closely related concepts such as immersion or engagement. It should include pure SoP evaluation. Otherwise, sense of presence would not be assessed properly.

The second type of validity, which has already been partly and briefly mentioned in Section 2.3.3, is content validity. It estimates the extent to which subjective test environment matches reality. For instance, test material should represent diverse classes of media contents, and the equipment used should be consistent with the ecosystem of the technology under test.

The third validity type, internal validity, is a concept close to construct validity. Internal validity indicates up to which degree we can consider correlation as causation. If two independent variables are isolated enough from other influence factors, any correlation can be viewed as causation. Otherwise, correlation is not causation. If correlation implies causation, part of construct validity is fully established.

The fourth type of validity is the external validity, which informs if obtained results can be generalized to a broader context than the experimented one. For instance, it specifies if we can apply conclusions to different populations, settings and other independent variables. External validity is usually considered under two perspectives, population validity and ecological validity, both fundamental to appraise experiment design strengths. Population validity is discussed in Section 2.3.4. Ecological validity identifies and determines the influence of testing environment on subjects' behaviors.

Apart from validity considerations, several effects are widely recognized as influencing subjective results, to mention but a few, experiment bias, learning effect and subjects cognitive load.

Experiment bias arises when the experimenter, usually unconsciously, leads subjects to generate desired results instead of providing independent opinions. The unconscious indication of the expected answer can occur when formulating the questionnaire, when giving oral instructions, through body language or voice tone or even by the assessment methodology. This explains the importance of designing subjective evaluations carefully to obtain subjects' impartiality. This effect is closely related to the Hawthorne effect in which individuals who are aware that they are observed tend to modify their behavior.

Learning effect results from subjects identifying the test design. Their answers are based on the knowledge acquired during previous stimuli presentation, instead of focusing on current stimuli experience. For instance, a subject can learn that a stimulus without audio will always generate a less satisfying experience, even though other parameters (such as resolution) have increased in quality. Learning effect is also related to presentation order, as experiencing a condition A and then a condition B could potentially exhibit better performance of condition B. This emphasizes the importance of randomizing the order of stimuli presentation for each subject.

The cognitive load refers to the extent of mental effort required in terms of memory and information processing when realizing a task. A heavy cognitive load

can have adverse effects on task completion. This explains the reluctance to design a subjective test implementing sequential presentation of video pair-comparisons.

Standardization committees provide researchers with numerous recommendations necessary to run valid subjective tests. From laboratory settings to test duration and content selection, many contextual biases are removed from evaluation following these guidelines. Also, these recommendations are broad enough to cover numerous subjective test designs, regardless of the experiment technology or purpose. This triggers the fact that there is no description of how to formulate the question asked to subjects during the evaluation. Experimenters are free to develop their questionnaires (such as in gaming evaluations). According to ITU guidelines, subjects should answer only one question.

However, even if quality could be investigated with one question (and this fact is disputable), QoE would require further questions. For instance, additional components such as subjects' prior knowledge, their former experiences or their present state of mind may be examined. One can argue that these pieces of information can be investigated before and after the experiment, while one can also include several questions in the test session to study the experience resulting from every stimulus.

The remainder of this document reports activities done in view of providing guidelines for conducting subjective evaluations for context-based QoE in immersive technologies.

Chapter 3

Support the transition from SDR to HDR

Contents

To get accustomed to best practices when designing a subjective test and identify challenges to address when considering emerging technologies and QoE, two works have been done on HDR imaging.

Contexts of both studies are alike and consist of a partnership with Dolby Laboratories.

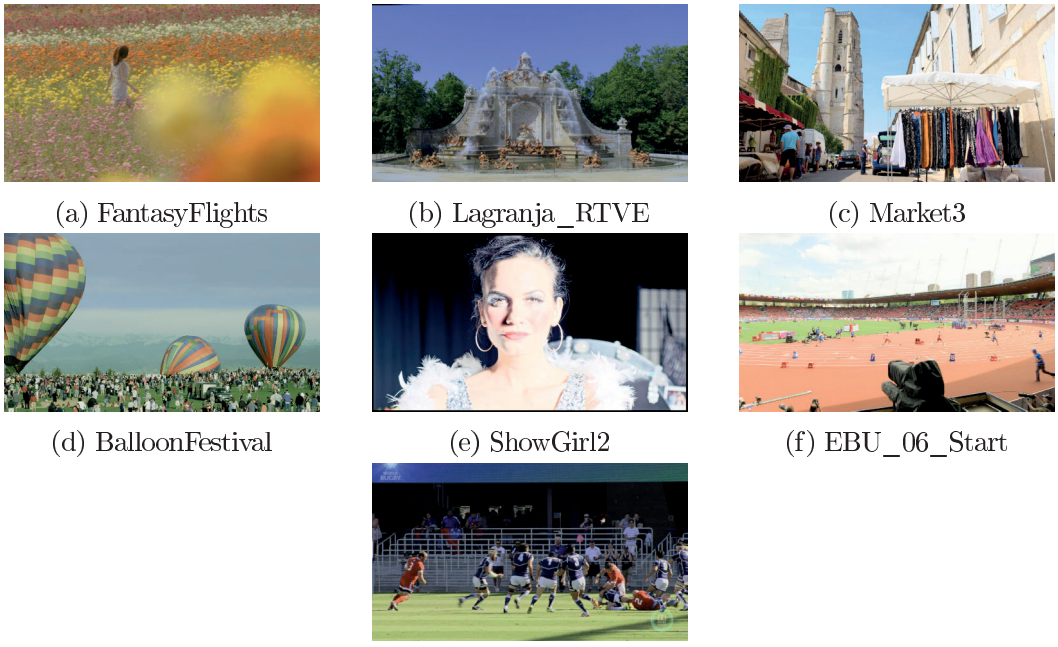
Dolby has dedicated many resources to research on HDR and WCG. The resulting contributions are numerous and most are held by various standardization bodies such as Advanced Television Systems Committee (ATSC), SMPTE, Digital Video Broadcasting (DVB), ITU, ETSI and European Broadcasting Union (EBU). Among the main contributions are the PQ transfer curves described in Section 2.1.4.2 and an end-to-end provider solution: Dolby Vision™. This framework implements what Dolby named "better pixels". Pixel enhancement is permitted through the use of emerging technologies which better represent a scene, namely HDR and WCG. Combining both technologies results in richer experiences as increasing the dynamic range enables the representation of deeper blacks and brighter highlights and expands the contrast. Enlarging the color gamut offers more refinement in the representation of colors as well as having more vivid colors. Dolby Vision™ includes solutions for content production (acquisition and edition), distribution (encoding and delivery)

and playback (display and EOTF). In addition to their end-to-end solution, Dolby has proposed a new color space, ICtCp, which most compelling strengths come from the luminance channel hue linearity and lack of correlation with color channels.

Our works find themselves in the validation (or not) of solutions proposed by Dolby with the perspective to provide guidelines for the use of HDR and WCG. As mentioned already, the construction of the ecosystem for these standards is ongoing.

The first work investigated the differences between ICtCp and Y'CbCr for HDR and WCG imaging. An EETF, named reshaper was included in this analysis. This conversion optimizes the content description given the targeted screen DR for its compression. This is the first time that both the ICtCp color space and the reshaper are evaluated jointly and subjectively.

The second work studied the efficiency of Dolby Vision^{TM} backward-compatible compression solution compared to typical uncompromised SDR and HDR delivery. To the best of our knowledge, this is a first attempt to compare HDR and WCG compression solutions for both SDR and HDR delivery.



(a) FantasyFlights (b) Lagranja_RTVE (c) Market3

(d) BalloonFestival (e) ShowGirl2 (f) EBU_06_Start

(g) Rugby

Figure 3.1: First frame of test sequences used in HDR representation and compression evaluations.

It should be stated that Dolby Laboratory has developed the evaluated solutions in the following works (ICtCp, reshaper, and codecs). The company has provided the test material (which involves the content selection and preparation) and the equipment (playback solution and SDR and HDR displays).

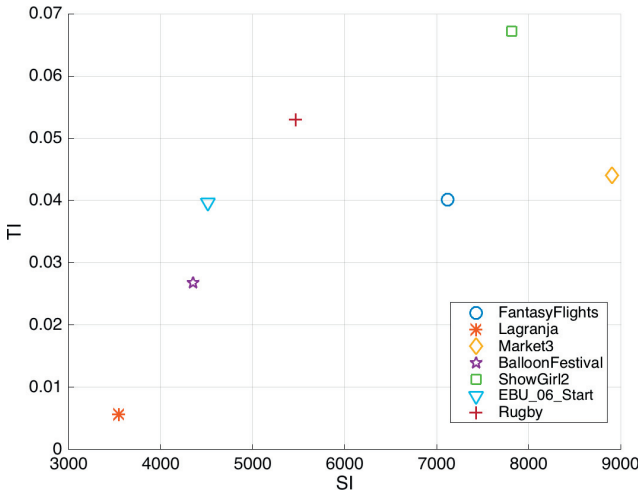


Figure 3.2: SI and TI of HDR representation and compression evaluations sequences.

Contents

Test sequences used in conducted subjective tests come from MPEG and Dolby collections. Three MPEG contents, namely $Market3$, $BalloonFestival$, and $ShowGirl2$, and four Dolby sequences, namely $FantasyFlights$, $Lagranja\_RTVE$, EBU_06_Start and $Rugby$ were used. $Rugby$ and EBU_06_Start sequences have been added due to their representativeness of sport event broadcasting contents. Figure 3.1 shows the initial frame of each sequence. They were selected for good testability regarding the presence of textures (high and low frequencies), various spatial and temporal complexities.Figure 3.2 indicates the spatial and temporal complexities of sequences (see equations 2.3 and 2.4).

HDR contents were available in 4:4:4, BT.2020, 1080p (HD), 16 bits, 50 fps. $FantasyFlights$, $BalloonFestival$ and $ShowGirl2$ contents are originally 24, 24 and 25 fps, respectively. Frame rate consistency is mandatory for the rendering devices and software used in the studies. To cope with the 50 fps constraint, each original frame of these three sequences was duplicated. The modification of frame rate in those sequences caused temporal jerkiness. SDR contents have the same color sampling, resolution and frame rate as HDR contents but were represented with 10 bits, BT.709 in a BT.2020 container. Contents were generated given the use of the PQ transfer function.

To be free to use any subjective test methodology, reference contents should be available in both SDR and HDR formats. This is a severe constraint as few contents are acquired in SDR and HDR. If not available, one can reliably and accurately derive SDR reference streams from HDR original sequences through tone-mapping and color grading. In this thesis, when this process was required, experts manually performed the color grading.

Test equipment

HDR display systems, made available by Dolby, were two 1080p, 2000 cd/m^2 Dolby Maui monitors. These monitors were calibrated accordingly to the SMPTE ST.2094 [SMP 2017], which sets the minimum peak luminance of Liquid Crystal Display (LCD) panels for HDR at 1 000 cd/m^2. As recommended in Rec. ITU-R BT.2100 [ITU 2016], displays' black level were set to 0.005 cd/m^2. The PQ was used as EOTF to render HDR contents [Litwic 2016].

SDR content rendering system comprised two HD Dolby professional reference PRM-4220 monitors, set to 100 cd/m^2, the typical SDR peak luminance [SMP 2017].

The two monitors of each display system were positioned in a side-by-side fashion. In this layout, two subjects can simultaneously attend the same test session. Dolby Vision^{TM} was used for playback. Rec. ITU-R BT.2022 and BT.2100 [ITU 2016, ITU 2012b] were followed to define the viewing distance to 3.2H, for both systems. Separate rooms were set up to perform SDR and HDR experiments in parallel.

3.1 HDR representations

The interest in HDR and WCG is unquestionable as it promises an increase in the quality of experiences provided to end-users. Yet, current practices do not guarantee the optimal use of these technologies. Such imaging is usually built on top of legacy ecosystems such as color spaces or compression schemes, not always compliant with new technologies. For instance, the EOTF used in SDR imaging, the gamma correction, is specifically designed to match CRT TV characteristics to encode images using an optimized number of bits. This correction attributes too many code words to high luminances. This constraint was not an issue for SDR but is for HDR. Consequently, HDR and WCG development involved creating new TFs, for instance the HLG and the PQ, tackling legacy TFs limitations. Various aspects, such as TFs, may be improved and optimized for HDR and WCG use.

Dolby Labs proposed solutions which consist of a new color model and an EETF. The former intends to cope with Y'CbCr color space limitations, impacting the amount and type of artifacts introduced during compression operations. The latter, namely reshaper, is optimizing the representation of streams before their compression, reducing quantization errors from a perceptual perspective. The evaluation of those solutions is crucial, as it impacts the future development and deployment of these emerging technologies. Subjective evaluations had to be carefully designed and conducted to provide valid conclusions on the impact of those solutions on HDR and WCG optimal use.

In Dolby's end-to-end framework, three compression schemes are offered. This enables backward-compatible or not delivery of enhanced streams. Choosing between coding schemes for HDR delivery may jeopardize broadcasters services. Indeed, no comprehensive evaluation of the quality of SDR and HDR streams has been conducted. Broadcasters do not have the information necessary to understand the

trade-off made when selecting one compression strategy. It is mandatory to provide those insights to guide the selection of a compression approach.

In the following subsections, the solutions of Dolby Labs to improve HDR and WCG representation are presented first. Subjective test design addressing the issues mentioned above and their results are then indicated. Finally, conclusions drawn from experiment are reported. The same structure is followed to present our study on HDR compression schemes. An overall conclusion summaries the works outcomes at the end of this chapter.

3.1.1 ICtCp vs. Y'CbCr

Color space selection to represent multimedia contents impacts the entire production and delivery chain, especially content compression. A NCL color space has chrominance information present in the luminance channel signal. Color artifacts, created during the quantization and chroma subsampling operations of content compression, will also impact the luminance channel. When using HDR and WCG imaging, color, contrast and luminance precisions are increased. Streamed contents will thus present more noticeable compression artifacts.

Faithful representation is needed to represent the extensive amount of information conveyed by HDR and WCG. It is the opportunity to change color models to address the lacks of traditional representations. Dolby Labs seized this opportunity and developed the ICtCp color space.

To conduct a fair evaluation of a color space, we must compare it to a state-of-the-art widely used representation, the NCL Y'CbCr, presented and defined in Section 2.1.3. When considering HDR and WCG, Y'CbCr color model limitations are listed as follows in Rec. ITU-R BT.2390 [ITU 2015d]:"

- quantization distortions (bit depth limitations)
- chroma subsampling distortions due to a perceptually uneven code word distribution
- color volume mapping distortions due to incorrectly predicted hue and luminance
- error propagation from the chroma to luma channel.

"

ICtCp color space has been designed by Dolby Labs to tackle above-mentioned Y'CbCr limitations. This color model is built upon the well-known IPT color space [Edge 2013], addressing non-constant hue lines (when varying saturation and luminance) by exploiting the LMS of human vision. ICtCp improves IPT by representing higher dynamic range (up to 10 000 cd/m^2) and wider color gamuts [Lu 2016a]. ICtCp representation is obtained by first calculating the LMS response to light, which is then fed to the nonlinear PQ filter. Finally, a color differencing process is applied through a specific transformation matrix (3×3 matrix).

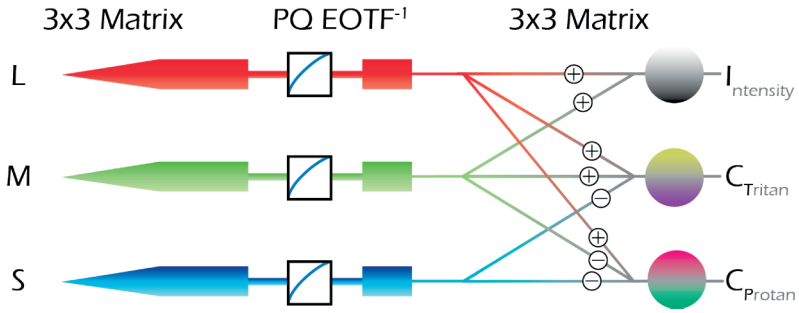


Figure 3.3: ICtCp color representation [Labs 2016b].

The computation of ICtCp values is described in details in [Labs 2016b] and Rec. ITU-R BT.2100 [ITU 2016]. It is illustrated in Figure 3.3. Intensity (I) corresponds to nonlinear pixels luminance, and yellow-blue Tritan (Ct) and red-green Protan (Cp) are the color channels.

ICtCp color space characteristics are exposed and demonstrated in [ITU 2015d, ISO 2016b, Lu 2016a], and summarized below.

- Achromatic channel I and isoluminance:
  As already mentioned, compression causes severe impairments should there be mutual information in luminance and chrominance channels. Decorrelating luma and chroma signals prevents the appearance of such discrepancies. Also, the intensity channel I corresponds closely to PQ luminance Y, by construction, indicating the isoluminance of ICtCp.

- Hue linearity:
  The hue linearity means that any variation in saturation or luminance does not impact the hue. ICtCp has shown straighter constant hue lines when compared to Y'CbCr.

- Perceptually uniform colors:
  The contours of MacAdam ellipses, which indicate chromaticity Just-Noticeable Difference (JND) [MacAdam 1942], form more circular shapes on the gamut representation of ICtCp than on that of Y'CbCr. Accordingly, ICtCp is perceptually more uniform than Y'CbCr and offers efficient chroma subsampling.

- Quantization to limited bit depth:
  Taoran et al. [Lu 2016a] showed that 10-bit ICtCp contains more information than 10-bit Y'CbCr and evaluated the gain at about 1.5 bit. Moreover, they demonstrated that ICtCp is less prone to quantization errors. Last but not least, ICtCp has a similar conversion complexity as Y'CbCr.

Overall, these characteristics favor ICtCp color space exploitation to represent HDR and WCG contents.

3.1.2 Adaptive reshaper

The adaptive reshaper, described in [ATS 2015, ISO 2016a], is an EETF. It converts reference display dynamic range (from 0 to 10,000 cd/m^2) to that of current display. Regarding these displays' limitations, the reshaper enables accurate rendering of smaller dynamic range contents. This operator can create SDR signals from HDR signals. However, this capability is beyond our evaluation scope.

The reshaper range mapping considers both color and luminance compression efficiency, by providing separate processing for luminance and chrominance channels. The model for luminance is a piecewise polynomial of the second order, with up to eight pieces. The chroma components model consists of a piecewise linear model with up to thirty-two pieces. The gain in coding efficiency is achieved due to the aforementioned parametric reshaping models which implement the following:

1. code words redistribution is adapted to pixel brightness and
2. luma and chroma signals are re-quantized, which eventually causes variations in the allocation of bit rate among the three channels.

The reshaper pre-processes the signal before analyzing the content and its artistic intent. Based on conducted analyses, an adaptive code word assignment is performed, from which is derived the reshaping curve. The curve is then modeled in luma and chroma polynomials, which characteristics are sent along with the signal. The signal is then fed to an encoder and compressed by normal means. Upon decoding, a Look-Up Table (LUT), used for retrieving pixel values, is derived from the polynomials. The post process LUT result is the original input signal representation. The reshaper process is fully described in [Lu 2016b].

3.1.3 Subjective assessment

To the best of our knowledge, presented solutions have not been subjectively evaluated. A benchmark of ICtCp performance using objective metrics for compression efficiency has been presented in [Lu 2016a]. Authors recommend ICtCp as standard color space in HDR and WCG applications. Subjective evaluations should confirm this result.

The reshaper compression efficiency has been objectively evaluated in [Lu 2016b]. Results show that using a reshaper improves the HDR HEVC Main 10 encoding efficiency. The objective quality benchmark implemented the DE100, L100, OSNR, and PSNR for X, Y, and Z components metrics to evaluate the reshaper and ICtCp compression efficiency.

Experiments should entirely and comprehensively appraise the benefits of ICtCp color space for HDR and WCG and indicate if the reshaper efficiency in terms of bit rate saving is not compromising the visual quality.

The test design must consider both evaluations at once while the analyses of results should clearly distinguish the color spaces comparison from the reshaper evaluation. Indeed, on one hand, we compare HDR sequences represented in Y'CbCr

and ICtCp perceptual quality. On the other hand, the reshaper operator, targeting three different bit rate saving levels, is evaluated on the two color spaces.

3.1.3.1 Test material

We propose the following naming convention to refer to the various systems under test:

- P00: Y'CbCr signal (without a reshaper)
- P10: Y'CbCr signal reshaped at a 100% bit rate of P00
- P11: Y'CbCr signal reshaped at a 90% bit rate of P00
- P12: Y'CbCr signal reshaped at an 80% bit rate of P00
- P20: ICtCp signal (without a reshaper)
- P31: ICtCp signal reshaped at a 100% bit rate of P20
- P32: ICtCp signal reshaped at a 90% bit rate of P20
- P33: ICtCp signal reshaped at an 80% bit rate of P20.

Testing conditions are sorted as proponents and anchors to make comparison process clear. Indeed, proponents are challenging the anchors, which represent the reference stream. P00 is an anchor; remaining testing conditions will be referred to as proponents. P20 is a proponent regarding the comparison of color spaces and is an anchor when evaluating the reshaper efficiency.

The percentages of bit rate savings targeted by the reshaper are 0%, 10%, and 20%, leading to proponents P10/P30, P11/P31, and P12/P32, respectively. The two anchors indicate reference bit rates; P00 for Y'CbCr stimuli and P20 for ICtCp stimuli. The content Rugby was reshaped at different bit rates to match the visual quality of anchors. The actual bit rates for this content correspond to 100%, 95%, and 90% of P00 and P20.

Test sequences are 10-second extracts of *Market*3, *BalloonFestival*, *ShowGirl*2, *EBU_06_Start* and *Rugby* contents. This duration was selceted based on Rec. ITU-R BT.1788 [ITU 2011a].

The efficiency of both solutions is evaluated on four bit rates. Four bit rates to encode contents C1 to C4 anchors were selected according to [ISO 2016c]. Dolby Labs selected bit rates for content C5. Table 3.1 summarizes bit rates for each content, anchor, and proponent. It should be noted that R1 is the highest bit rate while R4 is the lowest.

A complete description of the stimuli creation process, particularly the reshaper use, is shown in [Lu 2016b].

name	sequence	frames	fps	br [$Kbps$]	Y'CbCr w/o reshaper	Y'CbCr w/ reshaper			ICtCp w/o reshaper	ICtCp w/ reshaper		
					P00	P10	P11	P12	P20	P30	P31	P32
C1	Market3	400	50	R1	5694	5603	5192	4595	5716	5775	5163	4616
				R2	2695	2662	2391	2144	2695	2670	2410	2160
				R3	1725	1717	1535	1369	1735	1724	1552	1385
				R4	1268	1257	1134	1003	1277	1275	1147	1020
C2	BalloonFestival	240	24	R1	4345	4313	3880	3506	4466	4378	3957	3476
				R2	2579	2532	2299	2032	2608	2518	2317	2055
				R3	1569	1566	1404	1244	1595	1564	1388	1245
				R4	1237	1232	1116	1004	1261	1245	1112	989
C3	ShowGirl2	339	25	R1	3444	3392	3051	2786	3380	3377	3151	2805
				R2	1680	1663	1538	1353	1669	1687	1531	1346
				R3	1004	999	895	790	997	1007	918	809
				R4	602	601	534	492	596	589	539	476
C4	EBU_06_Start	500	50	R1	2675	2671	2367	2127	2657	2692	2374	2109
				R2	1623	1603	1458	1296	1631	1606	1434	1295
				R3	835	825	746	656	844	848	754	667
				R4	522	517	456	414	541	517	468	414
C5	Rugby	430	50	R1	4337	4362	4104	3757	4420	4281	4142	3888
				R2	2626	2692	2490	2364	2665	2692	2495	2377
				R3	1584	1623	1509	1439	1606	1578	1514	1453
				R4	971	960	922	863	989	972	934	874

Table 3.1: Characteristics of test stimuli: name, frame rate and bit rates in Kbps.

3.1.3.2 Test methodology

The test methodology influences the type of results and conclusions possibly drawn from those. Color space evaluation indicates the necessity to compare Y'CbCr and ICtCp in a SbS fashion. Indeed, a single stimulus methodology or sequential pair comparison will require a considerable cognitive load from subjects. Besides, both color spaces enable an adequate content representation, making single stimulus presentation highly likely to be inconclusive.

Color discrepancies between Y'CbCr and ICtCp variants can be hard to notice for naive viewers. Also, test stimuli are 10-second videos. Their short duration makes assessment difficult for subjects, mainly because of some contents' high temporal complexity (e.g., C1 and C3). The difficulty in perceiving the small differences between test stimuli, especially with regards to the short duration of stimuli, strongly advocates for a repetition of stimuli during the assessment.

Regarding reshaper evaluation, comparing reshaped signals with anchors makes sense. Previous comments about temporal complexity and content type also apply to this evaluation.

The analysis scope is to identify if Dolby's solutions make a difference concerning perceptual quality when compared to legacy practices, in similar conditions. A full pair comparison is time- and resources-consuming. Besides, Comparing proponents to anchors compressed with a different bit rate will indicate when two systems, under different conditions, are performing equally. This indication is slightly out of the scope of our research interests. Additionally, some comparisons between systems will not provide much in-depth knowledge for our purposes. Hence, it was decided to conduct a partial PC evaluation.

The selected test methodology is, based on previously exposed reasons, a partial SbS PC with repetition methodology, described in Rec. ITU-T P.910 [ITU 2008d]. A pair comparison, also named test comparison, consists of a set of two test stimuli. To remove any contextual bias, test stimuli position (left or right) and test stimuli order, were randomized. The asked question is reported in Table 3.2.

Two Maui monitors, positioned SbS, render test comparison stimuli in a synchronous manner. Figure 3.4 illustrates the settings used in our laboratory environment.

Overall quality: overall image quality, image artifacts, color representation, environmental aspects.			
	Which one has a better quality?		
Stimulus number	Left	Same	Right

Table 3.2: Questionnaire for HDR representations test

A ternary categorical scale (left, right, and same) offered subjects with the possibility to provide their preference concerning the quality of stimuli. Observers were

Figure 3.4: 2000 nits Maui monitors, set at 1000 nits

asked to pay attention to compression artifact and color fidelity.

The PC is partial because proponents are only compared to anchors which have the same bit rate. Therefore, the test design led to a total of 200 test comparisons. When comparing color spaces performance, ICtCp proponents, namely P20, P30, P31, and P32, were compared to the anchor P00. The combination of five contents encoded at four different bit rates and four proponents resulted in 80 test comparisons. The remaining 120 test comparisons are dedicated to the evaluation of the reshaper operator. In details, with X={1,2,3}, the comparisons between P00 and P1X appraised the reshaper for the Y'CbCr color space, when P20 and P3X assessed the mapping function on ICtCp representations.

Accordingly to Rec. ITU-R BT.500 [ITU 2012c], no viewing session was designed to last more than 20 minutes. Too long viewing sessions would lead to subject's fatigue or lack of attention, reducing the accuracy of the collected ratings. In consequence, the entire experiment was made of six test sessions. Sessions SA and SB were evaluating color spaces differences, that is they were comparing P00 to P20, P30, P31, and P32. SC and SD were responsible for the assessment of the reshaper when contents are Y'CbCr, so P00 was compared to P10, P11, and P12. The last sessions SE and SF appraised the reshaper benefits on the ICtCp color space, which means that P20 was compared with P30, P31, and P32. The experiment implemented a between-subjects test method, thus each participated only in two test sessions.

Prior to the experiment, a training session rendered training comparisons, one for each content. It gave the opportunity to the experimenter to describe the voting process and draw subjects' attention on severe distortions (e.g., blur and blockiness and color impairments). The 50 fps playback constraint led to temporal jerkiness

for some contents, which subjects were asked not to judge in their assessment.

When the experiment was completed, we had collected 20, 22 and 26 scores for sessions SA-SB, SC-SD, and SE-SF, respectively. The proportion of women compared to men was roughly a third. Table 3.3 describes the population samples.

	SA-SB	SC-SD	SE-SF	Overall
Number of males	15	14	16	45
Number of females	5	8	10	23
Average age	22.35	21.91	21.46	21.87
Standard deviation of age	3.39	3.12	2.55	2.98

Table 3.3: Characteristics of population samples, per test sessions

3.1.3.3 Validity

Several validity concerns can be raised about this evaluation.

First of all, the test has been run involving PQ EOTF all along the process (content production, ICtCp transformation matrix, reshaper fitting curve, rendering transfer function). This indicates that the test is showing ICtCp and reshaper performances in an optimal context (using the PQ). To fully assess solutions and to verify our test's external validity, a less optimal setting should be considered by using the HLG OETF function for rendering. This technically sound comment is considered as negligible since evaluation aims at a proof of concept and both PQ and HLG are standardized transfer functions.

Secondly, using test contents in a training session is arguable as it may be biasing subjects. The test design itself justifies this procedure: assessing differences between two short video contents, with hardly noticeable discrepancies, is difficult. It was essential to guide subjects to pay attention to specific areas within contents. This allowed participants to perceive impairments and to decide which one of the two contents is preferred.

One can also question the specificity of test sessions, which assess one aspect of the study (Y'CbCr vs. ICtCp, Y'CbCr reshaper and ICtCp reshaper evaluations). It can be said that if each test had been performed separately, this concern would not have been expressed. Moreover, between-subjects test methodology increases the entire test validity. It reduces the learning effect in subjects and runs the study on a broader population sample.

A last comment regarding the test design is that PC may be not sufficient to assess color artifacts. Indeed, a reference is needed to be able to evaluate color shifts. However, the studied use case focuses more on an evaluation of users' preference than an actual measure of color space faithfulness to the reference content. Also, it would be arguable to set Y'CbCr contents as reference, as this representation is known to be limited.

3.1.4 Results and analysis

To analyze collected preference ratings, we investigated differences among distributions of scores in three rating categories (left, same and right), normalized by the number of scores per stimuli. This analysis was performed for each content and overall.

For each test comparison, normalized distributions indicated the proportion of preference for the anchor or the proponent perceptual quality (left or right) as well as the non-preference or indistinguishable difference proportion (same).

To analyze our results further, we equally distribute "same" scores in the two remaining categories (left and right). Thus, a preference matrix is computed, enabling test conditions direct comparison. By combining results, we can successively conclude about the comparison between Y'CbCr and ICtCp color spaces and reshaper efficiency.

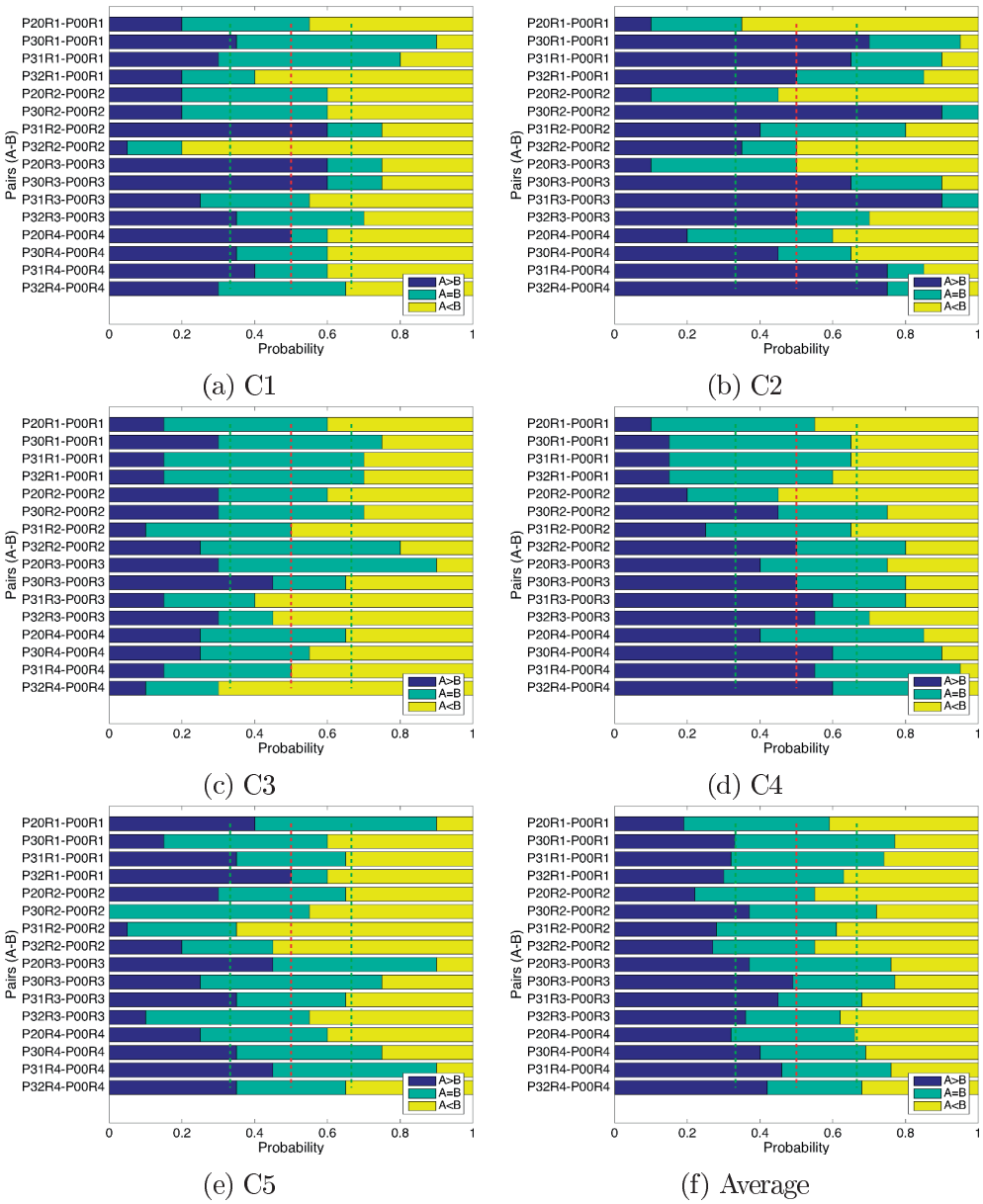
3.1.4.1 Y'CbCr vs. ICtCp comparison

Figure 3.5 presents proponents versus anchor preference and tie probabilities. Proponents P20, P30, P31, and P32 are compared to the anchor P00. No preference is observed neither for proponents or anchor. However, several trends emerge from the results. First, the proportion of "same" ratings in Figure 3.5f is such that the perceptual difference between the two color spaces appears negligible. Then, our results are content-dependent. For instance, C3 shows a preference for Y'CbCr whereas ICtCp is preferred to represent C2. C3 characteristics are its high spatial and temporal complexity. However, the preference for Y'CbCr visual quality for high complexity contents is not confirmed by C1 and C5, the two other contents presenting high spatial and temporal complexities.

Besides, the remaining contents' results show that ICtCp is preferred at lower bit rates, whereas Y'CbCr is preferred at higher bit rates. This finding is especially apparent for the proponents P20 and P30. Figure 3.6 presents the preference matrix for ICtCp proponents over P00, for all bit rates. Numbers indicate the reached average preference probability.

The previous finding, indicating the preference of Y'CbCr over ICtCp at high bit rates and the reverse for low bit rates, is stressed out in this matrix. At the bit rate R2, among all proponents, only P30 outperforms the anchor P00. As a matter of facts, P30 is preferred over P00 for all bit rates. However, P20 is the least performing proponent when compared to the other systems under test.

Overall, there is no clear advantage to use a specific color space, except if there are bandwidth constraints during the service delivery. Also, ICtCp shows more benefits when it is used with the reshaper. This last finding is promising as the reshaper enables bit rate savings while providing similar or improved perceptual quality.



(a) C1 (b) C2

(c) C3 (d) C4

(e) C5 (f) Average

Figure 3.5: The preference and non-preference probabilities of test comparisons. ICtCp proponents are compared to the Y'CbCr anchor P00. Green dashed lines split distribution graphs in thirds, whereas red dashed line delineates halves.

3.1.4.2 Reshaper evaluation

Figure 3.7 illustrates the results of subjects' preferences regarding the reshaper efficiency to deliver high-quality content. The proportion of scores among the three rating categories are similar. Hence, no proponent clearly outperforms the anchor, for both color spaces. Figure 3.8 depicts the average preference probability matrix of proponents for all bit rates and test comparisons. At R3 bit rate, anchors slightly

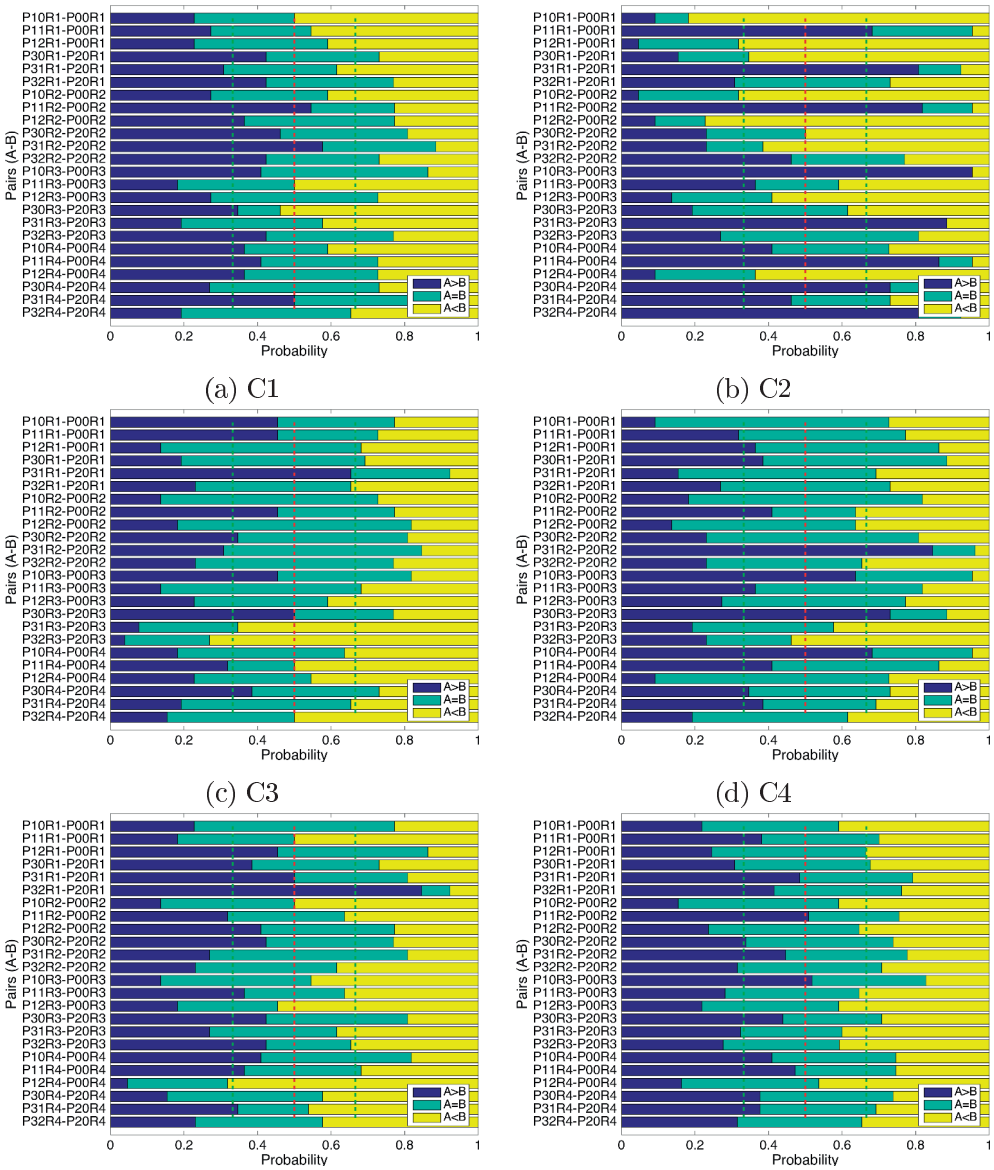
Figure 3.6: Preference probability matrix for each bit rate and for each investigated condition proponent vs. anchor.

outperform all proponents except P10 and P30. From the results, and apart from the first comment, one can observe that when it comes to low bit rates, P10 and P30 exceed their respective anchors. The proponents P11 and P31 achieve high preference probabilities for all bitrates, but R3, and are particularly preferred at high bit rates. P12, and, to a lesser extent P32, score the weakest performances. P12 is never preferred in overall.

The reshaper is more likely to be preferred over its anchor when using the ICtCp representation.

Given that the reshaper is saving bits while ensuring a similar or even preferred quality, P31, implementing the combination of ICtCp and a reshaper tuned at 90 % of the anchor's bit rate, seems promising for HDR and WCG representation and delivery.

Setting the reshaper at 100% of the input bit rate does not compromise the stream perceptual quality. This scenario may be of interest in specific use cases, such as improving the code words distribution transparently. Finally, a reshaping of 20 % is drastic and impacts the perceptual quality.



(a) C1 (b) C2

(c) C3 (d) C4

(e) C5 (f) Average

Figure 3.7: Individual pairs preference and non-preference probabilities. Anchors without reshaper P00 and P20 are compared to proponents with reshaper. Green dashed lines split distribution graphs in thirds, whereas red dashed line delineates halves.

Figure 3.8: Preference probability matrix for each bit rate and for each investigated condition proponent vs anchor.

3.1.5 Conclusion

In this work, we evaluated if the color space ICtCp is a more suitable model for HDR and WCG imaging than usual color space, namely Y'CtCp. We also verified if there is an interest to use the reshaper, the function which redistributes code words on the entire dynamic range of the content. This mapping may save bit rate by optimizing code words definition. We need to know if this bit rate saving impacts the perceptual quality to evaluate the interest of the reshaper.

In details, ICtCp was designed to considerably improve the legacy Y'CbCr, in terms of achromatic luminance channel I, hue linearity and perceptual uniformity. The abilities of this color space have been objectively evaluated as promising for HDR and WCG representations. The confirmation of this trend is discussed based on the conducted analysis of subjective scores.

The reshaper operator, adjusting the dynamic range to current DR constraints of HDR displays, enables bit rate savings. The perceptual impact of this compression operator has been appraised on two color spaces, namely ICtCp and Y'CbCr.

The run subjective evaluations implemented a partial PC SbS, with repetition methodology. Dolby professional HDR Maui and SDR PRM-4220 were the display systems used to render the 10-second stimuli. We collected more than 20 scores per stimuli before data processing.

Our results showed that the perceptual benefit achieved by ICtCp over Y'CbCr is relative: although there is no clear overall preference for a color space, perceptual quality of Y'CbCr is the highest at high bit rates while ICtCp performs better at lowest bit rates.

The reshaper operation realizes similar or slightly preferred HDR perceptual quality. Regardless of the bit rate and color space configuration, the reshaper saving 10 % of bit rate is noticeably preferred over the anchor. Finally, tuning the reshaper

at 20 % of bitrate saving should be prevented as it compromises the perceptual quality of contents.

ICtCp and reshaper combinations present equal or increased perceptual quality of stimuli, even though it allows bit rate savings.

Our recommendation about ICtCp and Y'CbCr depends on the bandwidth constraints of content providers. Under severe broadcasting constraints, enabling only limited bandwidth load, ICtCp representation should be favored, while Y'CbCr is indicated for uncompromised transmissions in terms of bandwidth. Reshaper use is recommended, especially when using ICtCp color space, at a maximum of 10% bit rate savings.

This study has been published by IEEE consumer electronics magazine of May 2018 [Perrin 2018a].

Our work has been extended in [Zerman 2017] in which they included a comparison to yet another color space, namely, Ypu'v'. They compared the compressed version of each color space with a reference and asked for the extent of visible artifacts. Their results corroborate ours when they say that there is no indication that ICtCp outperforms Y'CbCr, or the reverse. However, they observe the lower performance of Ypu'v'.

3.2 HDR compression

Broadcasting services have to cope with fast developments of technology and slow standardization processes. To face this major issue, broadcasters usually deploy new TV technologies while ensuring a smooth transition from present to future technologies. The last release of American ATSC 3.0 and European DVB broadcasting standardization committees regulated SDR BT.2020 [ITU 2015a] delivery, and they are expected to extend it to HDR BT.2020 in the near future.

In addition to suggesting a new color space and a pre-compression optimization operator, Dolby Vision^{TM} implements several compression schemes, targeting different use-cases of broadcasting community. The first one is a Single-Layer Backward-Compatible (SLBC) solution, which delivers high-quality SDR while retrieving HDR through a mapping based on metadata. The second one is a Dual-Layer Backward-Compatible (DLBC) approach. It is built on top of the previous coding scheme and ensures a high-quality HDR delivery by transmitting an enhancement layer, containing the information required to compensate for prediction errors. The last scheme is a Single-Layer Non-Backward-Compatible (SLNBC) HDR compression solution, delivering uncompromised HDR streams. SDR streams are retrieved through a metadata-based mapping of HDR signals. Overall, Dolby Vision^{TM} is a comprehensive framework implementing the three approaches of the HDR compression ecosystem.

Broadcasters urge for backward-compatible systems, as their service must guarantee a high-quality delivery, whichever format of the delivered content is and whichever is the user's equipment. An ideal solution would be a backward-compatible system which realizes high-quality SDR and HDR. The DLBC is the solution most likely to tackle this issue. Regarding the settings of this codec, it is critical to identify bit rate allocations to the EL that (1) realize high-quality SDR, (2) deliver uncompromised HDR, (3) reach a compromise, in terms of perceptual quality, between SDR and HDR decoded streams.

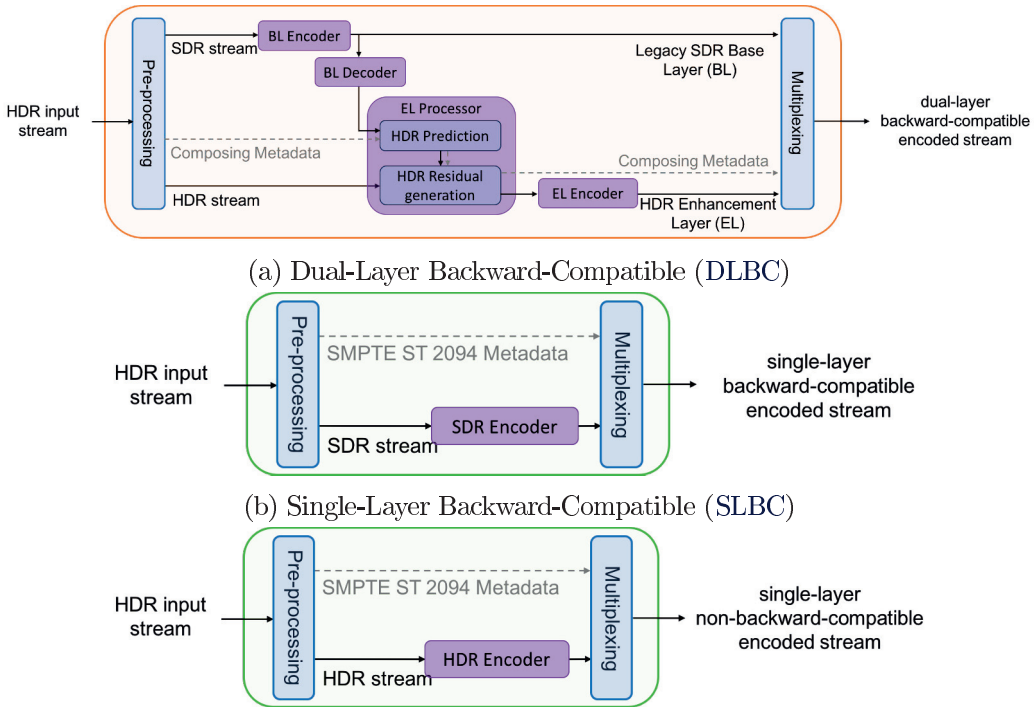
The essential points highlighted above were meant to be identified in this study. Dolby Vision^{TM}'s compression strategies were used to perform a comparison between the DLBC and uncompromised SDR and HDR delivery systems.

DLBC solution was compared to SLBC approach regarding SDR streams evaluation, and with SLNBC for HDR streams. Additionally, SLBC was included in the comparison for HDR streams to measure the efficiency of the metadata-based prediction, and to give a point of comparison regarding the benefit to transmit ELs in DLBC. The HDR compression ecosystem and related evaluations are presented in Section 2.1.4.3. Despite numerous works on HDR compression evaluations, no study considered SDR streams assessment as relevant for codecs comparison. However, it is critical for broadcasters to ensure high-quality delivery of SDR and HDR before the deployment HDR and WCG in the current market.

We propose to evaluate most representative compression solutions for HDR and SDR. For the first time, the evaluation is conducted on both SDR and HDR delivery. We draw recommendation to reach efficient compression solutions for HDR,

concerning compression scheme and bit rate settings.

3.2.1 Compression solutions



(a) Dual-Layer Backward-Compatible (DLBC)

(b) Single-Layer Backward-Compatible (SLBC)

(c) Single-Layer Non-Backward-Compatible (SLNBC)

Figure 3.9: Block diagram of codecs

Compression solutions evaluated in this study are following the specification of ETSI GS CCM [ETS 2017] and are developed upon HEVC-RExt. All three codecs are part of the Dolby Vision™ framework and therefore apply standardized processes such as PQ EOTF, parametric tone mapping [ITU 2016, SMP 2014a] and DMCVT [SMP 2017]).

The solutions are compliant with HDR and WCG formats, defined here as 16-bit depth, 4:4:4 chroma sampling and BT.2020 container. SDR backward-compatibility format refers to 8-bit, 4:4:4, BT.709, stored in BT.2020 container.

DLBC is made of a BL, which conveys the SDR stream, and a EL, which delivers the residuals to balance the prediction inaccuracy. Those residuals are computed as follows. The SDR stream is processed along with its decoded version, which is mapped to create an HDR stream. The mapping function can be polynomial or Multi-Mapper Resolution (MMR). Mapping settings, along with the flag indicating the drop of EL, are sent through composing metadata, based on DMCVT. The EL carries the quantized residual of the difference between predicted and input HDR streams.

When dealing with an SLBC, the EL is discarded, so the output HDR stream is the mapped SDR signal.

The SLNBC conveys an HDR signal multiplexed with DMCVT metadata, which carry supplementary color grading information for SDR reconstruction.

Diagrams in Figure 3.9 illustrate the three evaluated compression solutions. In Figures 3.9b and 3.9a, one can observe that SLBC is a DLBC when EL is dropped. More detailed descriptions of solutions are provided in [Diaz 2016] and [SMP 2017].

3.2.2 Subjective assessment

In this section are provided the experiments details such as material, methodology, equipment, and environment as well as test design and population samples.

3.2.2.1 Overall bit rate

Several bit rates are included in this study to investigate HD HDR delivery bandwidth constraints and provide guidelines to broadcasters regarding bit rate settings. Additionally, bit rates strongly impact the perceptual quality of streams, leading to the invalidity of conclusions if evaluating only one.

Ultra-low bit rates are not considered here as this attempt does not aim at the investigation of the limitations of compression approaches.

The selection of bit rates is based on typical broadcasting settings of UHD delivery, to provide a comparison point to content delivery services. Several studies indicated that UHD signal has an entropy about 3 to 3.3 times that of an HD. Hence, the envisioned bit rates of HD streams are 4, 5, 6 and 8 Megabits per second (Mbps), as they match typical UHD HEVC live broadcast bit rates (12, 15, 18 and 24 Mbps) [Le Feuvre 2014].

Those bit rates are allocated to the entire stream, hence are referred to as Overall Bitrates (OBs). Table 3.4 introduces the naming convention of OBs in its first raw.

	OB1	OB2	OB3	OB4
Mbps	4	5	6	8
CC1	4	5	6	8
CC2	3.6	4.5	5.4	7.2
CC3	3.4	4.25	5.1	6.8
CC4	3.2	4	4.8	6.4

Table 3.4: SDR bit rates (Mbps) for all combinations of OBs and Codec Configurations (CCs)

3.2.2.2 Configurations of compression solutions

From the codecs used in this work, five CCs are defined:

- CC1: DLBC codec, with 0% of OB allocated to the EL, assimilated as an SLBC solution. This configuration conveys SDR in its BL and reconstructs HDR through prediction. It gives a comparison point for uncompromised SDR in SDR evaluations and verifies if dropping the residual data for HDR reconstruction is decreasing HDR quality.
- CC2: DLBC codec, with 10% of OB allocated to the EL. This configuration and the two following investigate proportions of OB allocated to the EL that may yield to (1) uncompromised SDR, (2) uncompromised HDR, (3) a trade-off between SDR and HDR quality.
- CC3: DLBC codec, with 15% of OB allocated to the EL.
- CC4: DLBC codec, with 20% of OB allocated to the EL.
- CC5: SLNBC codec. This codec does not need parameter settings, therefore, stands as one CC. Its evaluation indicates the performance of a non-backward compatible solution which provides uncompromised HDR.

BLs may contain different bit rates, depending on the combination of CCs and OBs. Table 3.4 indicates the bit rates in BL for any combination of CCs and OBs. It is reminded that CC5 is not appraised in SDR evaluation.

3.2.2.3 Test methodology

The impairments introduced in the content during the compression operation are perceptually measured to discriminate media experiences. Concerning SDR stimuli, tiny differences between successive bit rates (between 0.2 and 0.8 Mbps) must be noted. To observe compression impairments, contents should be displayed along with their reference content.

Therefore, we selected the SbS DSIS version II with explicit reference methodology, in which one stimulus is the reference (original uncompressed), and the other is the compressed content.

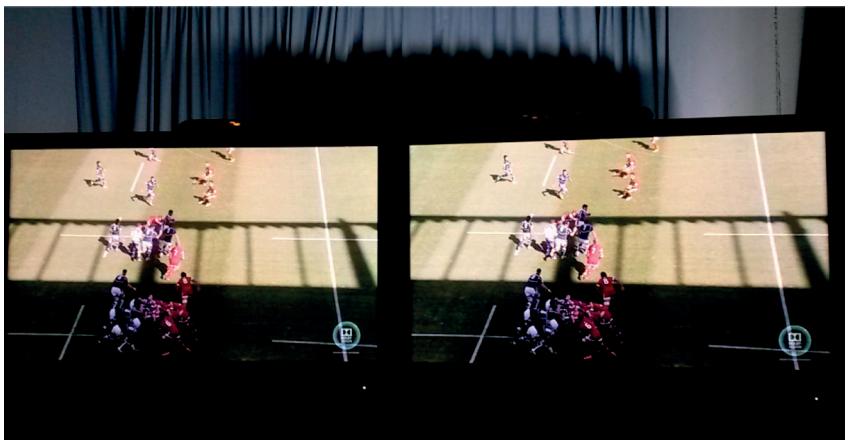
For the same reasons as above, the same test methodology was used for the evaluation of HDR streams. The laboratory environments are shown in Figure 3.10.

Each test sequence lasts for 12 seconds. The names of sequences are reported in Table 3.5. The diversity of SI, TI and DR ranges over chosen sequences is ensured. The legacy Y'CbCr color space was used to represent contents.

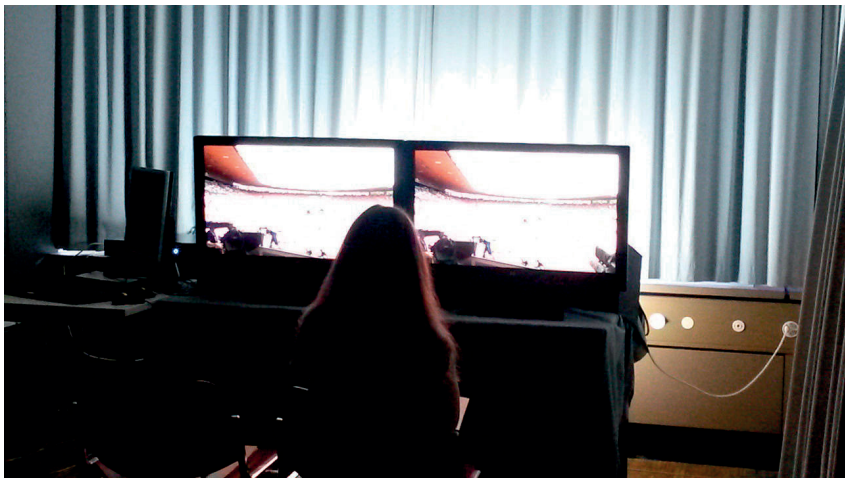
3.2.2.4 Experiment design

The combination of six sequences, four OBs and five CCs led to 120 test stimuli to evaluate in the HDR test. As CC5 was not included in SDR evaluations, only 90 SDR stimuli were under evaluation for this experiment.

Experiments were split into test slots. The HDR test slots comprised four viewing sessions while the SDR test needed only three. 30 stimuli were evaluated during one viewing session. Two dummy stimuli were presented at the beginning of each



(a) Professional Reference Monitor PRM-4220, set at 100 nits



(b) 2000 nits Maui monitors, set at 1000 nits. This SDR pictures hardly represent the DR rendered on HDR displays, this explains that bright parts of the content seem burnt

Figure 3.10: Laboratory environment setting.

viewing session to stabilize subjects' opinion. Viewing sessions were randomized with regards to content, the order of visualization, and position of the reference stimulus to reduce contextual influences. Also, they were designed to last at most 20 minutes to avoid subjects' fatigue and lack of attention.

The equipment constraints, especially an unforeseen loading time of the contents onto the rendering systems, increased the duration of test sessions to half an hour.

Four subjects are needed to run a test slot. Each subject participated in two viewing sessions, separated by 30 minutes of break. Each test slot collected two scores per stimulus.

3.2.2.5 Participants

Overall, 48 and 40 naive subjects took part in the HDR and SDR tests, respectively. Gender balance was achieved in both tests.

C	Name Sequences
1	FantasyFlights
2	Lagranja_RTVE
3	Market3
4	ShowGirl2
5	EBU_06_Starting
6	Rugby

Table 3.5: Sequences labels

A constraint of the experiment design was to evaluate SDR and HDR stimuli without interdependence between the two tests results. Hence, two different sample of the population assessed the SDR and HDR streams, contrary to [Litwic 2016].

We followed the Rec. ITU-R BT.500 [ITU 2012c] and included at least 15 subjects in our experiments to ensure a sufficient effect size for statistical analyses. Details about evaluated population samples are provided in Table 3.6.

Prior to the experiment, a training session gave the opportunity to the experimenter to describe the voting process and draw subjects' attention on severe distortions (e.g., blur and blockiness and color impairments). The 50 fps playback constraint led to temporal jerkiness for some contents, which subjects were asked not to judge in their assessment.

	HDR test	SDR test
Number of score per stimulus	24	20
Number of males	25	24
Number of females	23	16
Average age	22.9	22.4
Standard deviation of age	3.6	2.8

Table 3.6: Subjects' characteristics for HDR and SDR tests

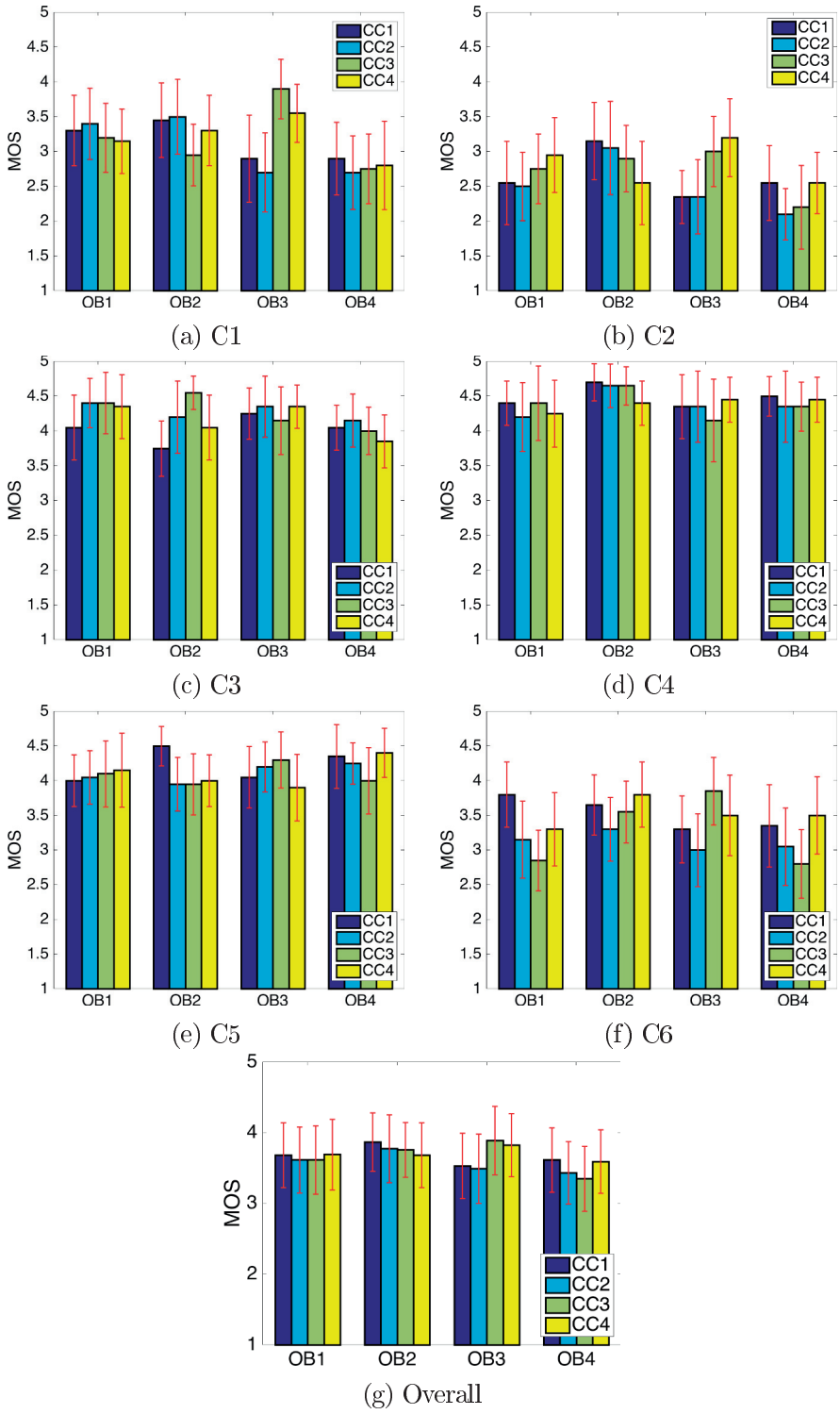
3.2.3 Results and analysis

Following the collection of impairments rating, usual statistical analysis involving MOSs and 95% CIs as well as ANOVA have been computed. Below, we analyze SDR and HDR results in details.

3.2.3.1 SDR

Figure 3.11 presents the computed MOSs and associated 95% CIs for each content and overall.

We observe that our results are content-dependent. For instance, C2 is a sequence with low temporal and spatial complexity with levels of impairments ranging



(a) C1 (b) C2

(c) C3 (d) C4

(e) C5 (f) C6

(g) Overall

Figure 3.11: MOSs and 95% CIs for SDR results for each sequence (3.11a-3.11f) and overall (3.11g)

from slightly annoying to annoying. Ratings express that artifacts in C3, C4, C5 are slightly annoying or almost not perceptible. These sequences are spatially and temporally more complex than C2.

It appears that subjects spotted impairments more easily in low complexity contents.

The spatial and temporal complexity of C1 and C6 is lower than that of C3 and C4 but higher than that of C5. Nevertheless, C1 and C6 ratings express that more impairments are detected, even though the level of impairments is low. Those contents contain large homogeneous areas, likely to produce block effects.

These results are characteristic of compression schemes evaluations: large homogeneous areas are prone to block effects, hence are challenging compression algorithm.

Overall, impairments are assessed as not or slightly annoying as shown in Figure 3.11g. This is a shred of evidence that encoded streams are perceptually highly faithful to their reference. The compression solutions under test are thus highly efficient.

OB2 and OB3 bit rates show the highest compression efficiency concerning faithfulness. Considering the CIs, their significant overlap does not allow any distinction of a system when compared to the others.

However, one can note the highest efficiency of CC3 and CC4 at OB3 compared to the other CC at the same OB. Similar observation appears for C1, C2, C6 contents, in addition to C5 for CC3 as well as C3 and C4 for CC4.

This result first emphasizes the high efficiency for SDR delivery of the DLBC codec. Second, the highest MOSs are reached by OB3CC3, OB2CC1, and OB3CC4, in descending order. This indicates that allocating between 4.8 and 5.1 Mbps to the BL seams to be a right balance.

Source	Sum Square	d.f.	Mean Square	F	p
OB	18.7	3	6.2	6.27	0.0003
CC	3.8	3	1.3	1.28	0.2801
Content	765.6	5	153.1	153.9	0
OB*CC	19.4	9	2.2	2.2	0.0215
OB*Content	25.9	15	1.7	1.7	0.0375
CC*Content	15.6	15	1	1	0.4011
OB*CC*Content	53.1	45	1.2	1.2	0.1869

Table 3.7: SDR ANOVA results

We ran an ANOVA and Bonferroni post-hoc multiple comparison tests [Armstrong 2014] to investigate whether there are significant statistical differences between evaluated stimuli.

Table 3.7 presents the results of the 3-way ANOVA. We find out that OB4CC3 and OB4CC2 are significantly lower than OB2CC1, OB3CC3, and OB3CC4 ($p < 0.01$). However, there is no statistically significant difference between CCs or all stimuli (combination of OBs, CCs and contents) ($p = 0.28$ & $p = 0.19$, respectively).

The non-significance of differences between CCs demonstrates the similarity of SLBC and the variants of DLBC codecs regarding their performance for SDR delivery. Finally, our results stress that systems allocating between 4.8 to 5.1 Mbps to the BL with an OB of 6 Mbps reached the highest performance.

3.2.3.2 HDR

The same analysis than SDR was conducted on HDR results. Figure 3.12 shows the obtained MOSs and corresponding 95% CIs.

There is a trend that fewer impairments are perceivable at higher bit rates. This makes sense as the more bandwidth is available, the more information can be sent; therefore the more refined and faithful is the delivered content.

However, the fact that CC5 (SLNBC) and CC2-4 (DLBC) achieve similar perceptual quality is surprising. It shows the DLBC codec high-quality delivery of HDR, but more importantly, evaluates the codec performance as good as uncompromised HDR delivery.

In details, there are specific behaviors worth discussing. In C2 and C3, CC1 evaluation is roughly the same whereas the behavior of other systems drastically changes. The under-performance of CC2-4, but especially the poor efficiency of CC5 in C2 is surprising, as this CC conveys a pure HDR stream. Hence, we attribute the observed behavior to the low complexity of C2.

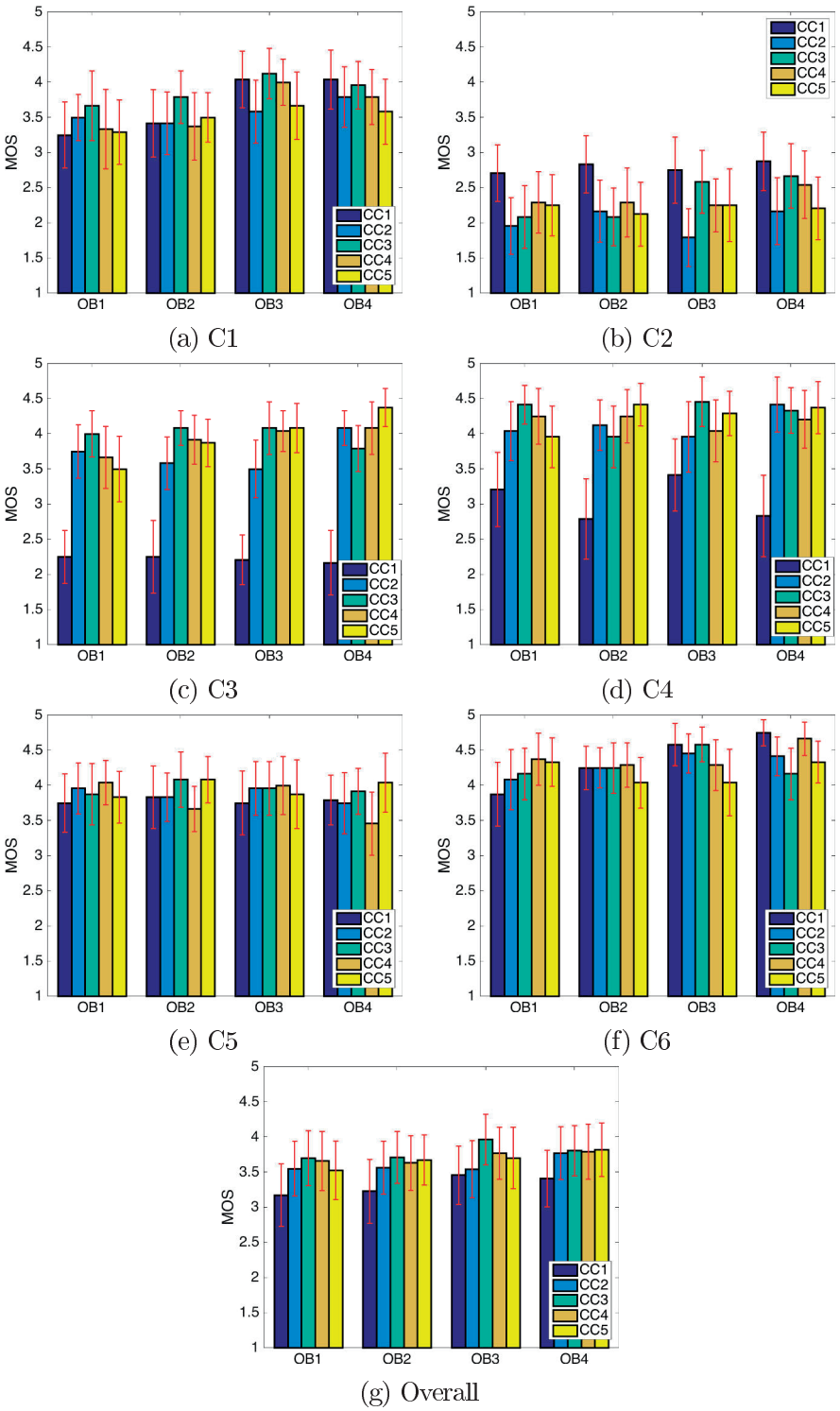
The high quality of CC2-5 and the poor performance of CC1 in C3 and C4 demonstrate the benefit of transmitting the residual errors of HDR prediction in the EL. It also stresses that HDR prediction is not sufficient for HDR delivery, especially for content with high spatial and temporal complexity.

The results of C1, C5, and C6 are similar: all systems appear equivalent and reach a high-quality delivery. This would indicate that for such contents (sport-events broadcasting) DLBC and SLNBC are efficiently codecs, regardless of parameter setting.

CC3, the system allocating 15% of the overall bit rate to its EL, achieves highest efficiencies for all contents and OBs in overall (see Figure 3.12g).

Results of the 3-way ANOVA are presented in Table 3.8. We observe that quality is content-, OBs- and CCs-dependent ($p < 0.01$). However, all stimuli are coming from the same distribution as our results failed to reject the H0 hypothesis (ratings come from the same distribution) ($p = 0.47$). Therefore, codec settings with specific bit rate allocation do impact the perceptual quality of delivered streams significantly.

HDR results indicate that at least a CC is significantly different from other systems. The identification of this difference is observed through a Bonferroni post-



(a) C1 (b) C2

(c) C3 (d) C4

(e) C5 (f) C6

(g) Overall

Figure 3.12: MOSs and 95% CIs for HDR results for each sequence (3.12a-3.12f) and overall (3.12g)

hoc tests.

No difference is statistically significant between the systems in sequences C1, C2, C5, and C6. In C3 and C4 and overall, the only statistically significant difference is that between CC1 and CC3. We can deduce three outcomes from these results.

- CC1 is the least performing system and is statistically different from at least another system in C3 and C4. This result advocates against the use of the SLBC.
- The non-significant differences between CC2-5 show that DLBC and SLNBC codecs perform similarly.
- To further discriminate CC2-5, we discuss the fact that the only system that differs significantly from CC1 is CC3. CC3 should then be favored as HDR compression solution.

Source	Sum Square	d.f.	Mean Square	F	Prob>F
OB	19.76	3	6.587	7.26	0.0001
CC	77.55	4	19.388	21.38	0
Content	1108.25	5	221.65	244.42	0
OB*CC	9.15	12	0.762	0.84	0.6081
OB*Content	18.24	15	1.216	1.34	0.1684
CC*Content	292.36	20	14.618	16.12	0
OB*CC*Content	54.54	60	0.909	1	0.4713

Table 3.8: HDR ANOVA results

3.2.3.3 Combination of SDR and HDR analyses

We summarize here and combine the findings of SDR and HDR analyses.

In the SDR analysis, even though all CCs are similarly performing, it has been observed that systems dedicating 4.8 to 5.1 Mbps to the BL and an OB of 6 Mbps achieved the highest perceptual quality. HDR analyses indicated that the SLBC coded should be prevented for HDR delivery, whereas DLBC and SLNBC codecs enable high quality HDR delivery. Besides, if a system should be favored, it would be CC3.

The combination of SDR and HDR analyses supports the CC3 at an OB of 6 Mbps as high efficiency compression solution. Indeed, this system fulfills all the constraints expressed above.

3.2.4 Conclusion

In this study, we compared a DLBC solution to two single-layer codecs realizing uncompromised SDR and HDR, namely SLBC and SLNBC. Our findings draw

recommendations for HDR delivery about (1) bandwidth allocation, (2) whether broadcasters should envision backward-compatible solutions or not, and (3) which approach is the most suitable for SDR, and HDR delivery.

The codecs mentioned previously have been assessed under several conditions. Four overall bit rates, representative of today's UHD bandwidth constraints, three enhancement layers configurations investigating the parameter setting for the DLBC codec, and six contents were the variables of our experiments. SDR and HDR evaluations have been conducted in parallel to consider both legacy and future delivery contexts. At least 20 grades per stimuli were collected from two gender-balanced populations of about 22.5 years of age on average.

Even though SLBC and DLBC codecs have a similar high efficiency for SDR delivery, our results advocate for the allocation of between 4.8 and 5.1 Mbps to the BL for all systems. With regards to the HDR delivery, the SLBC performs similarly of significantly less efficiently than the two other approaches. The use of this codec for HDR should thus be prevented.

The DLBC and SLNBC solutions show no significant differences. To further discriminate those two solutions, we looked deeper into which approach is statistically different from the SLBC solution. The scenario stressed in the process is the DLBC using 6 Mbps of overall bit rate and allocating 15% of this bit rate to the EL (resulting in 5.1 Mbps dedicated to BL).

By construction, the SLNBC delivers uncompromised HDR. Hence, it is surprising that this system did not outperform the others significantly, for the transmission of HDR streams.

Overall, the DLBC approach delivers perceptually uncompromised SDR and HDR and reaches the highest visual quality when being tuned with an overall bit rate of 6 Mbps, 15% of which is allocated to the EL. This finding is in line with the constraint of ensuring a smooth transition to HDR TV for broadcasting services. Besides, it provides a recommendation about fix bandwidth settings for HDR delivery.

The conducted investigation mainly focused on multimedia transmission from a perceptual perspective. A study on codec complexity (e.g., memory resources, computation complexity, especially in the decoder) or on the robustness of the streams to packet loss may be of interest to establish our recommendations.

This study has been published by the IEEE Transactions on Broadcasting, in the special issue on Quality of Experience for Advanced Broadcast Services [Perrin 2018b].

3.3 Conclusion

The HDR and WCG are highly promising formats, which were evaluated to be a revolution similar to the introduction of color in media contents.

Nowadays, tremendous research is dedicated to defining this imaging (e.g., representation of content, compression schemes or transfer functions to match the DR

of the display). Several solutions have been standardized regarding TFs. Indeed, Dolby and BBC proposition of PQ and HLG functions have been accepted by the community. However, many other issues are not tackled yet.

We have conducted two studies necessary for the development of the HDR imaging ecosystem. In the first one, we addressed the comparison of legacy color space (Y'CbCr) with ICtCp, designed to accurately and efficiently represent HDR. We also evaluated the interest to include a representation optimization, namely reshaper, before compression in delivery pipelines. In the second work, we provided an analysis of three compression approaches concerning the delivered perceptual quality for SDR and HDR streams.

With our research outcomes, we meant to provide recommendations for the design of the HDR ecosystems regarding content representation and delivery. Those two points are especially meaningful for broadcasters.

Our conclusions are the following, and are only dealing with perceptual quality concerns.

- Overall, there is no reason to favor Y'CbCr or ICtCp. If the design system is meant for low bandwidth constraints, then we recommend the use of ICtCp color model. On the contrary, if broadcasters are creating a high-quality service with high bandwidth allocations, Y'CbCr must be used.

- The reshaper mapping function showed promising results. Overall, similar or improved perceptual quality is reached with the reshaper when settings are tuned to achieve a bit rate saving up to 10%. However, setting the reshaper to 80% of the initial bit rate impacts the perceptual quality too much.

  We recommend the use of the reshaper in any delivery pipeline. We also want to stress out that the reshaper is even more efficient combined with the use of the ICtCp color space.

- Our results showed that the use of a DLBC system realizes uncompromised SDR (perceptual quality similar as SLBC) as well as uncompromised HDR (perceptual quality similar as SLNBC). We thus recommend this compression schemes for HDR and WCG broadcasting. This outcome is highly important as the transition from SDR to HDR can be operated smoothly.

The solutions we have evaluated were already assessed objectively and other subjective studies, conducted after our own, confirmed our conclusions. The guidelines provided above are thus reliable and are applicable for future HDR systems.

Let us now focus on the evaluation methodology approach.

Designing a test relies on taking into account a lot of parameters, especially design details that can improve the test accuracy and robustness. However, those

decisions are subtle and can result in a trade-off between various effects of the considered design.

It is worth mentioning that the performed tests described in this chapter are quality assessments and assume ideal network conditions (e.g., no delay, jitter or packet loss impairments) for content transmission. This enabled the evaluation of only one quality measure such as preference or perceived amount of impairments.

Conducted tests have implemented two SbS pairwise comparison test methodologies. In the first experiment, ICtCp and Y'CbCr provide highly realistic representation, and so the reshaper (in a lesser extent when saving bitrate). Conducting an SS test wouldn't offer a comparison of the systems. Concerning the second test, the aim clearly was a comparison of compressed contents with raw signals to evaluate how reliable is the delivery.

Accordingly to the purpose of the evaluation, the first work investigated the preferences of subjects over two stimuli. The second gathered levels of impairment detected between a stimulus and its reference.

Besides, both studies implemented only partial pair comparisons. A first concern is that some comparisons do not exhibit information interesting for the analysis. They are thus not included in the evaluation. A second concern is the importance of an experiment to be of acceptable duration (about one hour and a half) to be valid.

Methodologies to evaluate videos may include repetition. However, repetitions increase the learning effect in subjects and the test duration. Thus, whether to include repetitions is a design parameter which could strengthen or weaken test validity. This explains that video stimuli repetition should only be used with hardly noticeable impairments or difference across stimuli. This is critical in these scenarios as results may be inconclusive or show no difference when there is.

It is interesting to note that, depending on the gathered information, different processing should be applied. For instance, there is no possibility of giving a MOS to a preference grade; or there is no analysis possible if a pair-comparison has not been included in the test design (e.g., in the first experiment, including P10 vs. P30). This means that it is imperative to define which information we want to analyze before the subjective test design. This fact will be particularly crucial for the evaluation of QoE.

The experience I gained during these two works had a considerable impact on the definition and construction of my strategy regarding QoE evaluations.

Based on the subjective test recommendations of standardization committees, it has been emphasized that a specific questionnaire should be developed for QoE assessment. Despite the introduction of the QoE definition in ITU-T Rec. P.10/G.100 in 2017, no recommendation about question(s) to ask about which QoE influence factors is provided.

If perceptual quality is understood by subjects and can be assessed with one score (and this is debatable), asking people "What is the quality of your experience" is too

vague. In what sense is an experience of quality? More importantly, discrepancies between subjects are not bringing information: what is the indisputable reason for the difference of one level in the scale in terms of QoE? What is meant by quality of experience? etc.

In the literature, influence factors and related questions are defined on the case-by-case basis. Merging influence factors of all multimedia technologies may not be representative of QoE in all contexts. For instance, asking about depth cues for 360° or holographic contents make sense, while it is not a significant factor to assess HDR or high frame rate content. Besides, the constraint of reasonable subjective test duration prevents to ask about all influence factors that can be identified for multimedia experiences.

Consequently, it is important to identify which influence factors are representative of QoE in a specific context. Also, one must define what to ask about these parameters or overall experience. Recommendations lack guidelines regarding those two crucial steps in subjective evaluation design.

We decided to address the challenge to indicate best practices regarding the identification of influence factors and the way to evaluate them for most immersive multimedia. I insist on specifying that we consider "most" technologies, as multimedia experiences with interactions from the consumer are not studied in this thesis.

In the two following chapters 4 and 5, we present our works conducted on the identification of influence factors and how to evaluate them.

First, HDR 360° imaging presented an excellent opportunity for examining the influence factors of a format which is both immersive and realistic.

Then, with regards to what characteristics of experience or influence factors to assess, the Qualinet definition of QoE indicates the expectations and the enjoyment regarding the user personality and current state. I focused on the assessment of expectations for QoE evaluations in the second chapter.

Chapter 4

Contextual- and system-based influence factors for HDR 360 ° imaging

Contents

The performed studies on quality evaluation for HDR and WCG prescribe several points to take into account in QoE assessment. The first important one is the fact that asking a single question is probably not precise enough and does not indicate the strengths and weaknesses of the evaluated solution from the viewers' perspective. Multi-variate analyses would cope better with the multi-dimensionality of QoE.

An analogy with Personal Navigation Assistant (PNA) service is here used to illustrate issues to tackle when performing QoE evaluation. PNA services help when going from point A to point B. The service needs at least a geolocation system (such as a Global Positioning System (GPS)) and map information. Let PNA1 be a simple system having the only requirement to indicate any navigation path from point A

to point B. Let us assume that PNA1 does not have accurate mapmaking (e.g., lacks traffic directions). The recommended path proposing navigation from A to B fulfills service purpose. A first PNA1 marginal weakness is the possibility to lead the driver in illegal conduct, by taking, for example, a one way in the wrong direction. Also, PNA1 fails to consider user needs (e.g., fastest, shortest or most economical navigation). It investigates neither user equipment constraints (e.g., the height of the car, its consumption and fuel level or motorway toll requirement) nor real-time or predicted traffic. This path is representative of subjective evaluation using quality as a single variable. We used a similar process in the two previous experiments exposed, in Chapter 3, where gathered information was preference or impairment level. Even though it provided the required information to our research question, it did not give further knowledge about behavior causes and if users were satisfied.

Let PNA2 be a smarter navigation system which proposes a way from A to B, according to the desired type of path, user's equipment and accurate network information (from a detailed map with navigation directions to traffic prediction). This service is entirely meeting the users' requirements. However, its design defines its limits: the service focuses on material and equipment constraints. Despite being a high-quality service, consumers' additional needs are not covered, such as users' demand to stop for lunch or desire to do sightseeing on the way. This example represents QoS, which evaluates technical infrastructure, user material, and multimedia content. The inclusion of user-centric constraints to the service will allow the creation of an even higher quality service fulfilling consumers' needs.

It should be recalled that service quality does not always imply users' satisfaction. Let us consider a simple example: watching a video on a bus during the rush hour. The noisy and uncomfortable environment prevents the users' satisfaction regarding the multimedia experience, even with high-quality content and delivery. Using the same service at home, with a proper lighting environment, in a calm room with a comfortable sofa is thus answering user's needs and may exceed consumers' expectations which ultimately lead to satisfaction. Another consideration expresses the substantial impact of expectations on experiences. Users are aware of environmental constraints which are included in their evaluation. That is a less high-quality experience may be more accepted in a noisy and uncomfortable environment. Thus, to evaluate QoE in light of users' context and expectations is the researcher's ultimate goal. This leads to the last more sophisticated PNA service.

PNA3 service has all PNA2 features and also includes a user profile, based on pre-defined questions investigating consumer preferences, expectations, previous experiences and possibly feedback reports. This service is then elaborated accordingly to the consumer and his context. The system improves future experiences by implementing an evolving recommendation system. The inclusion of users' context targets a dedicated service which improves user's satisfaction. QoE corresponds to this latter scenario, in which expectations, previous experiences, and contextual information of users result in a more user-centric service. QoE evaluation is a more broad measure than quality and QoS for it considers intrinsic service characteristics:

quality, network performance, specificities and equipment of users.

The introduced analogy points out improvements that should be implemented to achieve QoE: (1) investigating the user profile: state of mind, expectations, prior knowledge, and contextual information, (2) including subjects' evaluation regarding specific service features (3) asking more questions about experiences. These steps lead to evaluating the quality along with essential influencing factors to further understand QoE, its construction, and possible improvement.

In this chapter, we address points 2 and 3. point 1 is left for further research presented in Chapter 5.

In details, this chapter presents an exploration of influence factors extraction in the context of 360° HDR formats. We combined both HDR and omnidirectional technologies to realize rich experiences. HDR provides realistic and faithful representations, while omnidirectional contents are immersive and engaging. This work performs multi-variate QoE analyses, in studying several quality aspects during stimuli assessment.

4.1 Related works on HDR and omnidirectional contents

HDR is becoming a popular type of visual content in both professional and consumer markets. With advances in capture and display technologies, the HDR imaging pipeline can potentially transmit the full range of light information of a scene. This characteristic makes it overcome several physical and perceptual limitations of SDR imaging systems. Over the last three decades, HDR content processing, coding, and quality assessment received particular attention in imaging.

Omnidirectional or 360° imaging is an emerging format in the area of immersive multimedia. 360° cameras allow the capture of the entire field of view that covers a full sphere. This content is usually visualized using near-eye HMDs so that the viewer is free to change the direction of his/her sight across the omnidirectional scene. However, to the best of our knowledge, current omnidirectional systems are based on SDR imaging systems, i.e., contents are captured with a single exposure and viewed on display with limited dynamic range. This makes current 360° imaging subject to the limitation of SDR systems especially from loss of information due to under- and over-exposure. This could lead to a less immersive experience. HDR imaging is a potential technology to resolve this issue. Therefore, in this study, in addition of how to extract influence factors for QoE evaluation, we set out to identify whether there is a qualitative benefit of applying existing HDR imaging workflows to the current SDR 360° imaging pipeline.

A first challenge to overcome is 360° HDR absence of display. Current 360° rendering systems are coping with legacy SDR representations. This forces to tone-map 360° HDR contents to be compliant with omnidirectional displays. Any 360° HDR content will thus be tone-mapped to match SDR formats.

A second challenge is the lack of contents dedicated to 360° HDR evaluations. On one hand, 360° HDR imaging is in its infancy. This explains the reduced number

of publicly available content. On the other hand, the available professional 360° HDR contents are not necessarily compliant with the evaluations to be conducted. According to features of both technologies, captured scenes to be used for QoE appraisal should include various illumination areas, with definite highlights and deep shadows while containing rich scenes information in different view directions. Artifacts due to the HDR acquisition, especially ghosting effect, blur, and noise resulting from the fusion of multi-exposure images must be avoided. 360° camera-consumers are not able to capture the full dynamic range of the scene. Nevertheless, they allow different exposure captures. Ultimately, merged exposure images recreate the captured environment in HDR 360° image.

We faced a third issue: the lack of pair comparison methodologies for 360° imaging and testbed. Our intent was to compare original SDR contents to 360° HDR tone-mapped contents (i.e., SDR variants). We meant to highlight the interest in using HDR imaging in combination with 360° formats. Additionally, we aimed to carry out an in-depth analysis of the strengths and weaknesses of current 2D tone-mapping operators for omnidirectional contents. We wanted to draw guidelines regarding new 360°-based tone-mapping operator creation. The subjective evaluation design was mostly focused on this purpose.

The following section introduces the work performed in [Perrin 2017a] investigating 360° HDR imaging impact on users' QoE, while tackling previously mentioned issues. The study results from a fruitful collaboration with the Advanced Media Coding (AMC) lab of Institute of Research and Technology (IRT) b<>com.

4.2 HDR 360° consumer-camera contents

The purpose of this study was to give an overview of advantages and drawbacks of current HDR 360° imaging when compared to SDR omnidirectional contents. For this matter, we have decided to create a camera-consumer dataset dedicated to evaluating HDR 360° images subjectively.

Regarding current acquisition systems for HDR contents, Nikon, Canon, Arri, Panasonic, and Sony, to name but a few, have created high-end DSLR cameras or specialized professional video cameras able to capture HDR. However, most consumer cameras only capture 8-bit images. They are far from reproducing the scene full luminance range. To overcome this limitation, one can merge multiple exposure images to gather more details in darkest and brightest areas.

Hence, we focused on omnidirectional systems enabling the capture of multi-exposure contents. Such systems are already used in various applications of omnidirectional HDR images.

A widely spread application is the Image Based Lighting (IBL). 360° HDR contents provide the environment information to recreate the illumination of 3D objects [Debevec 2008]. Capturing contents for IBL is often a cumbersome and expensive process. Specialized equipment is necessary for such capture: professional DSLR cameras, mobile rigs, stitching software, and dedicated processing. To overcome

such constraints, we chose to position this study in the context of real-world scenarios using consumer 360° cameras and consumer HMDs, and recreate HDR images from multi-exposure image sets.

The Ricoh Theta S 360° panoramic pocket camera has been selected for capturing the contents. This consumer camera produces omnidirectional images in 8 bit, 5376x2688 (14 Megapixel (MP)), and RGB. The capture system is combining two 190-degree field of view cameras. Seamless stitching is applied to generate a single 360-by-360 degree sphere image.

Regarding the camera specifications, Ricoh theta S allows shutter speeds ranging from 1/6400 to 1/8 seconds, International Standards Organization (ISO) sensitivities ranging from 100 to 1 600 with a wide lens's aperture (F2.0) and a focal length of 1.31. An automatic exposure control sets shutter speed and aperture to capture a scene with various Exposure Values (EVs), ranging from -2 to 2 by steps of a third. Captured spherical contents are compressed and represented as equirectangular projection before being stored. Raw contents are compressed with the JPEG codec, with a quality factor of 95. The stability of the camera and the alignment of multi-exposure pictures were ensured during the acquisition process thanks to a tripod, when not exploiting the environment as camera holder.

The Ricoh Theta S is controllable through Bluetooth connection with any handheld device through the Ricoh Theta S application, enabling the photographer not to be on every acquired picture. However, the short sensing range (about 3 meters) forces cameraman ingenuity, such as capturing scenes with at least one object on the forefront (to hide behind). Specific attention was drawn to having as few as possible multi-exposure bracketing impairments such as ghosting effects and noise.

With regards to acquisition and bracketing artifacts, the presence of individuals in a scene was challenging in several aspects. Any movement from an individual between scene acquisitions will result in ghosting effect when merging multi-exposure pictures. The same effect may appear during a long exposure image capture, for all moving objects (including clouds or cars). This explains why most acquired contents did not contain moving objects.

The 360° HDR dataset aims to give the opportunity to evaluate an extensive collection of scenes. This implies to collect a wide variety of shooting conditions and environments, e.g., with various overall contrast and brightness, diffuse and specular reflections, deep shadows and bright highlights. Specific attention has been paid to capturing scenes aiming specifically at omnidirectional and HDR evaluations. For instance, there should be interesting elements in several directions of the spherical scene (at least two different viewports on the HMD). This addresses the requirement of having 360° contents rich in structures spatially. Different under- and over-exposed areas, with specular or diffuse reflexions, are compulsory to capture scenes with high dynamic range and to challenge TMOs.

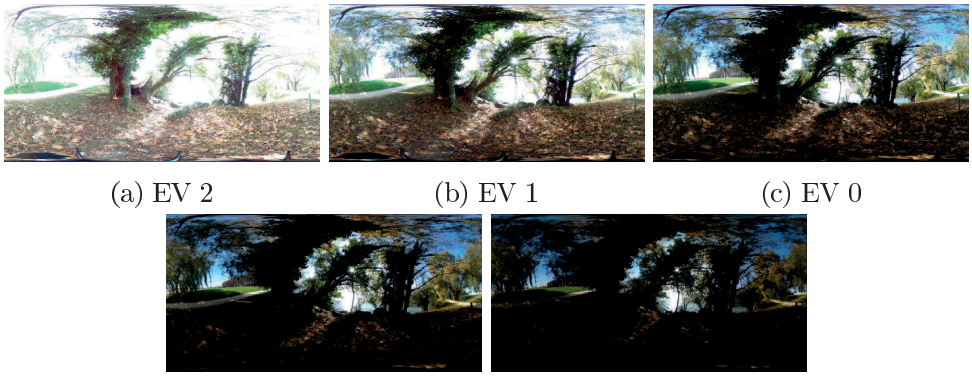
The selected acquisition system and procedure enable only the capture of natural scenes, in different contexts. Night contents are especially indicated for HDR imaging as they present intensely bright and dark areas. Also, a common problem faced with usual SDR sensors is that one cannot capture details in both illuminated

areas and shadows. So, outdoor scenes in daylight, with clear sky and shadows areas are valuable. Indoor views are relevant in dark rooms with windows through which one can see a bright environment.

Eventually, we captured 43 scenes. They are classified into three content types, namely *Indoor* (9 contents), *Outdoor* (24 contents) and *Night* (10 contents). Scenes' dynamic range is covered by five perfectly aligned multi-exposures pictures, with EVs ranging from -2 to 2 by step of 1.

After capturing five different exposures, there are many ways to generate an HDR image. However, based on the study of Akyuz et al. [Akyüz 2013], state-of-the-art reconstruction algorithms are not statistically better than each other concerning accuracy. Hence, we simply apply the classical method proposed by Debevec and Malik [Debevec 1998] using a triangular weighting function. This is one of the first methods that combined multiple exposures to produce an HDR image whose pixel values are proportional to the physical radiance values of the light captured.

One concern of the dataset is that images dynamic ranges vary from 7,59 to 12,27 f-stops, while profession HDR contents aim at about 16 f-stops. We believe this limitation to be due to the Ricoh Theta S sensor at extreme EVs. One can observe over-exposed and under-exposed regions in the reconstructed linear HDR scenes which are atypical in HDR imaging. Nevertheless, this dataset remains useful as it is the best possible dynamic range that can be captured using current consumer omnidirectional cameras, which is the addressed use-case.



(a) EV 2 (b) EV 1 (c) EV 0

(d) EV -1 (e) EV -2

Figure 4.1: Multi-exposures of the *Lake* content

4.3 Tone mapping operators

It has been previously mentioned that current HMDs are not able to render HDR contents. Such content visualization is possible if HMDs are provided with SDR contents, which are mapping variants of the HDR representation. This indicates the need to apply exposure fusion or state-of-the-art tone-mapping operators on the above-described dataset.

The exposure fusion workflow is based on the 'HDR mode' we find on mobile devices. This process is akin to the work by Mertens et al. [Mertens 2009] which consists of blending multiple exposures into a single image by skipping the HDR reconstruction step. TMOs are applied on HDR reconstructions coming from multi-exposure contents.

The TMOs considered in this study are based on 2D imaging, mostly because very little has been done in omnidirectional HDR tone mapping. Amongst the few works in literature, Hausner and Stamminger [Hausner 2004] present an extension of the Photographic TMO which adapts to the user's central field of view by using tracking information. Yang et al. [Yang 2012] and Mikamo et al. [Mikamo 2016] display two different tone-mapped images of the same HDR input image on each eye of a binocular display (such as an HMD). When seen through a binocular display, the combination of images is richer and more detailed than single tone-mapped versions. Yu [Yu 2015a] suggested a solution built upon a global operator. He has implemented a real-time operator, mapping the dynamic range on the current viewport displayed on the HMD.

These works are potential directions for research in omnidirectional HDR imaging for HMDs. However, none of these solutions are widely used nor accepted by the research community yet. Hence, we considered them as beyond the scope of this study. We only studied existing TMOs for 2D HDR workflows for qualitative assessment against SDR 360° workflows.

For this study, we have chosen commonly used TMOs that have been freely available on commercial software [Mantiuk 2007] for many years and are fundamental to HDR imaging. For this purpose, we used the pfstools software [1]. TMOs and exposure fusion method are described after this.

- The luminance nonlinearity introduced by many imaging devices can often be described as a simple gamma correction function $f(I) = I^{\lambda}$, where I$\in$ [0 1] is the normalized image pixel intensity. An EOTF is defined in [ITU 2007a] following a gamma curve. [ITU 2013a] encourages to use such functions to adjust contents to the display full dynamic range. We applied this widely used mapping function on the normalized HDR radiance. The λ value is set to 2.2, corresponding to common gamma correction value. This global operator is referred to it as **linear TMO**.

- **Photographic TMO** [Reinhard 2002] is a local operator inspired from two HDR photography techniques, zone system and dodging-and-burning [Johnson 2006]. Its principle is to reduce the dynamic range by averaging the luminance using the luminance response of neighbor pixels. This algorithm should be tuned specifically based on the considered scene. In our context, we used only the global components (the neighborhood is the entire image), turning the used implementation into a global operator.

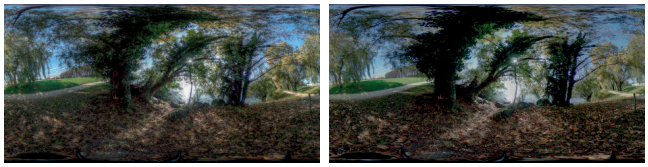
[1] http://pfstools.sourceforge.net

- [Mantiuk 2008] developed an elegant global operator that adapts content dynamic range to display capabilities. The operator is a piece-wise linear tone-curve which solves a minimization problem. This optimization aims to reduce the difference between HVS responses to original and rendered images. HVS responses are estimated through a contrast sensitivity function. This operator is named **Display adaptive TMO**.

- [Mantiuk 2006b] proposed a local algorithm that minimizes the distance between a set of contrast values (Gaussian pyramid levels), which specifies the desired contrast with the original image contrast. We call this operator the **detail preserving TMO**.

- **Exposure fusion** [Mertens 2009] is a local operator, which fusions bracketed exposure images. This technique reconstructs a high-quality SDR, without retrieving the HDR representation of the scene. The fusion is based on image Laplacian decomposition and a Gaussian pyramid of weight-maps. Weighted maps are generated depending on contrast, saturation, and well-exposedness criteria. This algorithm is not a TMO as the HDR image is never required in this method.

(a) Linear TMO (b) Photographic TMO (c) Display adaptive TMO



(d) Detail preserving TMO (e) Exposure fusion

Figure 4.2: Tone-mapped and exposure fusion pictures of the *Lake* content

For all the TMOs, we followed the recommended settings suggested by the authors, except for the display adaptive TMO where we applied the settings of surround lightings specific to our HMD. It must be noted that all mapping operations were done on contents represented as equirectangular projection.

The local characteristic of both *Detail preserving* and *exposure fusion* operators implies that the left and right sides of the equirectangular projected image will have different processing. As a consequence, a vertical line will appear on the HMD. This clearly indicates the stitching region of the picture, which may affect the user's judgment during the subjective study. We resolved this issue by doing additional processing for local operators. This involved three steps: 1) concatenating the columns of the left part of the projection to the right part and vice versa, 2)

applying the local operator and 3) removing the additional columns and retrieving the locally processed image.

4.4 HDR 360° Dataset

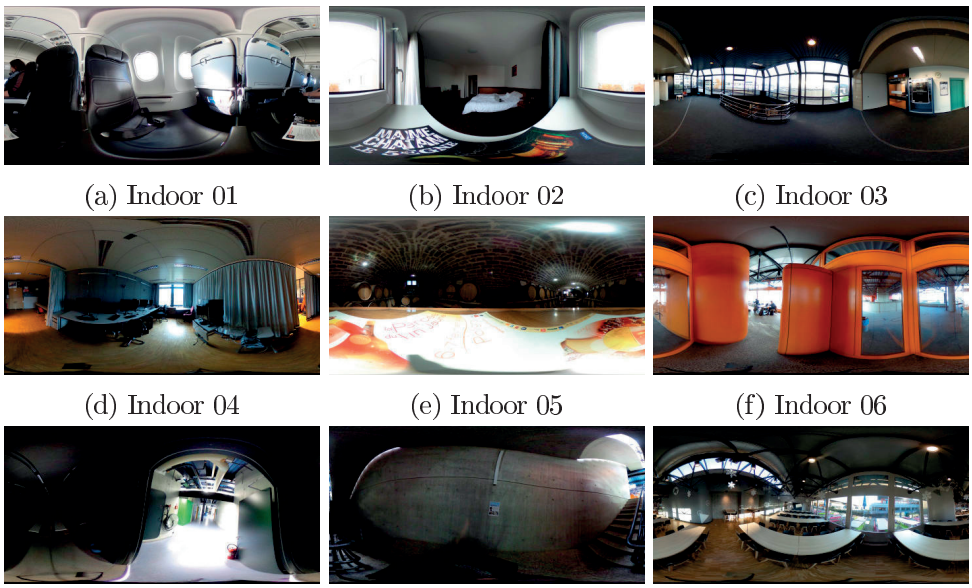
The 360° HDR dataset contains 43 scenes. Each scene is represented by five multi-exposure pictures, an HDR reconstruction as well as tone-mapped and multi-exposure fusion variants of the scene.

As an example, Figure 4.1 presents the five multi-exposure images of the scene *Lake*. A large part of the full range of luminance of the scene is thus captured. More precisely, the lowest EVs picked up the details in bright areas while the highest EVs acquired the details in dark areas.

Figure 4.2 illustrates the SDR images resulting from the four TMOs and Exposure fusion described in Section 4.3. Finally, Figures 4.3, 4.4, 4.5 present the mid-exposure equirectangular projection images of the indoor, night and outdoor scenes, respectively.

The introduced dataset is publicly available[2]. It should be noted that even though five exposures enable the capture of a large part of the dynamic range of a scene, the sensor properties of the Ricoh Theta S consumer camera are limited. Other limitations, such as movement between the multi-exposure captures or non-perfectly aligned pictures, lead to artifacts such as noise or ghosting effect. The scene captures showing such artifacts have not been included in the test, and so only 33 HDR reconstructions are made available. This choice shows our intent not to create artificially the artifacts that will discriminate the comparison between SDR and HDR contents. The multi-exposures of these impaired contents are available in the dataset, for future study on HDR bracketing artifacts in omnidirectional imaging for instance, but no HDR reconstruction are made available. Along with the dataset, information gathered during the subjective evaluation, described in the following, are provided.

[2] http://mmspg.epfl.ch/360hdr-consumercamera



(a) Indoor 01 (b) Indoor 02 (c) Indoor 03

(d) Indoor 04 (e) Indoor 05 (f) Indoor 06

(g) Indoor 07 (h) Indoor 08 (i) Indoor 09

Figure 4.3: Mid-exposure of every indoor scene included in the HDR 360° dataset

(a) Night 01 (b) Night 02 (c) Night 03

(d) Night 04 (e) Night 05 (f) Night 06

(g) Night 07 (h) Night 08 (i) Night 09

(j) Night 10

Figure 4.4: Mid-exposure of every night scene included in the HDR 360° dataset

(a) Outdoor 01 (b) Outdoor 02 (c) Outdoor 03

(d) Outdoor 04 (e) Outdoor 05 (f) Outdoor 06

(g) Outdoor 07 (h) Outdoor 08 (i) Outdoor 09

(j) Outdoor 10 (k) Outdoor 11 (l) Outdoor 12

(m) Outdoor 13 (n) Outdoor 14 (o) Outdoor 15

(p) Outdoor 16 (q) Outdoor 17 (r) Outdoor 18

(s) Outdoor 19 (t) Outdoor 20 (u) Outdoor 21

(v) Outdoor 22 (w) Outdoor 23 (x) Outdoor 24

Figure 4.5: Mid-exposure of every outdoor scene included in the HDR 360° dataset.

4.5 Content selection

To conduct a reliable subjective study test images should be containing a large variety of scenes. In HDR imaging, we often consider scenes with extreme variations in lighting levels. Keeping this in mind, we have selected eight different HDR images (shown in Figure 4.6 along with log2 histograms) which are representative of a diverse set of contents with indoor, outdoor and night scenes, under varying lighting conditions.

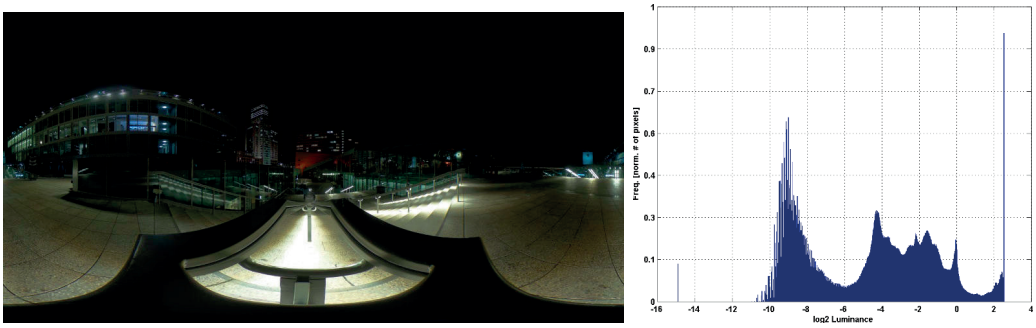
From the 43 contents of the created database, 11 were shortlisted based on a pilot study conducted by expert viewers. Images with minimal artifacts (e.g., ghosting, noise and over/under exposure) were selected for both the single exposure and the processed images. Table 4.1 summarizes the selected contents and their characteristics.

We considered a number of different global statistics over the entire sphere for each image. The statistics for the experimental images are DR (see Equation 2.5), key value (see Equation 2.7) and SI (see Equation 2.3) that can be found in Table 4.1.

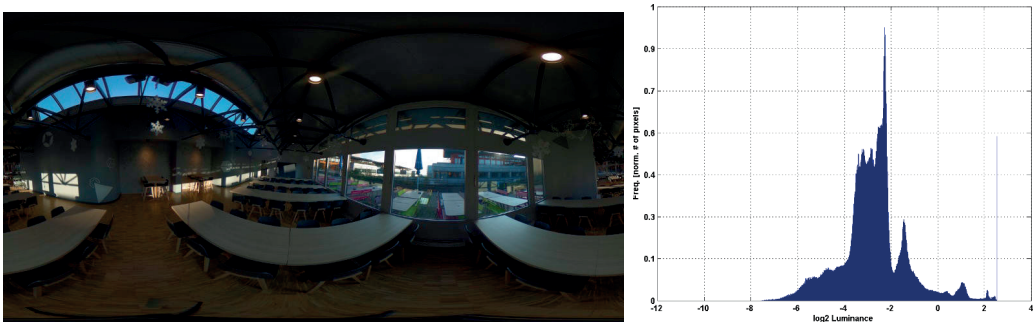
The various statistics do not show a clear indication that the chosen images contain a large variety of DR and spatial complexity. This is not to say that content is not diverse. To put into evidence the diversity of the selected omnidirectional content, we were inspired from the work in video tone mapping of Boitard et al. [Boitard 2012]. If we consider the viewport as video frame and the movement of the observer as the motion of the camera, it is possible to study key value variations as the user navigates across the 360° image. To simulate this, we extract the viewports along the center of the omnidirectional images and calculate the key value per viewport as seen in Figure 4.7. The brightness not only varies between the pictures but differs significantly within the entire sphere in each image.

		Reference in database		Exposure time		Statistics		
Set	Name	Content type	Reference	min	max	DR	Key value	SI
Training Set	Trail	Outdoor	01	1/6400	1/400	7.92	0.46	0.21
	Bridge	Outdoor	05	1/6000	1/125	7.59	0.46	0.12
	Sculpture	Outdoor	21	1/4000	1/200	7.93	0.59	0.15
Test Set	Room	Indoor	02	1/200	1/30	9.88	0.54	0.10
	Lab	Indoor	04	1/30	1/8	9.70	0.54	0.14
	Cellar	Indoor	05	1/180	1/30	7.86	0.55	0.10
	Cafeteria	Indoor	09	1/1500	1/40	8.22	0.41	0.11
	Vase	Night	01	1/30	1/10	10.59	0.52	0.19
	Berlin	Night	02	1/60	1/8	12.27	0.42	0.26
	Lake	Outdoor	02	1/5000	1/160	7.97	0.44	0.23
	Rolex	Outdoor	23	1/6400	1/500	9.58	0.57	0.13

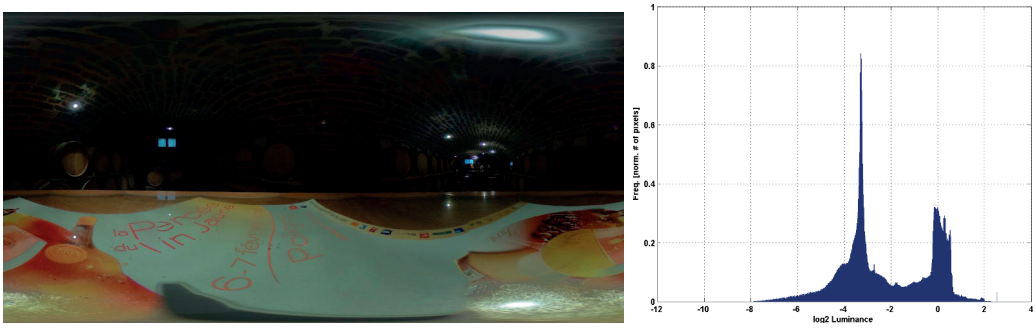
Table 4.1: Characteristics and global statistics over the entire image



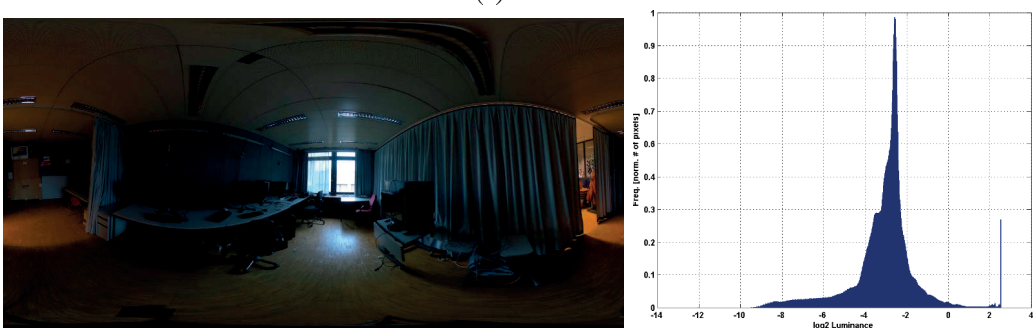
(a) Berlin



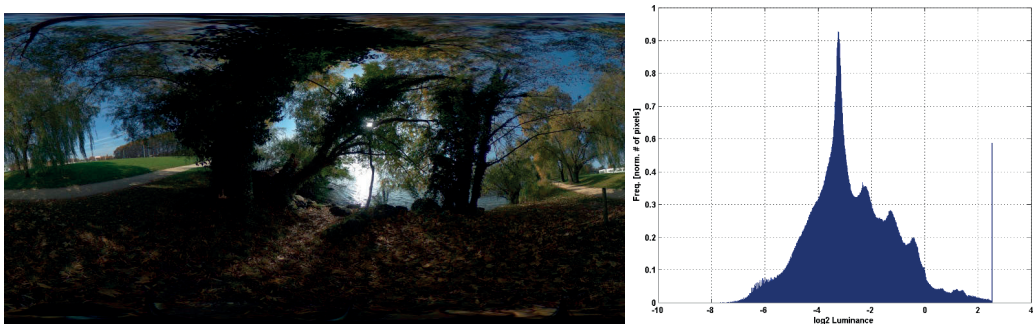
(b) Cafeteria



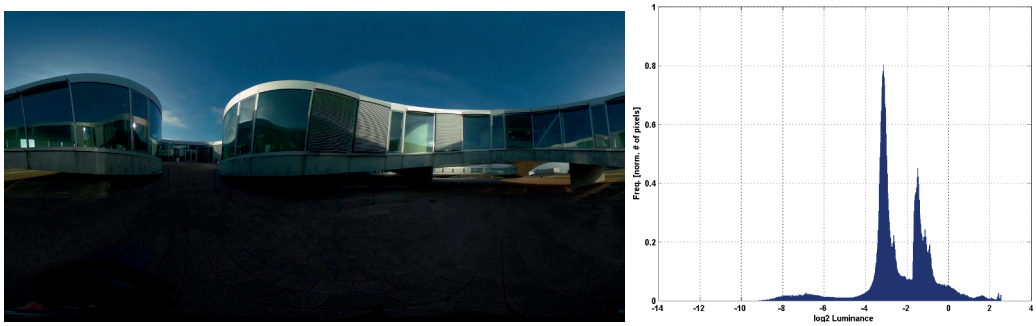
(c) Cellar



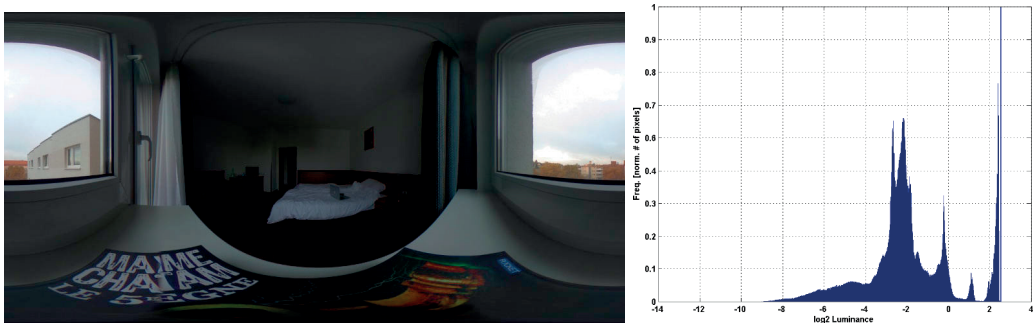
(d) Lab



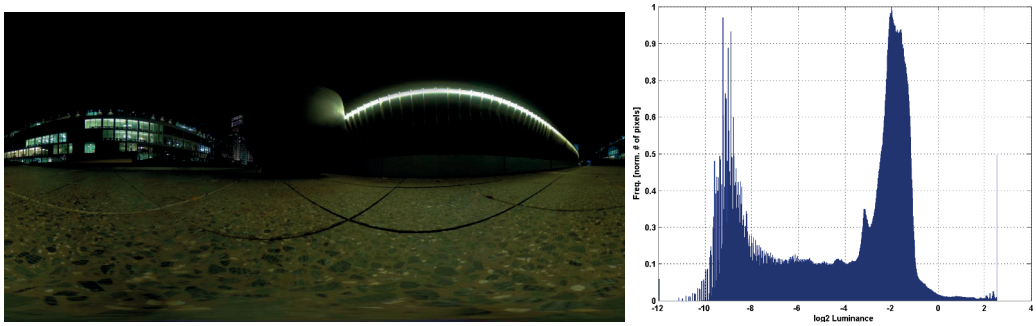
(e) Lake



(f) Rolex



(g) Room



(h) Vase

Figure 4.6: Equirectangular projections and histograms of test images. The linear TMO has been applied for each of the images above. Histograms are computed over the entire sphere of the omnidirectional image, showing the relative frequency of the $\log_2$ luminance of pixels.

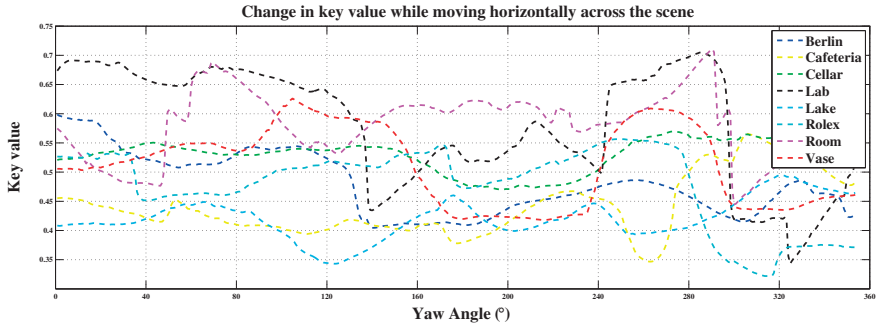


Figure 4.7: The variation in key value of the omnidirectional HDR content used in the experiment. The key value is calculated for a given viewport across the entire yaw angle of the scene while fixing the pitch and roll to 0°. This graph shows the content selected was diverse with varying luminance levels to challenge TMOs.

4.6 Equipment

Limited work exists on qualitative assessments of omnidirectional imaging. Authors of [Yu 2015b] proposed two objective metrics comparing different panoramic projections of omnidirectional videos. A quality metric, suggested in [Zakharchenko 2016], compares the Craster parabolic projection [Craster 1929] of different geometrical representations. Upenik et al. [Upenik 2016] introduced a testbed for single stimulus subjective evaluations of 360° images on HMD.

We build upon the testbed [Upenik 2016] to introduce a new dual stimulus evaluation method for omnidirectional imaging to conduct our experiments. This testbed originally enables test sessions by proceeding as follows: providing textual instructions, doing a training session, performing subjective evaluation using SS methodology, displaying a scoring menu and finally selecting the grade before continuing to the next stimulus. The testbed acquires information from every stimulus including the set of each subject's scores, the tracking of the direction of view and the duration of the stimulus visualization.

The direction of view is recorded by three coordinates, namely yaw, pitch, and roll, representing the angles formed with the normal, lateral and longitudinal axis, respectively. The frequency of acquisition is about 60 Hz, and the precision of timestamps is 10^{-7} seconds.

The equipment required to perform the evaluation is a hand-held device (e.g., iOS mobile) combined with an HMD. We have used two iPhones 6 and 6S, compatible with the testbed previously described. The devices are 4.7 inches diagonal with an HD resolution (326 ppi). Their maximum brightness is 500 cd/m^2, and their contrast ratio is 1400:1. The color space representation of those displays is full sRGB. There is no difference between the displays of both devices, and the followed calibration process was strictly the same.

To the best of our knowledge, there are no peak brightness recommendation or standard for HMDs. Therefore, to ensure the comfort of the subject as well as an optimal visualization setup, it was decided to set the display luminance to

100 cd/m^2. This decision is based on a few pilot tests as well as on recommended brightness settings for SDR TV in broadcasting services, the Rec. ITU-R BT.2035 [ITU 2013a]. The peak brightness was determined using a white screen and manually adjusting the iPhone's brightness slider. Measurements were taken using the x-rite i1 Display Pro[3] and the i1 Profiler Software 1.1.1.

The experiments were conducted using HMD Merge virtual reality[4] headset. This HMD is compatible with Android and iOS smartphones. Its dual inputs (named buttons in the following) facilitate and expand possible interactions within the testbed. The adjustable lenses lead to a more comfortable experience as they are designed to fit one's specific eye distance. This HMD guarantees a field of view of 90 degrees. The entire system has a precision of 8.3 ppd.

4.7 Evaluation methodology

Several studies in the past have evaluated the visual quality of tone mapping operators. Ledda et al. [Ledda 2005] conducted a subjective evaluation of tone mapping operators with a reference HDR display, while Yoshida et al. [Yoshida 2005] presented a study evaluating tone-mapped images with real-world scenes. Kuang et al. [Kuang 2004] did a pairwise comparison experiment between TMOs on a single SDR display. Evaluation of tone mapping for HDR video has also been well covered by Eilertsen et al. [Eilertsen 2013] and more recently by Melo et al. [Melo 2015].

Despite extensive studies done on this topic, the underlying conclusion among existing works is that the preference of TMOs is very subjective. In a subjective evaluation more aligned with our work, Narwaria et al. [Narwaria 2014] compared tone-mapped images with single exposure images. The study concluded that observers saw no significant differences between tone mapped and single exposure content. The authors explain this unexpected result citing many possibilities, including details in the bright or dark areas, unnatural colors, overall contrast, naturalness of the scene, etc. We will also consider these perceptual factors for our experiments.

In the rest of the section, details of the subjective test are discussed, from the design of a methodology for omnidirectional content pair-comparison to the definition of our experiment and the creation of a comprehensive questionnaire.

4.7.1 Pair Comparison approaches

In this study, we addressed a direct comparison between SDR content and HDR-based representations. This implied using a pair-comparison methodology. Figure 4.8 explains how HDR variants were generated, as described in Section 4.3, and defines the mid-exposure content as the SDR reference.

To the best of our knowledge, only SS methodologies have been used in previous subjective tests on omnidirectional contents. In PC methods, the relation between

[3] http://www.xrite.com/categories/calibration-profiling/i1display-pro
[4] https://mergevr.com/goggles

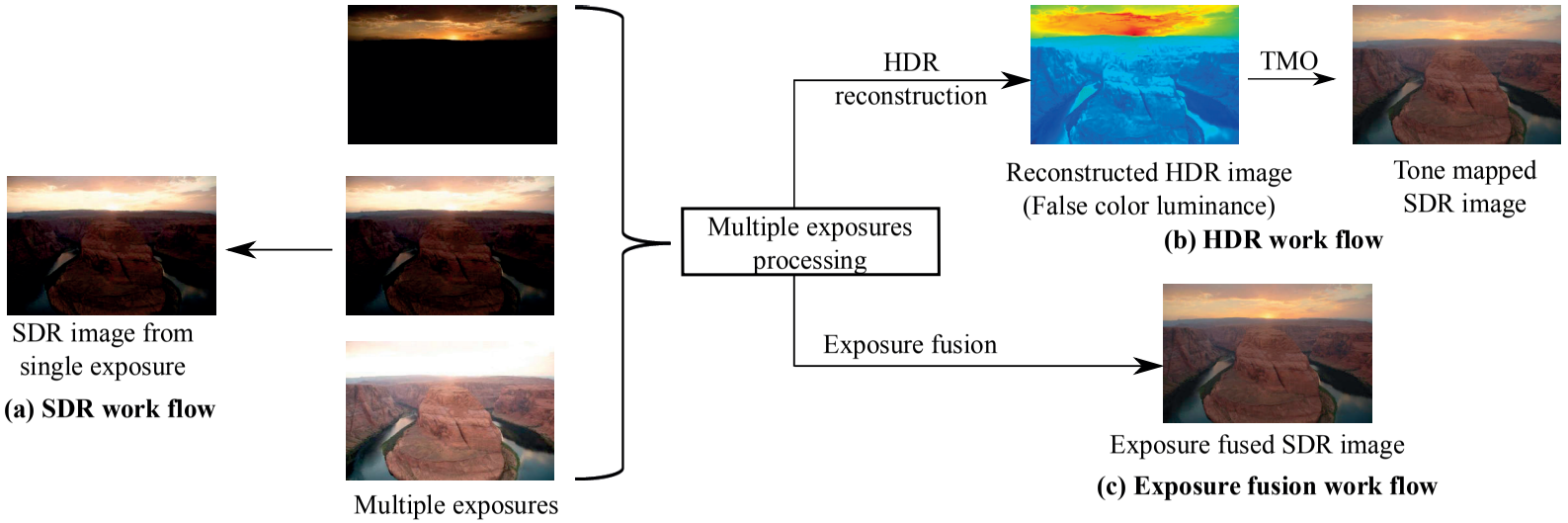


Figure 4.8: Three imaging workflows are shown. The example in (a) shows the SDR workflow based on capturing the mid-exposure of the scene. The HDR workflow in (b) describes the typical HDR pipeline consisting of bracketing, HDR reconstruction and tone mapping. The exposure fusion workflow is seen in (c).

two images or sequences of images is evaluated. The set of stimuli is usually presented as juxtaposed, or sequentially. Even though those are two preferential judging methods, as emphasized by the ITU recommendation [ITU 2012c], the choice of using SS or PC is based on the context of the analysis and the aim of the experiment. SS sorts contents by level of impairments or quality in an absolute way while PC permits to discriminate one content when compared to another, determining the relative quality of the impaired content. As we conduct here an analysis of the relative improvement of the perceptual quality of multi-exposures over single exposure workflows, a PC methodology is more suitable.

The first idea that was explored was to recreate SbS methodologies for PC in an omnidirectional environment. So, it was initially planned to split the screen into two parts and propose a spatial comparison. This rendering would have reduced the cognitive load required from subjects in sequential presentations.

Various variants of pair comparison methodologies were envisioned. This included well known side-by-side implementations such as butterfly and split screen. The main disadvantage of these approaches is lack of naturalness of the scene, especially when seen through an HMD. In Figure 4.9, we demonstrate the limitations of various side-by-side implementations for the *Room* image.

As the approaches mimicking the spatial pair-comparison methodology would not provide a valid assessment methodology, another solution was conceived. To genuinely compare the single and multi-exposure-based contents, instead of modifying the material to have both versions in the omnidirectional content, it was reasonable to enable the switch from one content to the other. It was then decided to introduce this new interaction in the testbed: the toggling.

(a) Split Screen

(b) Horizontal Butterfly

(c) Vertical Butterfly

Figure 4.9: Approaches considered to reproduce SbS PC methodology in an omnidirectional environment are presented. Using a split screen, as seen in (a), forces the user to evaluate different parts of the same content, which violates the construct validity of the experiment. Similarly, the butterfly comparison in (b) and (c) will result in very unnatural environments which may bias the assessment.

4.7.2 Toggling

A pair-comparison toggling approach means that the user can visualize both contents by switching between them. We build upon the testbed described in Section 4.6 to introduce this new dual-stimulus evaluation methodology for omnidirectional imaging.

The evaluation starts by first displaying the test stimulus. By pressing the right button on the merge HMD, the reference stimulus is presented on the current viewport. By pressing the right button multiple times, the user can toggle between the test and reference stimulus. Each scene is labeled either as 'T' for test or 'R' for reference at the viewport center, indicating to the subject which stimulus is currently being displayed. It is mandatory to toggle at least twice before scoring. No maximum limit on the number of toggling is set.

The vote is enabled only on the test stimuli so that the last visualized content is the one to be assessed. The left button of the HMD is used for the voting process. By first pressing the left button, a stationary vote menu is displayed. With the help of a red disc at the viewport center, users can select their score on the voting menu by head movement. Once the red disc has been correctly positioned over the preferred score, the user can cast his/her vote by pressing again on the left button. Immediately after, the next pair-comparison evaluation starts.

The order of presentation of pair-comparison stimuli is randomized in such a way that same-content pair-comparisons are not assessed successively.

The new testbed collects scores and tracks the direction of views, just as the single-stimulus testbed. It also has the feature of recording toggling information, by storing the timestamps and the viewport directions when users toggle. Such information is meant to be used to investigate the comparison process of subjects, and to observe the regions of interest for HDR omnidirectional contents. This could potentially help to identify areas of improvement for future omnidirectional TMOs as well as validate this new subjective assessment methodology.

4.8 Experiment design

We defined our pair-comparison subjective evaluation as an adjectival categorical judgment method on a 5-points grade (1: T worse than R; 2: T slightly worse than R ; 3: T same as R; 4: T slightly better than R; 5: T better than R). T refers to the test stimulus (TMO or exposure fusion image) while R is the reference stimulus (single exposure image). The rating scale is displayed on the voting menu of the testbed. The reference image is chosen as the single exposure image with EV = 0. This choice is based on photographic principles, as the mid-exposure usually permits the acquisition of bright and dark areas without favoring either highlights or shadows.

With regards to the identification of influence factors, we created a post-questionnaire investigating multi-dimensions of HDR 360° experiences. We focused on dimensions specific to both technologies and wondered about the impact of con-

tent aesthetic characteristics on subjective evaluations. Our approach to extract and select influence factors is presented below.

1. Influence factors related to HDR:
   The perceptual advantages of HDR imaging, when compared to SDR, are numerous. Among them are the abilities to represent more extensive ranges of luminance and colors, which eventually exhibit enhanced contrast and colors, and present more details in bright and dark areas.

   Accordingly, HDR images are evaluated according to various criteria, such as image contrast, naturalness, colorfulness, and overall brightness. Given our aim to deliver recommendations for the development of omnidirectional TMOs as well as for future questionnaires appraising 360° HDR contents, we investigated what factors influence the assessment.

   We aimed to explore all influence factors considered in HDR studies comprehensively. Narwaria et al. [Narwaria 2012] reviewed the criteria on which the differentiation of tone mapped images is based. The recurrent and most used factors were selected for our questionnaire: *Details in the bright areas*, *Details in the dark areas*, *Unnatural colors*, *Ghosting*, *Noise*, *Overall brightness*, *Overall contrast* and *Naturalness of the scene*.

   Additional empty spaces were added in case subjects wanted to make propositions.

   The factors listed above were extracted from numerous evaluations on HDR images and videos. It makes sense to evaluate each factor. We thus did not ask for the importance and relevance of a factor in the assessment of contents (on a Likert scale). Such an evaluation would not have given additional knowledge for the development of HDR 360° imaging.

   When designing this test, we had a hard time selecting HDR-related factors. We wanted to include as many most descriptive and representative factors as possible. However, there was no clear indication of how to choose influence factors.

   This issue is part of the main challenges faced by researchers when designing an experiment. My research interests in this thesis were to provide guidelines and information for influence factors identification. Factors selection among a set of influence factors is part of our research focus.

   Our perspective of evaluation is thus more to understand which factors are most considered in the assessment of HDR 360° images. Accordingly, subjects were asked to rank influence factors per likeliness to be examined during the preference choice. These criteria were ranked from the most considered (1) to the least considered (8) when making the evaluation decision.

2. Interest and liking in content:
   Contents affect an experience significantly. In perceptual evaluations of quality, it is hard for assessors to separate aesthetic aspects from visual quality. It

is easy to relate visual quality to the level of degradation (e.g., blur or noise). However, it is more complex and challenging to link characteristics usually viewed as aesthetic features (e.g., color, composition or lighting conditions) to visual quality. People are often facing bias due to personal tastes. This is why we investigate here aesthetic aspects of contents to see if they are linked to subjects behaviors.

Typically, aesthetic visual quality usually combines the analysis of low- and high-level features. For instance, exposure, sharpness, contrast, colorfulness, texture are regarded as low-level features. High-level aspects are representative of the content layout (e.g., composition, naturalness or being visually appealing) [Marchesotti 2011, Li 2009a].

In [Li 2009a], authors recommend to include color, brightness and composition concepts. Marchesotti et al. [Marchesotti 2011], state the importance of illumination for HDR images. For typical contents, their outcomes indicate that color enables to differentiate high-quality and low-quality images the most compared to other aesthetic features. Their predictor of pleasing or not pleasing content based on images blurriness and noisiness is quite accurate.

Apart from factors in aesthetic literature, we had a look at judging criteria in photography contests. Factors are only high-level features such as composition, color/lighting/exposure/focus, level of details, "wow! factor", visual appeal, memorability, overall artistic impression, and originality [5] or creativity, photographic quality, genuineness and authenticity of the content [6].

In our context, we considered the use of aesthetic appreciation to relate certain behaviors of subjects to content characteristics when scoring.

Regarding the factors included in the appraisal, we made the following decisions.

- Low-level features are difficult to assess in our contents. These were captured to present a variety of textures, composition, shapes and combine several expositions information. Thus, we have decided to include mostly high-level aesthetic aspects in content appreciation.
- Based on the literature, we had to include a color factor.
- Our contents have been created to present several different brightness/ illumination/ lightning in various viewports of the 360° content. This made impossible the evaluation of content as bright or not. We thus did not include the aspect of brightness in aesthetic factors.
- Various standard rules of composition in photography do not apply to 360° imaging. For instance, the well known "rule of thirds" is not relevant if the content format is not rectangular. Consequently, the composition was not viewed as a factor in the context of the study.

[5] http://www.mardenkane.com/articles/criteria-for-judging-a-photo-contest.html
[6] https://photography.nationalgeographic.com/contest-2015/rules/

- Our content selection included the rejection of contents presenting blur and noise. We must specify here that night contents contain a fair amount of noise, due to capture conditions. As we conduct a pair-comparison study, there is no issue to include those contents in the evaluation. We combined measures of distortions by asking if the raw content is "of quality".
- To our mind, other high-level features had to be included in the aesthetic evaluation. For instance, viewers may stop fully exploring a 360° content (after the first encounter) should it be viewed as boring. We relied on the work of Mansilla [Mansilla 2013] on quality of aesthetic experiences to include the aspects of boredom, interest in the content, familiarity, pleasure, and immersion.

Accordingly, the aesthetic appraisal of content was assessed through a self-made questionnaire investigating if a content was *Boring*, *Interesting*, *Colorful*, *Aesthetic*, *Familiar*, *Of quality*, *Pleasurable* and *Immersive*. Subjects selected the attributes they found relevant for each content. An additional open question inquiring the dislike or like of the content was added to get further insight into the reasons for content preference.

3. Sickness:
Virtual reality is known to generate sickness symptoms in viewers, especially nausea when the experience is active (e.g., the user interacts with the system) [Sharples 2008]. This effect is due to discrepancies between information transmitted to the brain by vestibular and visual systems [LaViola Jr 2000]. Physiologically, women are undoubtedly more affected by sickness than men [Reason 1978].

   It is essential to verify if participants were subject to sickness to validate our results and assessment methodology. Indeed, the less sickness effect is observed, the more reliable will be the gradings. Besides, the potential gender bias induced by sickness is interesting to study.

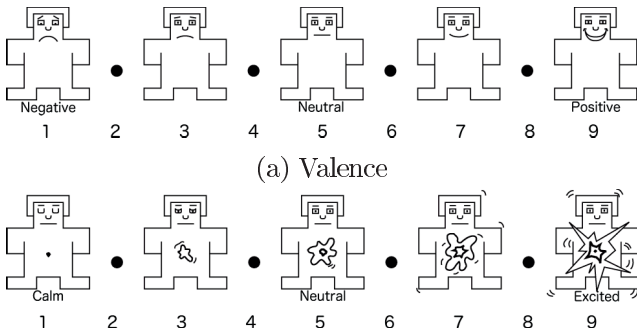
   One recognizes virtual reality sickness (also called simulator sickness) thanks to several symptoms, usually sorted into three categories, namely nausea, oculomotor and disorientation. In her work, Kolasinski related and adapted the concept of motion sickness [Kolasinski 1995] to simulator sickness in virtual environments. She had created the widely used simulator sickness questionnaire, using 16 symptoms to describe sickness. Among them are found general discomfort, fatigue, headache, dizziness, and vertigo. Kennedy et al. [Kennedy 1993] extended this questionnaire with new symptoms (for a total of 28) such as drowsiness, decreased salivation, depression, decreased/increased appetite, confusion and vomiting. In [LaViola Jr 2000], other symptoms are deemed appropriate for sickness evaluation, for instance, pallor, fullness of stomach and dryness of mouth.

As nausea symptoms are well described in the original simulator sickness questionnaire and that our test aimed at generating as low level of sickness as possible, we stuck to the widely used survey for the evaluation of virtual reality sickness.

Accordingly, to accurately evaluate the degree and type of sickness generated by our experiment methodology, the simulator sickness questionnaire by Kolasinski [Kolasinski 1995] was included in our questionnaire. The symptoms *General discomfort*, *Fatigue*, *Headache*, *Eye strain*, *Difficulty focusing*, *Increased salivation*, *Sweating*, *Nausea*, *Difficulty concentrating*, *"Fullness of the head"*, *Blurred vision*, *Dizziness with eyes open*, *Dizziness with eyes closed*, *Vertigo*, *Stomach awareness* and *Burping* were evaluated on a 4-points scale from none (1) to severe (4).

4. Virtual reality and overall experience:
After considering HDR and aesthetic influence factors, it is high time to include omnidirectional aspects in the assessment. More general considerations are investigated too.

   - VR and omnidirectional contents aim to provide more immersive contents in showing a new environment information covering a full sphere. Immersion and isolation are thus undisputed factors to include here. The proposed experience is not especially engaging as the only interaction with the content is its exploration. If this dimension should have been evaluated for experiences like VR gaming, we determined that this context does not require the assessment of engagement.
   - It is current to consider subjects' state of mind, particularly boredom, when designing a subjective experiment. For instance, apathy indicates a possible lack of consistency in given scores at the end of the test resulting from the low attention of subjects.
     We decided to include an evaluation of subjects mood after the test to verify the validity of collected scores. To do so, we examined several emotion models to represent subjects state of mind following the experiment. Dimensional models are available to represent moods. The self-assessment manikin [Bradley 1994] represents emotions following two independent dimensions, namely arousal and valence. Though this two-dimensional space represents most emotional variations, a third dimension, dominance, may be included to indicate if one feels empowered and in control. Miller mood map highly similarly categorizes emotions following a two-dimension space (energy and valence). Another strategy is to assess both negative and positive affect as in Positive and Negative Affect Schedule (PANAS), a widely used questionnaire in psychology, which evaluates ten items for both aspects [Watson 1988]. Instead of appraising positive and negative affects, one can categorize emotions in primary emotions as done in the Plutchik's wheel of emotions [Plutchik 2001].



(a) Valence

(b) Arousal

Figure 4.10: Self assessment Manikin valence and arousal dimensions evaluation. 5- and 9-point scales are possible with this representation.

Height basic emotions, namely ecstasy, admiration, terror, amazement, grief, loathing, rage, and vigilance, can be mingled to create any other emotion. That is, aggressiveness is a mix between vigilance and rage; boredom is low-intensity grief.

Our concerns are not to have a precise report of subjects emotions but to get an overall state of mind. We do not need complex representation of emotions through multiple dimensions or categories. Thus, we decided to include two questions for valence and arousal following the self-assessment manikin space. Figure 4.10 presents 5-point scales representation.

- VR and HMD were pretty new to subjects. To have an exhaustive profile of users to analyze our data, we included a question demanding subjects' familiarity with VR. We wanted to categorize subjects in naive, casual and expert viewers for VR. We assumed that naive users have no or less than five experiences with VR.

  As seen in photography factors, a "wow" effect, also named novelty effect, may happen when encountering an experience for the first times. We found interesting to inquire about the possible influence of the novelty effect on experiences and to observe whether subjects are aware of this potential bias during the assessment.

Considerations about omnidirectional, emotional and novelty aspects of HDR 360° experiences led to the appreciation of *Immersion*, *Isolation*, *Enjoyment*, *Arousal/Excitation* and the *Enjoyment due to the novelty of the visualization*. A 5-point Likert scale from strongly disagree to strongly agree was used when asking subjects if they experienced the attributes mentioned above.

Valence was renamed enjoyment and arousal was coupled with excitation. This is explained by the fact that we use a Likert scale to fit with the assessment of immersion, isolation, and novelty. Also, we paid attention to the fact that excitation may be understood more easily by the tested

population, which mostly contains native French speakers.

Subjects often orally report additional insights on the experiment when they have completed it. Such comments are likely to help to define influence factors further or modify the test methodology to increase subjects' comfort or the system usability. We thought that it would be interesting to have a written trace of those comments. Consequently, in an additional open question we asked subjects to describe their VR experience in a few words.

The post-questionnaire, together with the definition of technical terms are included in appendix 8.1.

The study was carried out on eight contents, selected from the dataset introduced in Section 4.4 following the process described in Section 4.5, on four TMOs and an exposure fusion operator, exposed in Section 4.3. Overall, 40 pair-comparisons stimuli were evaluated. A single test session of about 20 minutes was needed for the assessment of all pair-comparison stimuli.

There was no need to conduct this experiment in a laboratory environment as the entire sight of subjects was covered by the HMD. We thus only made sure to perform the test in a quiet room. Figure 4.11 shows subjects taking part in the test.

Figure 4.11: Set-up of the test environment

Before the test, subjects were screened for correct color vision and visual acuity using Ishihara and Snellen charts. As recommended in the simulator sickness questionnaire [Kennedy 1993], participants reported whether they are in their normal

state of health. Any violation of the previous conditions results in the rejection of the subject from the experiment. Overall, 25 subjects, of whom 17 are males, participated in the test and they were 29.5 years old on average, ranging from 19 to 55 years old.

After being instructed with the study process, subjects were requested to read the information and consent forms, the post-questionnaire, and the explanation sheet defining technical terms of the questionnaire. Any question was answered before starting the experiment.

To get subjects familiarized with the assessment procedure and to reduce grading discrepancies across-subjects, instructions were provided by the testbed and a training session was taken before the test session.

Three pair-comparison stimuli were rendered during the training session, in the following order: the photographic tone mapped *Bridge* content, assessed as worse than R (1), the display adaptive tone mapped *Trail*, evaluated as equivalent as R (3), and the exposure fusion of *Sculptures*, assessed as better than R (5). Different contents were used in the training session when compared to the test session material, to prevent any bias introduction.

The test session starts right after the completion of the training session. It is recalled that subjects had to toggle at least twice for each pair-comparison and could only vote on the T stimulus. No time-constraints were defined per stimulus or test session. Once the test session is over, subjects are required to fill the post-questionnaire. The test was carried out in a calm environment, free from any disturbances. A spinning chair was used during the assessment for subjects comfort as well as to ease their 360° navigation within the omnidirectional contents.

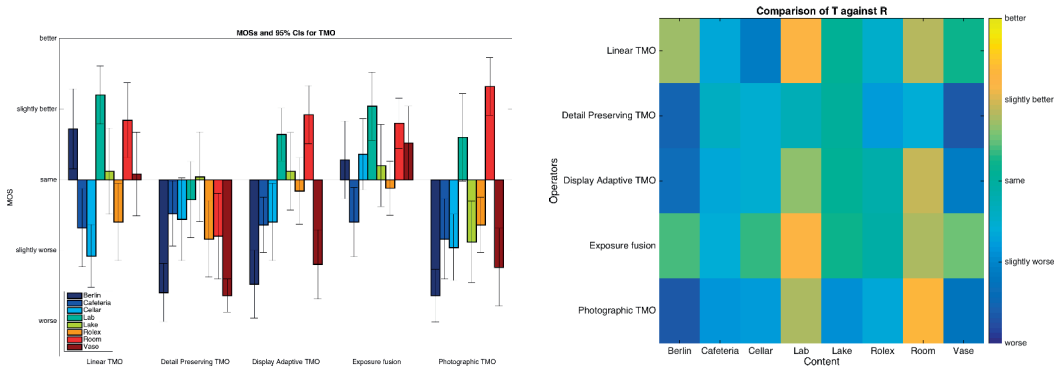
4.9 Results and analysis

The analysis covers three types of information, namely subjective scores, post-questionnaire data, and view direction tracking, combined with toggling information.

4.9.1 Subjective scores

After an outlier detection based on Rec. ITU-R BT.500 [ITU 2012c], which did not reject any subjects from this study, MOSs and 95% CIs were computed and are displayed in figure 4.12. Figure 4.12a shows the mean score variations across contents for each operator, while Figure 4.12b presents a comparison matrix depending on contents and operators.

Overall, the results do not show a clear preference for the processed content over a single exposure content. Indeed, MOSs mainly range from 2 (T slightly worse than R) up to slightly above 4 (T slightly better than R), indicating a moderate improvement or deterioration of the perceived quality. Based on the Figure 4.12b, the exposure fusion and the Linear TMO show better performance when compared to the other operators. The extent of the performance is a similar or slightly improved perceived quality. On the other hand, the photographic, display adaptive and detail



(a) MOSs and 95% CIs per content per operator

(b) Comparison of MOSs

Figure 4.12: MOSs and CIs analysis

preserving TMOs do not perform as well. This result does not come as a surprise and can be explained through the limitations of the consumer camera. Let us recall that the DR measure of the dataset spanned from 7.59 - 12.27 f-stops. These TMOs are designed for content with a DR greater than 16 f-stops. Thus, results suggest that TMOs require further evaluation using professional 360° HDR content having higher DR. As a result, the dataset is ideal for the exposure fusion algorithm, which works well for images consisting of equally spaced EVs. This method avoids the HDR reconstruction step and is thus independent of the DR of the scene. Also, it is well known in HDR imaging that a linear TMO often results in a dark image and loss of detail [Reinhard 2007]. The fact that we can visualize Figure 4.2a also highlights the lack of DR in the content.

In addition to this, the differences across contents are significant, as several non-overlapping CIs demonstrate. This illustrates the variety of the chosen contents. It is worth noting that the variation of key values across omnidirectional content is a relevant indication for content selection. All operators, except the detail preserving TMOs, are particularly preferred over the reference for the contents *Lab* and *Room*, two indoor contents with an outside view through a window. No trend in the statistics found in Table 4.1 and Figure 4.7 seems to justify this behavior. This demonstrates the need for a precise questionnaire investigating specific content-related criteria.

The non-normality of distributions of scores prevents us from running a repeated measure ANOVA, especially considering our sample size (=25), which is not sufficient to overcome the violation of the assumption. As the aim of this work is not to differentiate TMOs but to investigate the need of a dedicated TMO for omnidirectional contents, no non-parametric test has been performed as they would be hard to interpret in the context of this study.

4.9.2 Post-questionnaire answers

- Criteria considered during the evaluation
  To identify which criteria are the most impacting the assessment process, the Borda count method [de Borda 1781] was applied. First grades were reversed ($\hat{MOS} = 8 - (MOS - 1)$) in order to have a higher number for most considered criteria. Then, a weighted average of $\hat{MOS}$ has been computed. These weights typically represent the rated importance of criteria. Criteria that have been the most assessed in first considered criteria are thus favored

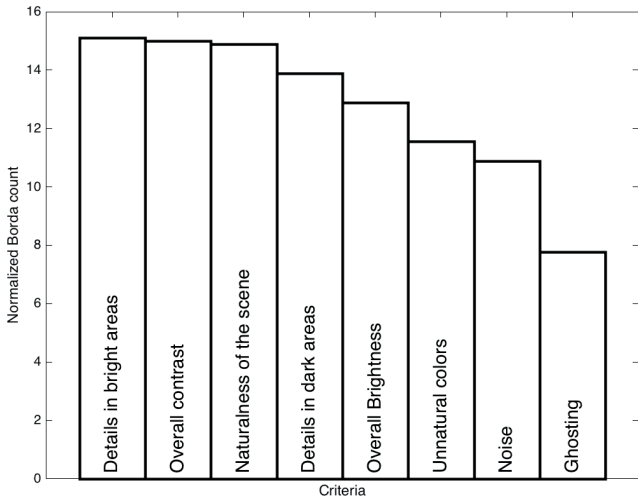


Figure 4.13: Rank of evaluation criteria

The results are presented in Figure 4.13. From the figure, we see that the details in bright areas, overall contrast, and naturalness of the scene are the most critical factors for the subjects.

- Interest and appreciation of the content

The Figure 4.14 illustrates the subjects' choices of suitable adjectives describing the contents. Contents showing the best results, *Lab* and *Room*, were both assessed as dull by more than 50% of the subjects. This is a curious finding regarding the content selection in subjective tests, that could be further investigated. When considering the content *Lake*, the results of its MOSs indicate a similar perceived quality from T to R stimuli for all TMOs, except for the Photographic TMO. More than 80% of subjects assessed the content as colorful, and the operator mentioned above introduced unnatural colors. Even if the criteria concerning the unnaturalness of colors was among the least considered, we can conclude that this criterion is still of importance in the evaluation of tone mapped omnidirectional HDR content.

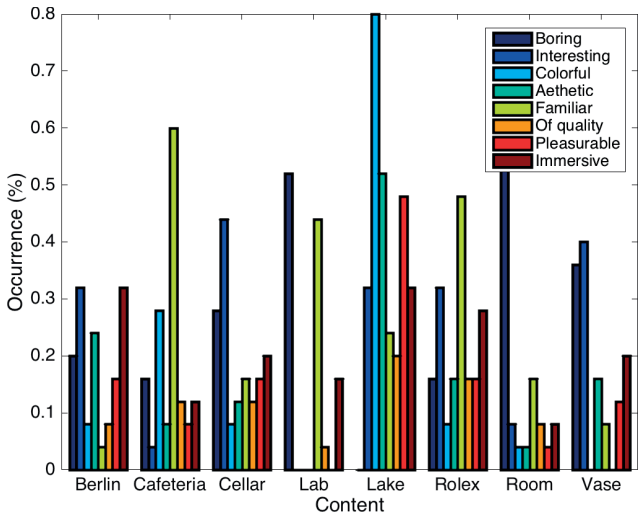


Figure 4.14: Content type

- Sickness

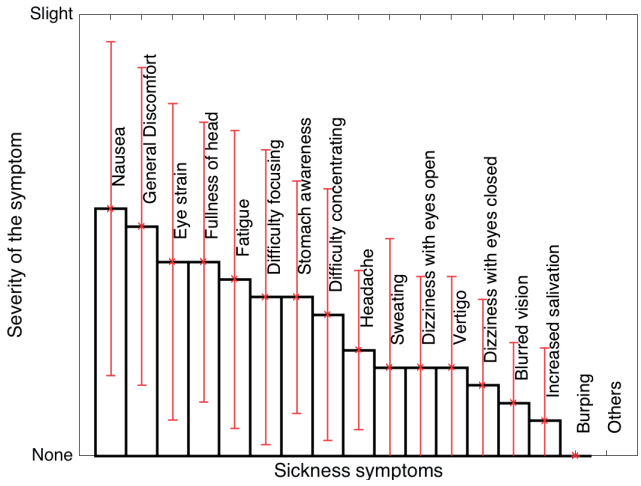


Figure 4.15: Sickness symptoms

Various works in literature have reported that a virtual environment can generate sickness [Cobb 1999]. Subjective evaluation is prone to bias if users are in discomfort or tired. The questionnaire aims to certify the validity of the test. Figure 4.15 reports the degree of severity of the sickness effect in subjects. We can observe that the extent of sickness is from none to slight. Also, 48% of our subjects reported not having experienced sickness. These two facts confirm the proper design of the test for pair-comparisons, concerning the sickness effect. Overall, when considering the severity and occurrence of sickness symptoms, slight nausea and general discomfort characterize the sickness experienced by

our subjects.

In terms of gender bias, about half of males and two third of females subjects experienced sickness. We verify here the fact that women are more prone to feel sick than men.

- Virtual Reality Experience

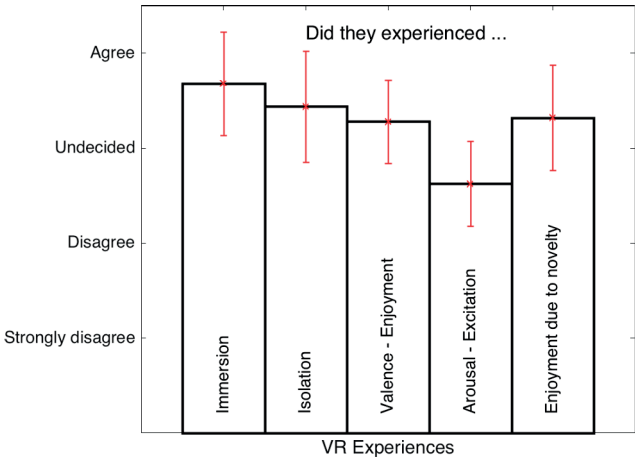


Figure 4.16: VR experience

The investigations carried out on subjects' experiences, and feedbacks are reported here. Before the test, 36% subjects had never used an HMDs, while 16% had used it more than five times. Remaining participants had few experiences with VR. We can conclude that our population sample is representative of the population when considering their level of familiarity with VR experiences.

The evaluations of subjects' experience showed a moderate immersion, isolation, and enjoyment. Subjects acknowledged a lack of arousal. Those results are mainly explained at the end of the post-questionnaire in the open question, which asks for a summary of the subjects' experience. Indeed, more than a quarter of the subjects mentioned that their excitation level was counterbalanced by their perception of the low-quality level of the images, especially when pixelation was perceived, or by the experience of sickness.

The reported feedbacks from the subjects in the open question indicate discrepancies in the appreciation of enjoyment and perceived quality. Spontaneously, nine subjects reported a funny or great experience, while five did not thoroughly enjoy the experience. Regarding the quality of the display, two subjects mentioned an excellent image quality, while three others found the quality (or resolution) not as high as expected or not sufficient to provide a truly immersive experience. These findings explain the modest experience of enjoyment reported in MOSs.

Subjects complained that it was tiring to keep their hands lifted to their face to control the two buttons of the HMD. This complaint must be examined in future experiments to reduce even more the likelihood of tiredness and lack of attention of subjects. Controlling the HMD with a controller (e.g., game-pad) could tackle this issue.

Subjects also gave indications to improve the immersion and isolation experience: one mentioned that he was "experiencing contents as a ghost instead of being physically there", another stated that "it would be more immersive to zoom in and out." Despite the soundness of these comments, interaction with content is out of the scope of this study. Also, implementing it will inevitably introduce various bias without providing significant new insights research-wise.

1. To enable zoom in and out interactions will make results harder to analyze and control. In addition, implementing interaction, knowing that only two buttons are available, will increase the complexity of the instructions provided to subjects.
2. To include a representation of subject's body in the 360° view is complicated and adds a constraint to content acquisition process: only pictures taken with a stable tripod of the same height across scenes should be used. This constraint is not always satisfied in the created database. Also, the bias introduced by having a body which is not the subject's one (e.g., height and skin color) or implementing the representation of a subject's body will be needing an appraisal.

Some subjects indicated that the duration of the test was correctly set, as they were starting to get bored and annoyed when visualizing the two or three last stimuli. Considering this comment, a test session (including explanations and training, if applicable) should not exceed 20 minutes, meaning about 40 pair-comparisons stimuli.

One subject mentioned the approach he carried out during the experiment, especially regarding the visualization: "At the beginning, I tended to move in all directions, and then, I progressively focused on areas of "interest", typical of each content" and had the "tendency to navigate more horizontally than vertically, and then to balance in both directions." These are essential indications towards the toggling information analysis. It is worth noting that the second assertion is consistent with the findings of Rai et al. [Rai 2017] about the quasi-isotropic trend of the distribution of gaze fixation orientations.

A few subjects claimed that choosing between the test and reference had been a challenge. They found cases where they preferred the test content over the reference content in bright regions, while not so in dark regions, and vice-versa. This suggests that further work is needed in 360° TMOs design to overcome such challenges.

4.9.3 Toggling and fixation locations processing

The experiment was designed to enable the comparison between test and reference stimuli. We meant here to process tracking and toggling information to observe and understand subjects behavior while comparing two omnidirectional HDR contents. Two approaches were considered while processing the toggling information.

- The first one is based on the assumption that subjects straightly compare stimuli where they have toggled. Indeed, the two contents are solely directly compared during the action of toggling, when the reference content is displayed immediately after the test content, or vice versa. Thus, the first analysis of toggling results consists of the observation of toggling locations. This approach is rather simple and straightforward. It consists of displaying the toggling locations of every subject for each content after having re-sampled those locations in cells of 1x1 degree. Toggling positions having occurred only on one of test or reference stimuli have been discarded. It is recalled that those locations indicate the center of the current viewport when toggling from one to the other of the paired-stimuli.

- The second approach consists in assuming that fixation locations present in test and reference paired-stimuli can be considered as content comparison areas. This method is more complicated, as it introduces new considerations such as the recency effect as well as the memorability of the paired-stimuli, which increase subjects' cognitive load. The interest of this second method resides in the fact that same reference stimuli will be visualized several times, precisely five (four TMOs and the exposure fusion evaluations). This can lead to possible subject behaviors which were not considered in the previous approach: subjects may visualize the entire reference stimulus, to refresh their perception of key areas of this stimulus, before comparing it with the test stimulus.

  The implementation of this second approach is inspired mainly by Upenik et al. [Upenik 2017] and proceeds as follows. The available raw data consist of the tracking of head movements and toggling information, in the form of two arrays of view-port center's yaw and pitch coordinates along with timestamps. The head movement tracks are first split into test, and reference head movement tracks sets, according to which of the test or reference stimuli was visualized. If not mentioned otherwise, the processing described below is applied separately to each of the two sets. We have defined fixation locations as being the locations where the angular velocity of an observer's head does not prevent the subject's ability to focus attention on an object.

  The angular velocity is obtained by computing the first order time derivative of yaw and pitch coordinates. To remove any digital differentiation noise the sets of head movement tracks were filtered in advance of the derivation with a second-order Butterworth low-pass filter with cutoff frequency of $f_c = 2$ Hz. The threshold defining if a track location is a fixation was set to 15 degrees

per second, complying with the work of Upenik et al. [Upenik 2017], and was verified as sound by a conducted analysis of head angular velocity.

The fixation locations of every subject were fused by summing all the locations in cells of 1x1 degree. The overall fixation locations indicate the fixation locations present in both test and reference fixation locations sets, resulting from the above-described processing.

Typical evaluations of 360° contents involve another step in processing. Fixation locations are not exactly representative of visual attention. Indeed, subjects rarely focus precisely on the same point [Upenik 2017]. Thus, it is common to apply a Gaussian filter to the fixation map to obtain a saliency map, representative of statistical areas of fixations [De Abreu 2017].

This step has not been performed so that we observe the actual distribution of fixation locations (centers of the viewport). Indeed, we do not want to observe visual attention but accurately observe where people looked at precisely. For instance, applying the Gaussian filter may prevent to see that one or many subjects paid attention to a specific highlight if it is located near several visual attention areas. It is then important to stress out that data analyzed are not visual attention areas but fixation locations.

The processing performed in this study only considers the head movement positions. As a future extension of our work, more accurate fixations can be computed based on the prediction of the eye gaze fixation from head fixation locations, applying the work of Rai et al. [Rai 2017], for instance. This has not been done in this work as no prediction model has been validated yet. Besides, the selection of the threshold indicating if the angular velocity of the observer's head prevents subjects focus of attention should be optimized. Perhaps the threshold should be computed per subject.

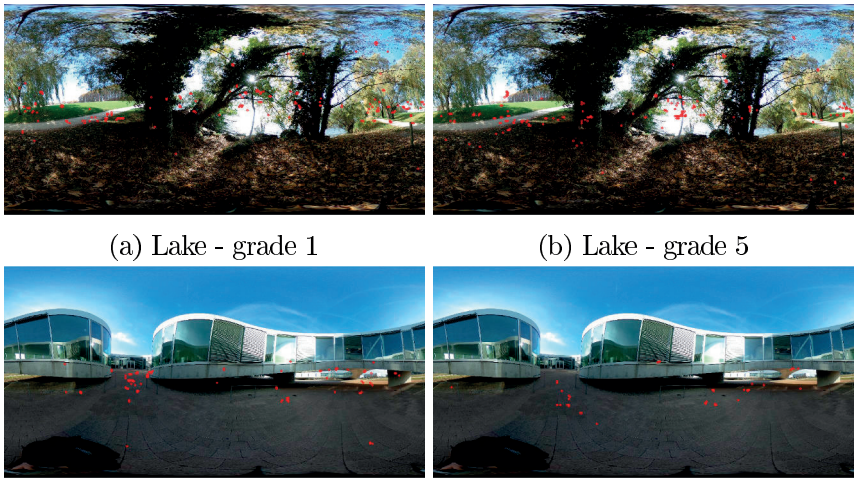
4.9.4 Toggling and fixation locations analysis

In this section, the results of the processing of toggling information previously-described are examined.

Figures 4.17, 4.18, 4.19 and 4.20 present variations in fixation results when considering different scores, gender, TMOs and if sickness has been reported.

Discrepancies between fixation results categorized by grades and TMOs do not show clear patterns among all contents. However, fixation locations of females are more spread on the pitch axis when compared to fixations of male. The same behavior is observed between subjects who have reported no sickness effect and those having experienced sickness effects. The sickness effect is preventing subjects to explore the omnidirectional content comprehensively as limited head movements reduce its impact. It also explains why female subjects are more prone to a slight sickness effect, as they tend to explore comprehensively 360° contents.

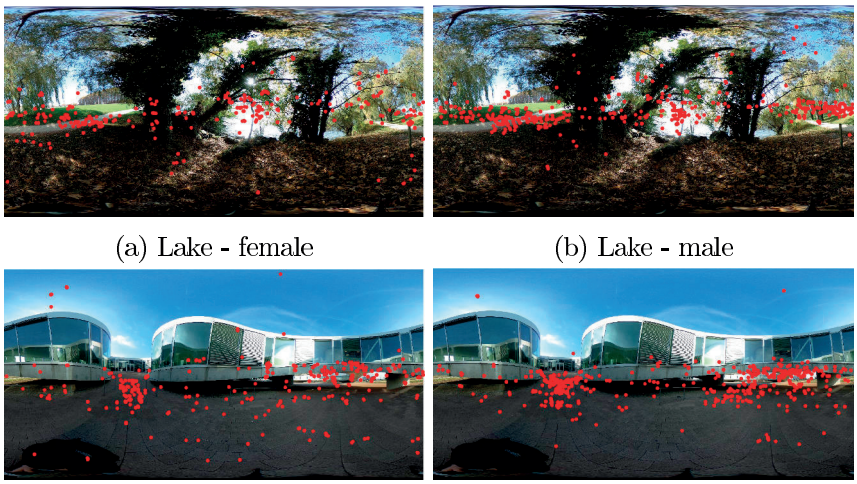
Figures 4.21 and 4.22 present toggling and fixation locations per content, respectively. The resulting toggling and fixation locations per content and per operator



(a) Lake - grade 1 (b) Lake - grade 5

(c) Rolex - grade 1 (d) Rolex - grade 5

Figure 4.17: Fixation locations (red points) for contents Lake and Rolex, discriminated per grade (1 and 5).

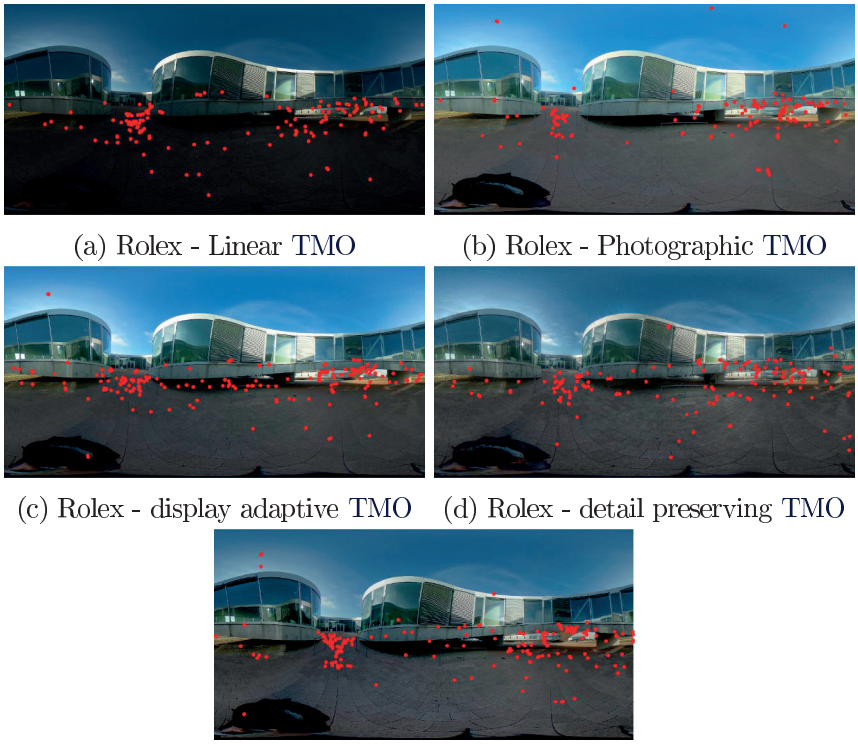


(a) Lake - female (b) Lake - male

(c) Rolex - female (d) Rolex - male

Figure 4.18: Fixation locations for contents Lake and Rolex, discriminated per gender

are not comprehensively reported as the behavior of subjects across operators shows minimal variations, especially for toggling locations. This fact is not in line with the behavior of subjects during the visualization of 2D contents, investigated by [Narwaria 2012] who indicates that every operator has a signature in visual attention maps. The toggling and fixation locations are thus not sufficient for an in-depth investigation of the behavior of subjects, leading to the need of introducing eye gaze tracking or at least eye gaze prediction. This conclusion is also confirmed by the results of toggling and fixation locations per content, per operator and per grades,

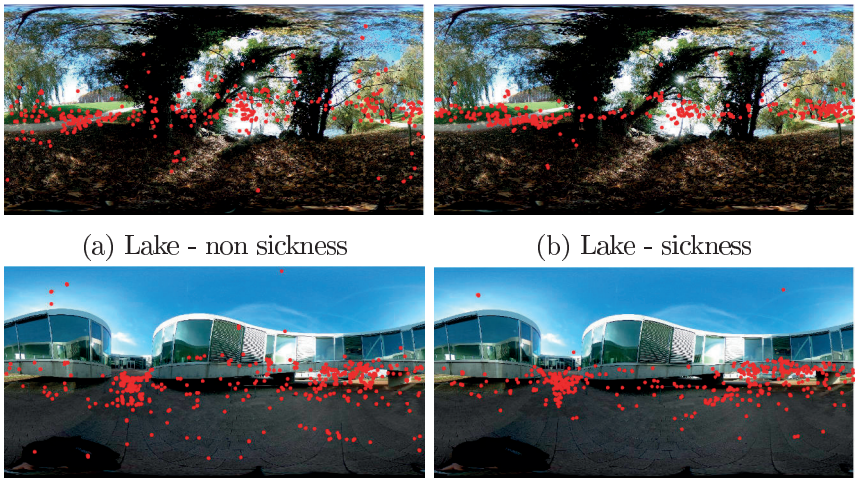


(a) Rolex - Linear TMO (b) Rolex - Photographic TMO

(c) Rolex - display adaptive TMO (d) Rolex - detail preserving TMO

(e) Rolex - exposure fusion

Figure 4.19: Fixation locations approach per TMO for Rolex content

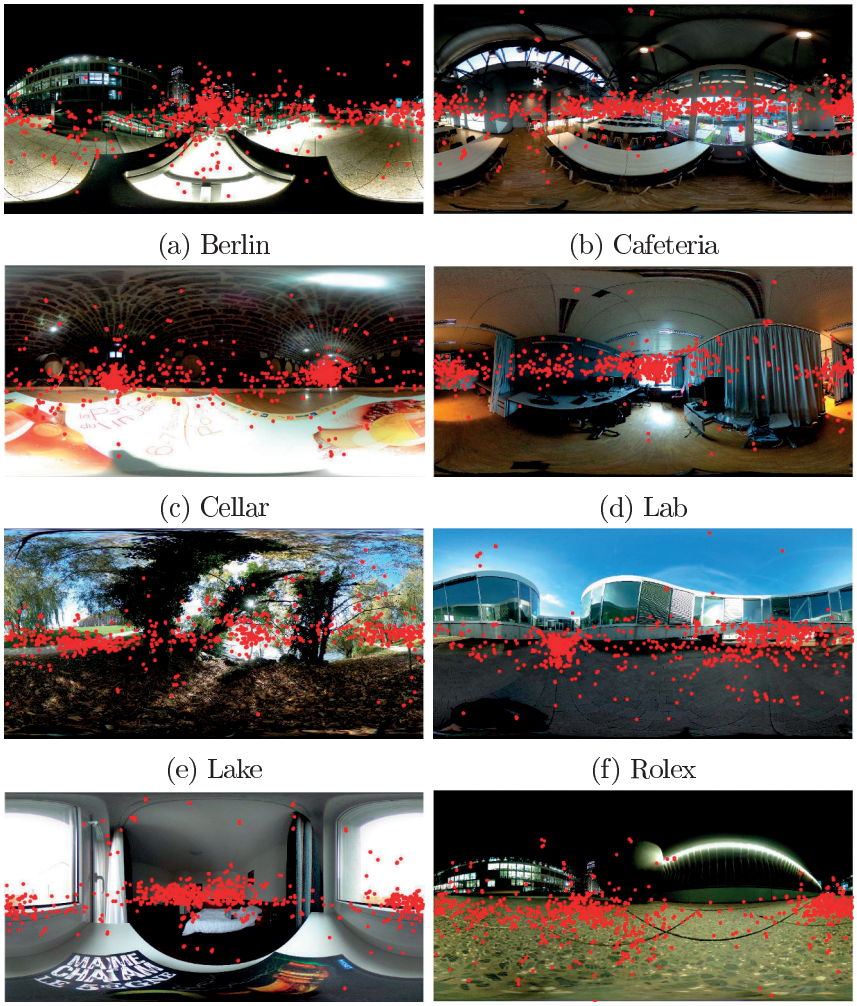


(a) Lake - non sickness (b) Lake - sickness

(c) Rolex - non sickness (d) Rolex - sickness

Figure 4.20: Fixation locations for contents Lake and Rolex, discriminated per reported sickness

which demonstrate that toggling and fixation locations are not sufficient to indicate the reasons of subjects' choices.



(a) Berlin (b) Cafeteria

(c) Cellar (d) Lab

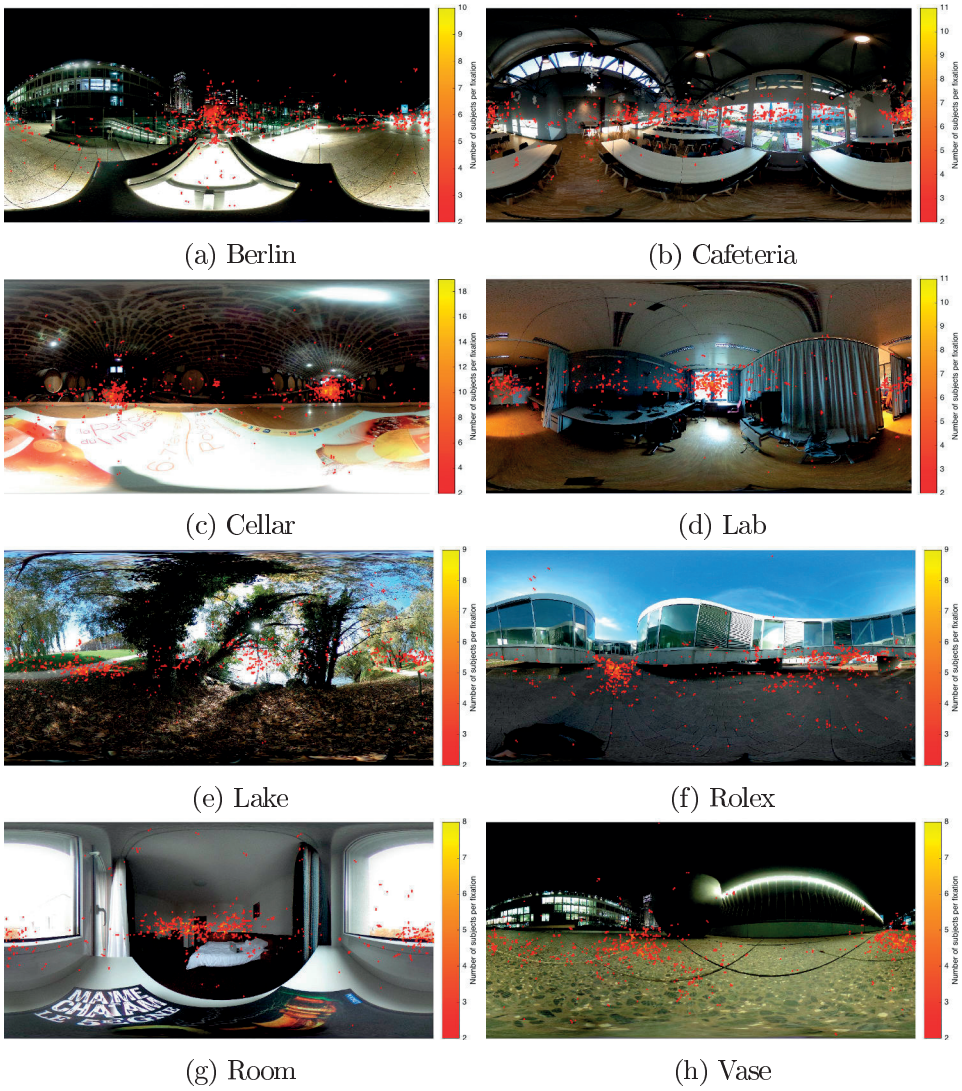
(e) Lake (f) Rolex

(g) Room (h) Vase

Figure 4.21: Toggling locations approach

The insights provided in Figures 4.21 and 4.22 clearly show the equivalence of using toggling and fixation locations. However, no guidelines are drawn from this study as both methods have advantages and drawbacks. The processing of toggling locations is not complex and emphasizes clear interest areas. However, the considerable variation in toggling coordinates prevents to see trends across subjects, as well as impedes the differentiation of two regions of toggling in terms of saliency. The fixation locations provide these insights, as they indicate weighted fixation locations based on the number of subjects sharing a specific location. However, this method requires more complex processing, which introduces a non-optimized parametrization (e.g., angular velocity threshold) and filtering. The results of this method can then be questioned as being less accurate or biased.

In the remainder of this section, we are using the results of both toggling and fixation locations to conclude about the behavior of subjects. The term "areas of



(a) Berlin (b) Cafeteria

(c) Cellar (d) Lab

(e) Lake (f) Rolex

(g) Room (h) Vase

Figure 4.22: Fixation locations (Test & Reference) approach, provided with the number of subjects per fixation locations

interest" refers to the areas with a high density of toggling or fixation locations.

The first result of interest is that there is no influence of the initial viewport on subjective scores. Indeed, contents *Cellar*, *Lake*, *Rolex*, and *Vase* are showing few or no toggling or fixation locations at the initial viewport position, indicating it is not a region of interest. We can thus conclude that the initial viewport is part of the regions of interest for *Berlin*, *Cafeteria*, *Lab* and *Room* contents.

Locations marked in results are mostly longitudinally distributed at a latitude close to the equator. Additionally, areas of toggling and fixation are also mainly centered at the equator latitude. In fact, the first quartile, average and third quartile of the set of pitch coordinates are 28.6, 1.9 and -23 degrees, respectively. It seems that this range of latitudinal head positions is comfortable and natural for the users.

When considering the distortions introduced by the equirectangular projection of a spherical content, we can conclude that, overall, each content was fully visualized, including the upper and lower parts of the spherical material.

Regarding the areas of toggling and fixations, it seems that subjects are paying attention to locations with artifacts, indicating a loss of information. The highlights (brightest areas) of several contents, such as the outside information through the windows of *Lab* and *Room* as well as the building and the sky over the grass field in the background of the *Lake* content, are clear toggling and fixation areas. Their toggling and fixation locations correspond to saturated regions in SDR contents, which present more details in HDR contents.

However, subjects did not focus on other highlighted areas saturated in SDR contents. Little attention has been paid to the reflection of the sun on the lake or windows in *Lake* and *Rolex* contents, light-emitting areas such as light bulbs in *Cafeteria* and *Cellar* contents, or buildings' rooms having the light turned on in night contents. In those locations, the loss of information is usually considered by expert viewers to verify the representation faithfulness and smoothness from bright to dark regions. A possible explanation is that, for naive observers, the importance of the lost information of reflective and light-emitting areas is not significant compared to other areas when considering the entire scene.

One can also notice the trend whereby observers focus more often on distant details than on close ones. In the *Cellar* content, two clear areas of interest are the extremities of the cellar. Overall, in all contents, main areas of toggling and fixation coincide with the locations of furthest objects. This behavior of subjects is interesting and can lead to new visual attention models as well as provide clues for the development of 360°-dedicated TMOs. Thus, this finding needs to be confirmed by further analysis with more various contents, as well as with contents with a more significant dynamic range and on other population samples.

To summarize, the insights provided by toggling or fixation locations are equivalent. They are, however, not precise enough to identify the variations in subjects' behavior depending on operator used. Regarding subjects' attention, we have seen that the initial viewport does not influence toggling of fixation locations. Also, viewers tend to browse through the content longitudinally with a reduced latitudinal range and are prone to focus on distant objects. At last, assessments of subjects seem to be based on the loss of information in specific areas.

4.10 Conclusion

360° contents are subjects to SDR limitations, for instance regarding contrast, colors or details in bright and dark areas. A burned area in VR is a disturbing experience as many visual data of the viewport are lost. To address this issue and make VR experiences more natural and faithful to reality, we explored the solution of using HDR representation.

We evaluated the interest in developing HDR for omnidirectional contents as

well as suggesting directions of improvement for dedicated TMOs. An end-to-end HDR pipeline evaluation using consumer equipment has been carried out.

I created a publicly available dataset composed of 43 multi-exposure images, acquired with a consumer-camera. The dataset also includes HDR reconstructions and SDR variants resulting from four well-known off-the-shelf TMOs and one exposure fusion algorithm. Eight of those contents, after careful selection, have been used to conduct the evaluation.

A new subjective test methodology, namely toggling, enabling pairwise comparisons for omnidirectional content, has been introduced. It allows switching between assessed stimuli and their references. This is, to our best knowledge, the first sequential pair-comparison methodology for omnidirectional contents.

The viewing material, compliant with the designed testbed, consisted of the mobile merge HMD combined with iPhone 6 and 6S. The environment of the test was a calm and quiet room, as no other environmental constraints needed to be addressed.

In a post-questionnaire, we studied several criteria to analyze our results further but also provide guidelines for future assessments of HDR 360°. Influence factors related to both technologies were included in the questionnaire. Also, we explored how to relate subjects' behavior to aesthetic characteristics of the content.

Our results exhibit that none of the evaluated operators show an apparent increase in perceived quality. Indeed, ratings go from slight preference of reference to slight preference of test content. However, the exposure fusion, as well as the linear TMO to a lesser extent, show promising results as being assessed as similar or slightly better than single exposure reference. This is in line with our expectations, knowing the limited dynamic range of the consumer-camera contents acquired. Accordingly, **HDR omnidirectional contents are promising. However, our study case showed that consumer-based material is not mature enough and must be improved.**

Discrepancies across operators and contents lead to the identification of a need for dedicated TMOs for omnidirectional contents. Subjects do not interact in the same way with 2D and omnidirectional contents. Behaviors attributed to 360° exploration, such as focusing on distant objects, change the approaches to follow to design future TMOs. Besides, contrary to expert viewers, subjects did not focus on visual cues that help to compare bright and dark details to realistic representations (e.g., light bulbs and specular reflexion on surfaces). Instead, they focused on the loss of information from SDR to HDR. **A future 360°-TMO should accurately represent furthest objects and then focus more on detail levels in highlights and shadows, than on realistic representation of illumination.**

The analysis performed on our self-made questionnaire emphasized that *Details in bright areas*, *Overall contrast* and the *Naturalness of the scene* are essential criteria to consider during the assessment. We were able to relate the poor performance of the photographic TMO to the crucial faithful representation of colors. This fact stresses that all HDR-related influence factors must be considered in future evaluations. Regarding VR experiences, some subjects reported sickness effects

with an extent ranging from none to slight, mostly felt like *Nausea* and *General discomfort*. This fact confirms the proper design of our methodology. Besides, evaluation of immersion, isolation, subject's state of mind and novelty effect was not as informative as expected. It seems that, except for the navigation through the content, HDR aspects prevail over omnidirectional content factors. This may be because these dimensions are evaluated in a post-questionnaire. Including influence factors in the main questionnaire (asked for every stimulus) may change this result.

Both the MMSPG of EPFL and the AMC of b<>com collaborated on this work. EPFL was responsible for content acquisition, making a 360° evaluation testbed available, create the post-questionnaire, run the test and perform the data processing. B<>com was in charge of preparing the content (e.g., applying HDR reconstruction and TMOs) and modifying the testbed. The content selection, test methodology design (toggling) and results analysis was joint work.

From the subjective test design perspective, numerous recommendations are outcomes of this study.

- We observed the strengths to evaluate multiple factors for QoE evaluation. The richness of the analysis is increased by collecting and examining more precise, distinct and specified characteristics.
- We propose the following three steps process to extract influence factors to conduct a study

    1. *Review the literature and extract all or most common influence factors.* Industrial documents as well as not scientific reports may be of great help. For instance, when evaluating a system, it is current to include factors that were stressed in the advertisements of the system as marketing strategies highlight the strengths of the described solution. From my personal experience, I would advise you to trust your intuition to add new or atypical factors.
    2. *Select most relevant influence factors for your study.* This selection can rely on theoretical models. You may also need a preliminary or exploratory experiment to achieve the selection. An example is to present experiences to a reduced sample of the population to observe which characteristics are of interest. A pre-test, or pilot test, helps to improve the quality of data significantly by detecting problems of formulation and design [Aziz 2015]. Pre-test usually stands also for validation. Keep in mind that your context of study will also drastically impact which factors to include in the experiment.
    3. *Validate your questionnaire.* This process is usually time and resources consuming. A large sample of the population (more than 30 participants) may be required. Statistical models are used to assess the questionnaire reliability. Usually,

Cronbach's alpha [Cronbach 1951, Rosa 2015, Capraş 2017] and intra-class correlation coefficient [Bartko 1966] compare whether quantitative measurements within a same groups resemble each other.

- We recommend the use of the following influence factors in HDR imaging, in this order of importance: Details in the bright areas, Overall contrast, Naturalness of the scene, Details in the dark areas, Overall brightness, Unnatural colors, Noise, and Ghosting. We have seen that even the unnatural colors factor is crucial for the assessment.

- Subjects reported that in some cases, they could not assess contents properly as their preference was going towards test or reference depending on the lighting of the area. We thus deduce that subject could be allowed to provide a different assessment regarding which content is preferred in dark and bright areas.

- Our results illustrate the benefits to include aesthetic characteristics of contents to analyze subjects' behaviors further when assessing an experience. It should be noted that we have selected the aspects Boring, Interesting, Colorful, Aesthetic, Familiar, Of quality, Pleasurable and Immersive based on our context. Those aspects were compliant with our study aims and could be reused for any research. Though, we encourage researchers to add or remove factors depending on their study purposes.

- When conducting a test with HMD which needs an interface for interaction (e.g., toggling, scoring), we support the use of a controller. Indeed, to guarantee subjects comfort is a requisite. Severe bias can occur if fatigue leads to lack of attention of subjects.

- Subjective tests with omnidirectional contents are tiring due to possible sickness effect and the exploration of every stimulus. 20 minutes was a good duration for such format evaluation. Subjects stressed out that they were starting to be less focused on last evaluations. This means that about 40 pair comparison can be evaluated per test sessions when 360° contents are under study.

Overall, this work lays a basis for the future development of HDR imaging for omnidirectional representations. It also provides numerous outcomes for designing future experiments on HDR and omnidirectional contents.

Regarding HDR 360° contents, replication of our study on contents with higher dynamic range or HMD with embedded or mobile HDR display should be considered in the near future. Experiments may include the influence factors that we have here extracted for this format assessment. Also, the processing of toggling information can be further refined by incorporating an estimation of eye position or by optimizing the thresholding of head movement speeds, possibly per subjects.

4.11 Future works

Our work can inspire and be the basis of numerous other applications or evaluations.

First of all, a similar study may be conducted on professional or computer-generated contents. The combination of all studies would be representative of the HDR ecosystem and provide an exhaustive evaluation of HDR omnidirectional potential.

Second, different rendering could be used for HDR 360° contents. We have seen that Yang et al. [Yang 2012] and Mikamo et al. [Mikamo 2016] proposed to display two different tone-mapped images of the same HDR input image on each eye of a binocular display. When seen through a binocular display, the combination of images is richer and more detailed than single tone-mapped versions. This promising rendering approach could be evaluated with various TMOs to understand which ones are the best to pair. Also, new rendering strategies can be compared to this latter.

With regards to binocular display, it would be interesting to see if capturing stereoscopic HDR omnidirectional contents is beneficial. Capture systems enabling stereoscopic imaging are often more complicated than 2D information. Knowing that developed HDR 360° capture systems are already quite elaborate, stereoscopy adds complexity which is maybe not worth it.

If we continue our reasoning about stereoscopy, one could develop a binocular display rendering. Each eye would see a stereoscopic representations of the scene captured at different exposure. This could be assimilated as a 360° fusion exposure technique.

Lastly, our study can be extended to HDR 360° videos.

Multiple other opportunities and future works are possible after our work as this field is relatively new.

Chapter 5

User-centric influence factors for VR gaming

Contents

In-depth studies of QoE and users' satisfaction tend to include more and more influence factors regarding the evaluated technology and its specific features. In the following, we investigate concepts which are common to all current and emerging technologies. We believe that QoE is directly linked to previous experiences and expectations of viewers. In facts, expectations are the first influence factor named in the QoE definition.

Besides, the novelty effect, a bias introduced when experiencing a technology for the first times, is assumed to impact subjective QoE evaluations. For instance, when encountering 3D contents for the first times, 3D effects were pleasantly surprising. After some time, this effect faded out and headaches or discomfort ruined 3D experiences and their acceptability. The novelty effect causes opinion alteration about technology features that have not been experienced before.

In view to include factors related to the user mutual to all media experiences, we investigate whether introducing novelty effect and expectations in subjective evaluations help to reach advanced knowledge on QoE.

In [Borer 2015b], the novelty effect is related to improvements in new immersive multimedia technologies. The novelty effect is also closely associated with expectations. More precisely, the authors in [Parasuraman 1985, Zeithaml 1993] indicate that expectations depend on expertise and familiarity with a service. When evaluating emerging technologies, most consumers have little to no experience with the service or product under consideration. Unfamiliarity or inexperience with product leads to novelty effect. This explains that we investigate the novelty effect before examining expectations.

The following sections introduce state of the art on those two concepts and describe advancements provided by the conducted research. In details, we started with an exploratory evaluation of the novelty effect. Mainly mentioned in studies about emerging technologies, novelty fades out with time. The unknown duration of the effect makes it difficult to observe. There were thus three stages in this study: (1) to observe the novelty effect, (2) to analyze on which aspects of novelty influence subjects results, and (3) to conclude about the usefulness and reliability of evaluating novelty effect in future experiments.

Then, to my mind expectations are one of the most interesting and complex dimensions of QoE. Indeed, expectations come from cognitive processes, they form the basis of perceptual evaluations and give indications about subjects profile. It is an exhaustive aspect of QoE. However, few evaluations of QoE contain expectations appraisal. Thus, we investigated how to evaluate subject's expectations. We also inspected expectations variation between before, during and after the experiment. The observation of such information opens the way to understand further and analyze subjective scores. Even more interesting, it is possible to observe the impact of an experiment on subjects. Ultimately, indications of the relevance of using expectations in subjective evaluations are given.

Both studies were about VR gaming. This is an emerging technology that has drastically progressed over these last years. Regarded as one of the present hot topics, participants will have expectations based on hearsay, perhaps previous experiences, and mostly their expectations. This made this technology perfect for our subjects of study, namely novelty effect and expectations analyses.

5.1 Novelty effect

During the works performed on HDR, we encountered several studies that mentioned the novelty effect, which could threaten the external validity of subjective experiments. People are indeed subject to get more excited or more dubious regarding a service, technology or tool. This state of mind unconsciously impacts someone's evaluation.

Also referred to as "honeymoon" or "wow" effect, it is part of the everyday

vocabulary and is rarely defined in non-dedicated research. It is however described in several different fields such as psychology, education, and management.

In [Gravetter 2018], authors discussed the differences in novelty effect depending on the evaluation context. For instance, in psychology, the novelty effect is "the tendency for an individual to have the strongest stress response the first time that an individual is faced with a potentially threatening experience. Over time, as the novelty wears off, the stress response decreases" while in the context of human performance, novelty effect is characterized as "the performance to initially improve when new technology is instituted, not because of any actual improvement in learning or achievement, but in response to increased interest in the new technology"[Meline 2009]. The source (threatening experience or new technology), impact (stress response or performance increase) and valence (positive or negative) of the novelty effect are different.

Based on the study performed on novelty in [Blythe 1999], the valence of the novelty effect can be negative and positive. Besides, a new effect source is introduced, the relative advantage (degree to which technology improves former technologies). The authors recalled that the novelty effect is highly subjective. It can be expressed in various behaviors and, therefore, must be thoughtfully defined and quantified. In [Wells 2010], authors agree with previous definitions of novelty effect whereas they indicate that excitement and various salient beliefs about the technology are causing the effect: "The perceived novelty of an information technology innovation is considered to be the degree to which a user perceives an innovation to be a new and exciting alternative to an existing technology that can vary by individual and, subsequently, should be investigated as a potential predictor of adoption".

In education, several studies compared a group who is using new instructional methods to an untreated control group. It appeared that excitement and enthusiasm among subjects affected the experimental group's response [Ary 2018]. In [Bracht 1968], enthusiasm and disruption caused the novelty effect. Disruption occurred when encountering a new and unfamiliar treatment.

In a different context, Love et al. [Love 2013] investigated the attendance at sports events when building a new stadium. Their research focuses on the duration of the novelty effect as attendances increases in the first years of the stadium opening to eventually reach the original attendance quota.

On online selling websites, novelty is "defined as the lack of experience of individuals in the organization with similar purchase situations" [McQuiston 1989].

To summarize the above,

- a careful analysis of definitions of novelty effect in state of the art revealed that only a few studies addressed the time span during which novelty effect remains dominant [Gravetter 2018].
- Although its subjectivity and variation across subjects were discussed, the fact that subjects are not fully conscious of this effect was rarely mentioned [Wells 2010].

- A positive valence effect was often expected [Meline 2009, Ary 2018, Gravetter 2018] but the impact of novelty could be positive or negative [Blythe 1999]. In [Bracht 1968] a negative novelty effect was referred to as a disruptive effect. We did not distinguish the two notions as the value of valence can discriminate them from each other.

- There were divergences regarding what causes the novelty effect. It is arguable to state that novelty effect only comes from an increased interest level [Meline 2009], as it seems important to incorporate affective reactions [Blythe 1999], salient beliefs and perceived usefulness [Wells 2010], lack of experience and unfamiliarity [McQuiston 1989] or strong stress response [Gravetter 2018].

- All definitions examined in the above were restricted to the exact underlying application, while a broader definition is probably more desirable.

Accordingly, we proposed a new definition of the novelty effect, defined which influence factors are to be evaluated, and conducted the experiment. We observed and analyzed the novelty effect and concluded about it.

5.1.1 Definition

To cope with issues mentioned above, we propose the following definition for novelty effect.

The novelty effect is the transitory and unconscious tendency of subjects to evaluate their experience differently the first times they are using a product, not because of the product intrinsic features, but because the user perceives some of those features as unfamiliar.

This definition

- reflects the instability and unknown duration of the effect,

- emphasizes that users are unaware of it,

- notifies that the effect can have a positive or a negative influence on the experience,

- indicates that its occurrences are not necessarily related to time (e.g., release date of a product),

- is broad enough to refer to experiences when using applications, technologies, services or devices,

- expresses explicitly that this effect occurs not due to features quality or efficiency but due to their unfamiliarity,

- points out the subjectivity of this bias,

- explains that the modification of at least one product feature can result in a novelty effect, and
- specifies the source of the effect.

5.1.2 Influencing factors

To conduct an in-depth analysis of the novelty effect, different factors influencing or being influenced by this effect must be characterized. From the literature, especially [Blythe 1999, Wells 2010], and previously mentioned state-of-the-art definitions, we have identified seven main influencing factors.

1. **Interest**: willingness to be involved with and to discover more about a product or a service. This factor is not related to the curiosity of trying (if one has that opportunity), which is discussed under another aspect called the current mood.
2. **Experience**: experience gained from using a product or a service in a specific usage context.
3. **Features**: intrinsic features present in a product or a service, independent from their usage context.
4. **Expectations**: appraisal prediction of inherent characteristics of a product or a service, independent of the usage context and based on prior knowledge and experience.
5. **Relative advantage**: comparison between a product or service and another similar product or service.
6. **Conscious newness**: degree to which a person consciously feels that a product or a service is unfamiliar.
7. **Current mood**: user's state of mind.

To avoid any misconceptions, these influence factors impact or are impacted by experiences regarding novelty. Such influence factors may or may not be the same for the evaluation of QoE.

5.1.3 Exploratory experiment

As emphasized in the introduction, the conducted evaluation targetted three aims (1) to observe the novelty effect, (2) to analyze which aspects of novelty influence subjects results, and (3) to conclude about the usefulness and reliability of evaluating novelty effect in future experiments.

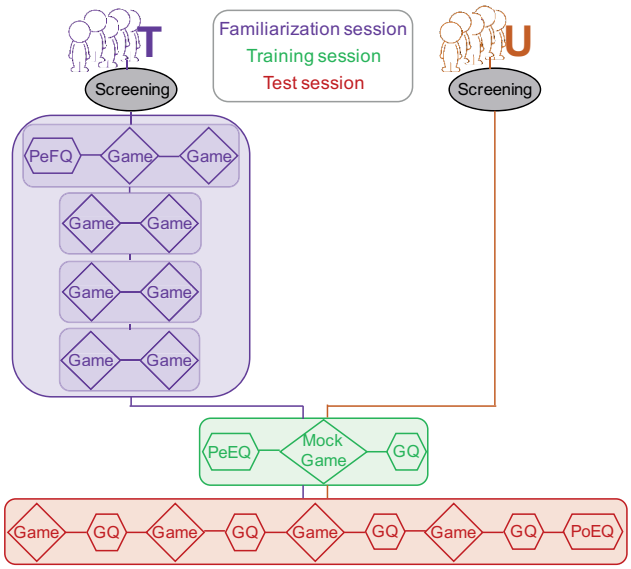


Figure 5.1: Procedure used in the exploratory experiment

5.1.3.1 Equipment

VR gaming offers the opportunity to study relatively new technologies and hopefully reach a population that never has encountered it. VR gaming also presents clear unfamiliar capabilities when compared to typical gaming experiences. For instance, it is rare to control games actions or cameras with head movements or to visualize streams on a head-up display. We took advantage of the newness and unusual capabilities of VR gaming

A Samsung Oculus Gear along with a Samsung Galaxy S6 constituted the VR gaming platform. The audio sound systems were over-ear Sennheiser headphones. When needed, a Thrustmaster controller was used.

The test room provided a controlled, calm and neutral environment, fulfilling multimedia tests requirements.

5.1.3.2 Evaluation methodology

To verify if the novelty effect can be observed and quantified, two populations were required. Originating from the same population (O), subjects were split into two sets. One set was untrained (U) and hypothetically impacted by the novelty effect. The second, referred to as trained set (T), went through Familiarization Sessions (FSs), designed to reduce the novelty effect.

Population O was composed of individuals with no experience with VR, especially with HMDs. This prevented introducing any bias in the experiment due to discrepancies between subjects level of familiarity with VR or HMD.

U and T populations were screened for correct visual acuity and color vision using Snellen and Ishihara charts, respectively.

Before FSs, subjects from T completed a first questionnaire, the Pre-Familiarization Questionnaire (PeFQ). They took part in four FSs, during which they experienced eight VR games (two per session), of 15 minutes each. At least a day separated two FSs.

From this point on, both populations followed the same experimental procedures. Subjects had to fill a questionnaire, the Pre-Experiment Questionnaire (PeEQ) in advance of the experiment. The PeFQ and PeEQ consisted in the same questionnaire. PeFQ notation was introduced to differentiate population T answers to PeEQ questions before and after FSs.

A training session was organized consisting of a 5-minute VR gaming experience followed by filling out the Game Questionnaire (GQ). This allowed subjects to get familiar with test and evaluation procedures, while population U was experiencing VR gaming for the first time.

The test session involved the presentation of a 10-minute VR gaming experience followed by the GQ filling, repeated four times. At last, subjects were required to fill in a Post-Experiment Questionnaire (PoEQ). Figure 5.1 illustrates the procedure described above.

The number and duration of gaming experiences in training and test sessions were defined based on timing constraints given in subjective tests recommendations by the ITU. Training sessions were designed not to exceed 45 minutes when completing the PeFQ and 30 minutes otherwise. Test session includes the filling of both pre- and post-questionnaires, training, and test sessions. Accordingly, four game-sessions were included for the actual assessment. This guaranteed a duration of less than an hour and a half for the entire experiment (apart from FSs).

Factors of gaming experiences were investigated through 9 questions. Seven of which were extracted from dimensions of the widely used Game Experience Questionnaire (GEQ) (competence, immersion, flow, tension, challenge, negative and positive affect) [IJsselsteijn 2008]. Due to the VR environment, a question on sickness [Hwang 2006], as well as a question on controls intuitiveness, were also added as they could impact the novelty effect.

Interest and conscious newness were composed of 3 questions each. Interest went through subjects' likelihood to read an article or do research on VR as well as their interest level in VR. Conscious newness measured the extent to which subjects perceived VR as familiar and new (from the gaming market and subjects perspectives).

Current mood was evaluated as a combination of energy and valence, following the Miller mood map [Miller 2009]. Similarly to the previous chapter, a more precise model is not needed because of the context of our study.

As proposed in the previous chapter on the extraction of features, we relied on HMD constructors documents to identify the technology features [1]. VR HMD established features impacted by or impacting the novelty effect were: visual, auditory, interface/usability, presence/immersion, distraction, enjoyment, cyber-sickness, Field

[1] http://www.samsung.com/global/galaxy/gear-vr/specs/

Of View (FOV) and HMD comfort and ease of use. 17 questions were created, inspired mainly by [Hwang 2006] Presence/Usability questionnaire.

Expectations questions consisted of reformulated questions about features mentioned above while relative advantage questions is a subset of features questions about visual quality, presence, enjoyment and intuitiveness of controls.

Every question was evaluated on a 5-point rating scale. This scale is sufficient regarding the fact that we wanted to observe the novelty effect, and then possibly analyze scores.

Table 5.1 summarizes how many questions and which influence factors were assessed in questionnaires. Remaining questionnaires details are presented below. The Miller mood map representation used in the questionnaire is illustrated in Figure 5.2. Dimensions evaluated for expectations, relative advantage and features are presented in Table 5.2 while these of experience are shown in 5.3. Table 5.4 reports all the investigated questions along with the factor that is assessed and the scale used.

Influence factors	PeFQ / PeEQ	GQ	PoEQ
Interest	3	-	-
Experience	-	9	-
Features	-	-	17
Expectations	17	-	17
Relative advantage	-	-	4
Conscious newness	3	-	-
Current mood	2	-	2
Novelty effect	-	-	1

Table 5.1: Number of questions for each influencing factor

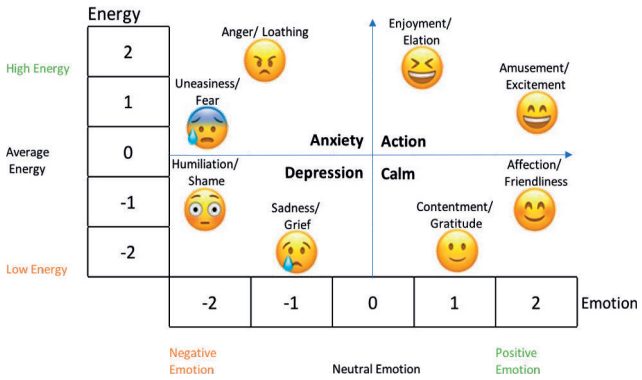


Figure 5.2: Miller mood map [Miller 2009]

Regarding participants, ten bachelor and master students composed the tested population. This makes five people per population set. Population O, T, and U were 22.3, 24.2 and 20.4 years old on average with a standard deviation of 3.30, 3.76 and 1.14, respectively. One woman was part of the T population while two were

Category	Number of expectation questions	Number of features questions	Number of relative advantage questions
Visual	4	4	1
Auditory	2	2	0
Interface/ Usability	4	4	1
Presence/ Immersion	1	1	1
Distraction	1	1	0
Enjoyment	1	1	1
Cyber-sickness	1	1	0
FOV	2	2	0
HMD	2	2	0

Table 5.2: Expectations, Relative advantage, and Features questions

Dimensions
Interface/ Usability
Cyber-sickness
Competence
Sensory and Imaginative Immersion
Flow
Tension/ Annoyance
Challenge
Negative Affect
Positive Affect

Table 5.3: Experience dimensions

Questions	Component	Scale 1	Scale 2	Scale 3	Scale 4	Scale 5
Situate your level of energy on the vertical scale (circle one number ranging from -2 to 2)	CM	-2	-1	0	1	2
Situate your emotional state on the horizontal scale (circle one number ranging from -2 to 2)	CM	-2	-1	0	1	2
If you come across an article about virtual reality, how likely are you to read it?	IT	Very unlikely	Unlikely	Neutral	Likely	Very likely
How often do you make researches about virtual reality on your own?	IT	Never	Rarely	Occasionally	Frequently	Very frequently
Do you consider yourself as passionate about virtual reality?	IT	Not at all	.	Moderately	.	A lot
When you wanted to perform an action in the game, did you have to think for a while about the way to achieve it through the HMD?	EP	Not at all	.	Moderately	.	A lot
Did you feel sick navigating through the game?	EP	Not at all	.	Moderately	.	A lot
Did you feel skillful?	EP	Not at all	.	Moderately	.	A lot
Did you find the game to be a rich experience?	EP	Not at all	.	Moderately	.	A lot
Did you forget everything around you?	EP	Not at all	.	Moderately	.	A lot
Did you feel frustrated?	EP	Not at all	.	Moderately	.	A lot
Was the difficulty level too high?	EP	Not at all	.	Moderately	.	A lot
Did you feel bored?	EP	Not at all	.	Moderately	.	A lot
Did you enjoy the game?	EP	Not at all	.	Moderately	.	A lot
Will the motions on the screen be fluid?	EX	Not at all	.	Moderately	.	A lot
To your mind, will the depth perception be good?	EX	Not at all	.	Moderately	.	A lot
Do you think that the screen resolution will be good?	EX	Not at all	.	Moderately	.	A lot
Do you expect to feel visually isolated from the ouside world?	EX	Not at all	.	Moderately	.	A lot
Do you expect to feel acoustically isolated from the ouside world?	EX	Not at all	.	Moderately	.	A lot
Is the sound going to help you feel like being in the game?	EX	Not at all	.	Moderately	.	A lot
Is the gyro sensor going to be a reliable measuring tool?	EX	Not at all	.	Moderately	.	A lot
Do you expect to have trouble using the controller?	EX	Not at all	.	Moderately	.	A lot
Is it going to be easy to control the virtual environment?	EX	Not at all	.	Moderately	.	A lot
To your mind, is it going to be easy to understand how to use the HDM?	EX	Not at all	.	Moderately	.	A lot
Do you expect to feel like being in the virtual environment?	EX	Not at all	.	Moderately	.	A lot
Do yo think that the HMD can distract you from the game?	EX	Not at all	.	Moderately	.	A lot
Will the navigation through the environment be enjoyable?	EX	Not at all	.	Moderately	.	A lot
Do you expect to feel sick navigating through the virtual environment?	EX	Not at all	.	Moderately	.	A lot
According to you, is the field of view to be sufficient for navigating through the virtual environment?	EX	Not at all	.	Moderately	.	A lot
Do you expect te HMD to be comfortable to wear?	EX	Not at all	.	Moderately	.	A lot
Do you think that the weight of the HMD will bother you?	EX	Not at all	.	Moderately	.	A lot
How likely is virtual reality to replace all other ways of gaming?	EX	Not at all likely	Slightly likely	Moderately likely	Very likely	Completely likely
Which device has the best visual quality? (TGD: typical gaming device)	RA	The oculus gear, definitely	The oculus gear, slightly	No difference	My TGD, slightly	My TGD, definitely
Which device makes you feel the most present in the virtual environment?	RA	The oculus gear, definitely	The oculus gear, slightly	No difference	My TGD, slightly	My TGD, definitely
With which device do you enjoy the virtual environment the most?	RA	The oculus gear, definitely	The oculus gear, slightly	No difference	My TGD, slightly	My TGD, definitely
With which device do you have most trouble controlling the virtual environment?	RA	The oculus gear, definitely	The oculus gear, slightly	No difference	My TGD, slightly	My TGD, definitely
Playstation, Wii, smartphones, PC or Xbox are common gaming devices today. What would you say about virtual reality?	NW	Not common at all	Slightly common	Moderately common	Very commun	Completely common
Were do you situate virtual reality in the history of gaming?	NW	Not at all new	Slightly new	Moderately new	Very new	Completely new
Does virtual reality stand out from your everyday experiences?	NW	Not at all	.	Moderately	.	A lot
Were the motions on the screen fluid?	FA	Not at all	.	Moderately	.	A lot
Did you find the depth perception to be good?	FA	Not at all	.	Moderately	.	A lot
Was the screen resolution good?	FA	Not at all	.	Moderately	.	A lot
Did you feel visually isolated from the ouside world?	FA	Not at all	.	Moderately	.	A lot
Did you feel acoustically isolated from the ouside world?	FA	Not at all	.	Moderately	.	A lot
Did the sound help you feel like being in the virtual environment?	FA	Not at all	.	Moderately	.	A lot
Was the gyro sensor reliable?	FA	Not at all	.	Moderately	.	A lot
Did you have trouble using the controller?	FA	Not at all	.	Moderately	.	A lot
Did you have trouble controlling the virtual environment?	FA	Not at all	.	Moderately	.	A lot
Was is easy to understand how to use the HMD?	FA	Not at all	.	Moderately	.	A lot
Did you feel present in the virtual environment?	FA	Not at all	.	Moderately	.	A lot
Were you distracted from your task in the virtual environment by the HMD new features?	FA	Not at all	.	Moderately	.	A lot
Did you enjoy navigating through the virtual environment?	FA	Not at all	.	Moderately	.	A lot
Did you felt sick navigating through the virtual environment?	FA	Not at all	.	Moderately	.	A lot
Was the field of view to be sufficient for navigating through the virtual environment?	FA	Not at all	.	Moderately	.	A lot
Was the HMD comfortable to wear?	FA	Not at all	.	Moderately	.	A lot
Did the weight of the HMD bother you?	FA	Not at all	.	Moderately	.	A lot
Do you think that at this stage of the experiment the novelty effect is (still) present?		Not at all	.	Moderately	.	A lot

Table 5.4: Novelty questions. The evaluated influence factors relate to current mood (CM), interest (IT), experience (EP), expectations (EX), relative advantage (RA), Conscious newness (NW), and features (FA). Questions highlighted in gray are negatively phrased

in U population. Before the experiment, subjects had never experienced VR. We verified subjects' color vision, and visual acuity made sure they are not prone to epileptic seizures. Additionally, any participant fluently spoke English.

5.1.3.3 Test material

A total of thirteen games were used during this experiment, eight for FSs, four during the test session and 1 for the training session. They have been sorted out as exploration, racing, shooter or strategy games to balance games quality and types between FSs and test sessions. This fact is shown in Figure 5.3. Test session games are denoted as $HeroBoundFirstSteps$, $405RoadRage$, $DuckHunter$ and $DefenseGrid$. The games presentation order was randomized to remove any ordering bias.

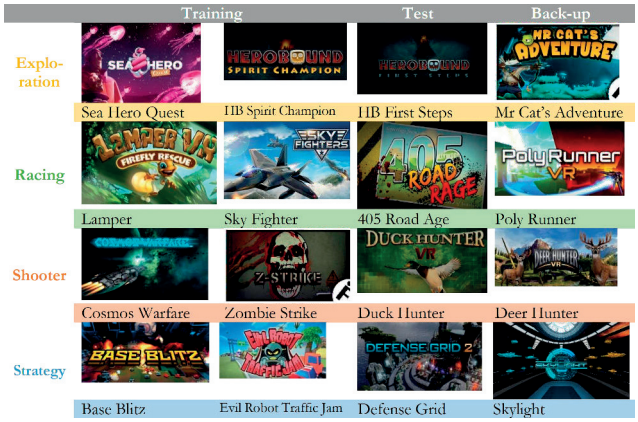


Figure 5.3: Test material

5.1.4 Results

Results were processed as a normalized sum of questions scores per influencing factor. Some items had to be pre-processed due to their negative formulation.

MOSs and corresponding 95% CIs were computed. Differences between scores distributions statistical significance were also investigated, using 1-way ANOVA. A p-value of 0.05 was selected to reject the null hypothesis.

First, a Sanity Check (SC) was conducted to verify that U and T subjects came from a same original population O. If so, ratings of PeFQT and PeEQU should come from the same distribution.

Then, the Intra-Group Analysis (IaG) verified the incidence of FSs on T scores.

Finally, the Inter-Group verification (IeG) indicated if a novelty effect was observed.

1. From SC, we observed a decrease in influencing factors perception from U population when compared to T. However, no statistically significant differences were obtained between these populations. This indicates our two samples are coming from the same population O.

With regards to subjects number in our analysis, to analyze further our results, we considered p-values under 10%. Under this condition, populations interest in VR was statistically different. This indicated that in future experiments, subjects should not choose which population they will be part of, due to a slight bias introduced by their interest in VR.

2. During IaG analysis, interest in VR increased while expectations decreased when comparing scores before and after familiarization and test sessions. It should be noted that T population interest in VR got stable during the test session.

   No statistically significant differences were obtained in this analysis. This indicates that FSs did not influence scores of the T set, based on used questionnaires. A more substantial population sample should be used to validate the FSs procedure.

3. When comparing scores of both populations, CIs are quite large due to the limited number of subjects per population. Nevertheless, expectations CIs are pretty narrow. This surprising finding possibly indicated common prior knowledge of VR gaming to all participants.

   A statistically significant difference between the populations' expectations was observed before the test session ($p = 0.04$). T expects higher experiences with VR than U. This fact supported the idea of bias in subjects allocation in U and T groups or emphasizes the expectations impact on the novelty effect.

   Overall, games ratings given by the U population were higher than the scores given by the T set, but not significantly. When performing ANOVA on influence factors sub-questions, frustration, and boredom were substantially lower in U population for Defense Grid game, as well as enjoyment for Duck Hunter game.

   The same reason can explain these two different behaviors. Participants U were not familiarized with the type of games designed for VR. In the case of Defense Grid, they were less annoyed by the 2D design of the game. Regarding Duck Hunter, subjects U were probably not enjoying as much as participants T due to their lack of expertise with head control.

The sanity check validated that our sets U and T derives from the same population O.

However, one must be careful when sampling population, not to bias the test by distributing participant based on their motivation.

Even though FS seems not to impact the population T, expectations of U and T sets were significantly different. When compared to U, subjects in T expressed higher expectations. This is the only influence factor that showed to be different.

5.1.5 Discussion

This exploratory test was subject to numerous constraints. Many points must be discussed before stating any conclusions.

With respect to the number of subjects, running the test requires at least 50 hours. This justifies that only ten subjects took part in the experiment, knowing that we were conducting a pilot test. This means that our observations must be processed with regards to the lack of external validity.

Gender balance was ensured in one population sample and there wes a male/female ratio of 80/20 % in the second. Having both sets balanced would have been better for two reasons, first because of possible gender bias, second because of the gender balance in gamers population [2]. We assumed here that gender bias was not significantly expressed in the collected data.

Concerning the sampling of population O in sets U and T, current practice are to assign most dedicated subjects to T set as the process is long and it is a loss of time and resources if a subject gives up. However, as our results indicated it, such sampling may influence the results and prevents the observation of an effect. Regarding the external validity concern, it seems reasonable to generalize this finding to all subjective tests.

The lack of proof regarding the influence of FS on T set may have several sources. First, possibly no novelty effect has been experienced. We can not observe an effect which is not present. This speculation is not highly likely as VR gaming presents numerous features subject to the novelty effect. Second, The number of FS sessions or gaming durations may not be sufficient for the novelty effect to wear off. Further experimentations need to explore this prospect.

We observed that CIs of expectations were less wide than for other factors. This would indicate that the tested population encounters similar information regarding VR. Accordingly, not only master student must be included in experiments. Population sample has to be more diverse.

Last but not least, the only significant difference between U and T observed in the experiment is about the factor expectations. Regarding the difficulty to design a test to observe the novelty effect, in the most valid way possible, it seems a lot of energy only to see variation of expectations.

The unknown duration of the novelty effect put the construct validity of the experiment design in jeopardy. Determining this duration is complicated if not impossible. Novelty effect is highly variant across individuals, is not apparent and shows no indisputable indication of its disappearance. Moreover, such information is of relative interest for the industry. Indeed, if marketer may found an interest in such information, numerous marketing models already take into account the novelty effect in their development. The Amara's Law [Amara 1960] and Gartner's Hype cycle [Linden 2003] perfectly show the pattern of over-enthusiasm and disillusionment that people goes through when experiencing novelty and also when this effect wears off.

[2] https://www.statista.com/statistics/232383/gender-split-of-us-computer-and-video-gamers/

For those reasons, it seems a better approach to evaluate subjects expectations and not precisely evaluates the novelty effect.

5.1.6 Conclusion

In this work, we proposed a new definition for the new novelty effect in multimedia applications, based on various fields state-of-the-art definitions such as management, marketing, education, and psychology.

We followed our recommendations for the identification of influence factors, extracted from VR HMD constructors specifications. That is, we evaluated interest in VR, gaming experience quality, VR HMD features, expectations about technological features, predicted relative advantage of VR gaming over typical gaming, and consciousness of newness.

We conducted an exploratory experiment which aimed to observe the novelty effect. In the event it is be seen, we aimed to understand which aspects of the novelty effect influences subjects results.

We designed our test as a comparison between two populations, one under the influence of the novelty effect, but not the other. That is, we compared scores of two population sets T and U. Originating from the same population of complete naive gamers in VR, people in set T were trained in various FSs for the novelty effect to wear off.

Five subjects took part in each set and were 22.3 years old on average.

No significant impact of FSs on population T was observed. This may due to the design of FSs (e.g., duration, number and variety of experiences) that did not cause the novelty effect to wear off. It can also be because no novelty effect was experienced.

No novelty effect was observed in a significant way except in variation of expectations. Population T expressed higher expectations regarding VR technical capabilities.

Some findings of this study were beneficial for our works and future explorations of QoE. Firstly, when sampling populations, it is better not to let subjects choose which population set they will be part of. This choice often reflects their involvement level. That is, subjects interest in the technology may bias results, especially if the design aims at a direct comparison of populations responses. Secondly, experimenters must make sure that subjects expectations regarding the evaluated system are various. Otherwise, the external validity of the experiment may be threatened.

This pilot test showed that it is complex to observe the novelty effect. But most importantly, efforts made to do so may be worthless. Indeed, based on our results, there were only discrepancies in expectations between groups. We thus settled for focusing on assessing precisely expectations instead of observing the novelty effect.

5.2 Expectations

We are interested here in the cognitive process of perceptual quality creation. We are precisely focusing on expectations, which are widely acknowledged as the source of QoE construction. Extensive research on expectations has been conducted in market studies when evaluating the quality of service, especially acceptability, usability, and reliability. On the contrary, scarce quantitative and qualitative analyses have been performed on multimedia QoE. Before conducting our evaluation, we reviewed important research matching our interests.

5.2.1 Definition

Not meeting consumer expectations typically results in dissatisfaction [Parasuraman 1985]. Satisfaction, however, is not strictly speaking the inverse of dissatisfaction. A service can be appraised as acceptable while not entirely fulfilling user requirements, leading to the incomplete satisfaction of consumers. Notions of dissatisfaction, acceptability, and satisfaction are linked to users' expectations: "expectations are important concepts because they form the frame of reference for satisfaction judgments" [Higgs 2005].

In the field of market analyses, expectations have played a considerable role since decades. Expectations have been empirically assessed and included in service quality models. Parasuraman [Parasuraman 2002, Parasuraman 1994, Zeithaml 1993] have done extensive research on expectations when designing the multiple-item scale for measuring consumers perceptions of service quality (SERVQUAL). He argues that "perceived quality is viewed as the degree and direction of discrepancy between consumers' perceptions and expectations" and defines expectations as "desires or wants of consumers, i.e., what they feel a service provider should offer rather than would offer".

Even though the proposed service quality modeling is questioned and discussed in numerous works [Teas 1993, Parasuraman 1994], most research, and experimental works endorse the satisfaction $expectation - perception$ (E-P) model, also named gap model [Parasuraman 1985]. This model indicates if the service has reached or exceeded customers' expectations.

Most disagreements come from the definition of expectations. In the above description, vague terms such as "should" or "desires or want" leave room for interpretation. This definition is not compliant with the complexity of the expectations concept. For instance, expectations result from different sources (e.g., previous experiences, cultural bias or advertisement) and are formed based on different service representation (e.g., ideal, desired or minimum acceptable service).

From the literature, we can distinguish four expectations categories: forecast, normative, ideal and minimum tolerable [Higgs 2005]. **Forecast expectations** are also known as expected or predicted expectations. They stand for consumers' beliefs about what will occur when encountering a service. **Normative expectations**, also named deserved or occasionally desired expectations, measure feasible and realistic

services that should be delivered. **Ideal expectations** refer to the maximum service performance reachable in one's opinion. **Minimum tolerable or adequate expectations** express the minimum service effectiveness that is acceptable from the consumer's perspective.

It is worth noting that forecast and normative expectations are specifically related to the considered service, while ideal and minimum tolerable expectations are broader constructs, considering a more extensive range of services. Also, by definition, respondents must have been at least once exposed to a service to construct ideal and minimum expectations. Yet, many empirical studies suggested that with only limited or no past experiences, one can form expectations.

Overall, from the marketing literature, one can understand that a lot of work has been done, though no standard study design or expectations definition has been widely accepted and used.

In the state of the art of multimedia quality and QoE assessment, despite few studies were dedicated to expectations, there are huge discrepancies from one work to another. Some works are in contradiction or present a lack of consistency with the state of the art. The research studies given in the following paragraph are works on which we relied on to design our evaluation.

Authors of [Perkis 2006] designed a QoE model including expectations as non-measurable entities. They established a correlation between attitude/ behavior and (not) meeting expectations. However, at the same time, they claim that multimedia services quality is less related to expectations than actual perception. They based their claim on the work performed by [Houghton 2005]. I have to disagree with their claim, as to my mind, they have inadvertently improperly generalized Houghton's findings.

Houghton indicated that E-P gap is measuring the opportunity to exceed expectations, which is a relevant way to investigate quality, as it is less relying on what people say and more on what people think. Presented findings even show that the perceived importance of influence factors (reliability, usability, and fun) is less relevant for qualitative assessment than perception-expectation differences.

In [Sackl 2012], authors do not define expectations but intend to observe them. They have although noted the eclectic use and complexity of the concept. To cope with this fact, they implicitly measured services expectations through a system providing a same delivery quality level, yet indicating wireline or wireless Internet connexion by means of a fake router. While the delivery offers the same service quality through the same wireline transmission, subjects think they are using different delivery connections. The conducted subjective evaluation is designed in such a way that participants are asked to manually change the router configuration from wireline to wireless connection or the reverse. Two services use cases are investigated: web surfing and video streaming. Wireless connexion is assessed as having a higher connexion speed, except for the streaming of poor quality video.

From the 49 subjects, only six people rated web services as same, while no

statistically significant differences are observed among grades in video streaming context. This finding would indicate expectations influence ratings in web services, even though not in video streaming. This approach is highly unusual. Authors did not ask subjects about their expectations but studied their behavior as suggested in [Manski 2004]. However, this practice is unfortunately not applicable to all services and could be impacted by an experiment bias. This effect and other validity concerns have been discussed in a similar work investigating users behavior in web video services [Robitza 2016].

In the continuity of their work on expectations, [Sackl 2014] measured users expectations in pre-questionnaires that participants fill in, in advance of the experiment. In two different studies, they investigated normative and minimum tolerable expectations. Findings show promising results as expectations-based QoE model reached higher MOSs prediction accuracies than expectations-free models. Significant advice provided in this study is that definition and evaluation of expectations should be tangible and concrete to gather meaningful information. In their study context, expectations were appraised on a list of relevant features.

Many works, in marketing studies and QoE research, considered only a subset of expectation categories. SERVQUAL implemented expectations regarding an ideal service [Parasuraman 2002], while [Higgs 2005] considered forecast expectations as "true" expectations. The model developed in [Zeithaml 1993] takes into account normative, minimum tolerable and forecast expectations. The drawn conclusions indicated the importance of normative and minimum tolerable expectations. It also indicates that adequate service expectations are the most likely to evolve.

During reviews of state of the art, the dynamic nature of expectations caught our attention. During subjective evaluations, expectations fluctuate in time due to the experience. An in-depth analysis of expectations variation could indicate users QoE. However, each expectations category is more or less prone to be modified. Ideal expectations, for instance, are more likely to stay stable over time, while forecast expectations can evolve after every single experience.

From the above, one learns that there are four expectations types, each of them is representative of specific service characteristics. Expectations are setting the reference to which the service experience will be compared. This reference can be one's opinion about what the service will or should offer, or what is the minimum acceptable or the highest service representation. It is not always necessary to evaluate all expectations types, as long as there is a technically sound reason to justify this practice. When subjectively investigating an expectation type, one should formulate precise, concrete and tangible questions to gather meaningful information. Expectations are processed through gap models: the evaluation of expectations is usually compared to another measure. Those measures, to name but a few, can be the perceptual quality, or revised expectations. The perceptual quality refers to the actual experience assessment, while revised expectations are subject's expectations after the experience.

5.2.2 Cross comparison study between VR and typical gaming platforms

An attempt to include expectation evaluation in QoE assessment has been carried out in the context of VR gaming. VR gaming is an emerging technology designed to provide richer experiences, especially regarding presence and immersion. This is an adequate use case to test new subjective evaluations such as expectations assessment. Our work also focused on how to find a balance between (1) thorough evaluation, taking the risk to get biased results by subjects' tiredness and lack of attention after answering lots of questions, and (2) single-dimensional evaluation, possibly inconclusive because of shallow insights extracted from subjects' answers.

The study is presented here and has been published in [Perrin 2017b]. This work has been performed in the framework of Marie Sklodowska-Curie Action QoE-NET, grant agreement No 643072, during a close partnership with the Quality and Usability Lab (QUL) in Deutsche Telekom (DT) Laboratories.

Tremendous progress has been made regarding the development of HMDs, leading to the growth of market share dedicated to VR in the gaming field. To explore further the capabilities of VR gaming, we inspected the benefits of this new gaming platform compared to conventional gaming devices.

Recently, high-quality HMDs with embedded displays, such as Oculus Rift DK2, HTC Vive or Playstation VR, were introduced on the consumer market. However, their prices seem to exceed the budget of many players, and the number of available games on such platforms is still limited. Affordable solutions, such as Google cardboard, Merge VR headset or Oculus Gear, permit the use of a mobile phone as the display. Due to the high quality of current mobile devices, the resolution of these systems equals or even improves that of embedded displays. However, the refreshing rate and field of view are slightly reduced. The reasonable price of mobile HMD, makes it more widely used.

Regarding gaming experiences, extensive research has been conducted to evaluate them, leading to numerous gaming questionnaires with multiple variables. On the contrary to what has been done in multimedia (evaluating perceived quality with a single question), gaming questionnaires include various dimensions such as flow, competency, degree of control, involvement, aesthetics, novelty, positive and negative affect [Möller 2013].

A higher sense of presence is expected from VR platforms when compared to computer and mobile platforms [Slater 1997]. It makes presence a mandatory dimension to evaluate during comparison across platforms. In addition to presence, another dimension influencing cross-platform assessment is the way to control games, as controls are varying widely from platform to platform. For instance, computers are controlled with mouse and keyboard; mobile phone is used through touchpad or Bluetooth controller; HMDs can rely on head movements and a hand-held controller. Additionally, an experience can drastically fluctuate because of controls intuitiveness. Hence, we believe that two essential dimensions of comparison across platforms are controls and presence.

	IPQ [Schubert 2001]	**PrQ** [Witmer 1998]	**PENS** [Rigby 2007]	**UEQ** [Laugwitz 2008]	**IEQ** [Jennett 2010]	**GEQ** [IJsselsteijn 2013]
Evaluated dimensions	• Spatial Presence • Quality of immersion • Involvement • Drama • Interface awareness • Exploration of VE • Predictability and Interaction • Realness	• Realism • Possibility to act • Quality to act • Quality of interface • Possibility to examine • Self-evaluation of performance • Sounds • Haptic	• Competence • Autonomy • Intuitive control • Presence/ Immersion	• Attractiveness • Perspicuity • Efficiency • Dependability • Stimulation • Novelty	• Cognitive involvement • Emotional involvement • Real world dissociation • Challenge • Control	• Competence • Sensory and Imaginative immersion • Flow • Tension/ Annoyance • Negative and Positive affect
Number of items	14	19 (or 24 with evaluation of sound and touch)	18	26	32	33 (Core), 14 (In-game)

Table 5.5: Comparison of state-of-the-art questionnaires

State-of-the-art gaming and presence questionnaires were compared in terms of evaluated dimensions and number of items in Table 5.5. We observed that most widely used gaming questionnaires, namely User Experience Questionnaire (UEQ) [Laugwitz 2008], Immersive Experience Questionnaire (IEQ) [Jennett 2010] and GEQ [IJsselsteijn 2013], investigate numerous questions. However, they are not assessing presence but immersion.

Presence questionnaires, such as IPQ [Schubert 2001] and Presence Questionnaire (PrQ) [Witmer 1998], are not evaluating important gaming aspects such as control.

To be able to select questionnaires for subjective evaluations, two studies compared individual assessments of questionnaires to metacritics scores [Johnson 2014, Lau 2015]. They have demonstrated that PENS [Rigby 2007] achieved the highest accuracy over the other evaluated questionnaires (GEQ and IEQ). Also, presence dimension is assessed in this questionnaire. Hence, PENS seems adequate for a cross-platform comparison for gaming experiences that includes VR HMDs.

The questionnaire is composed of four dimensions, namely *Competence*, *Autonomy*, *Intuitive Control* and *Presence/Immersion*. Competence is the feeling of being in control and excelling while doing a game action. Autonomy refers to the level of influence that player's actions have on the game story. Intuitive controls measure the degree of convenience of the set of commands that represent game actions. The notions of presence and immersion have been introduced in Section 2.2.3.1. The PENS also includes a *Relatedness* dimension, which implies interactions with other users. As the test comprised only single-player experiences, relatedness was excluded from the questionnaire.

Concerns about participants' boredom and lack of attention, due to numerous items in questionnaires, have been raised in [Jennett 2008]. Gaming core questionnaires rarely include less than 20 questions. Several researchers claimed that many dimensions could be assessed with fewer items. The concept behind this is to remove dependent questions and keep most independent items. Consequently, less precise answers are collected but in some cases these pieces of information were useless for the data analyses. Additionally, to diminish the number of questions is to decrease rating times. Such practices are particularly indicated in subjective evaluation as long as it does not prevent a proper and valid analysis of data.

Authors of [Jennett 2008] proposed a solution to reduce inter-relations between gaming dimensions and reduce the number of items in questionnaires. They reported a strong correlation between a single overall question and a set of questions about one dimension, namely immersion. Hence, it is valid to inquire a single overall question about immersion. This study process could be extended, and other important dimensions number of items in questionnaires could be reduced in a similar way.

Previous works that have investigated VR gaming experiences and cross-platform comparisons are reviewed here.

The modality of the display is under review since years in the gaming field. Games can be presented differently based on the platform that is used, such as 3D displays, typical conventional 2D screens, small mobile devices, and attention-

grabbing HMDs. Although this critical factor might have a significant impact on user experience, only preliminary studies have been devoted to the comparison of user experiences between common gaming platforms and VR HMDs.

To the best of our knowledge, the very first study towards understanding VR influence on gaming experience was conducted by Tien Tan et al. [Tan 2015] using the Oculus Rift DK1. At that time, the Oculus Rift DK1 provided lower image quality than current HMDs, but was still able to provide a higher degree of flow and deeper immersion when compared to common gaming platforms. Later, [Hupont 2015] investigated user experiences between conventional 2D displays and Oculus Rift DK2, over a forklift truck simulator. Again, higher presence and usability ratings were reported for the HMD. In [Egan 2016], authors carried out an investigation on the influence of HMDs on user QoE. To measure QoE, beside a self-designed questionnaire, authors recorded two physiological signals, Heart Rate (HR), and EDA. Results showed an increased overall quality, measured in terms of MOS, when using HMDs for VR presentation. Furthermore, it was reported that EDA is more representative of user QoE than heart rate when exploring VR experience. All performed analyses ran on HMDs with an embedded display, while most VR gaming consumers have more affordable versions through mobile display HMDs.

Other cross-device comparisons, not involving HMDs, were also performed. In [Schild 2012], a study assessing user experiences of stereoscopic games compared to monoscopic games is presented. This study was conducted with 60 participants, each participant playing one of the three selected games. Results showed richer immersion, presence and higher sickness for stereoscopy games. The impact of the display size on gaming experience was investigated in a study conducted by [Beyer 2014]. Four mobile devices ranging from 3.27 to 10 inches were used. Analyses showed that overall quality and immersion were significantly affected by display size of lower than 5'.

The conducted study aimed to investigate the added value of HMDs when compared to two popular gaming platforms, namely computer and mobile phones, by doing a controlled user study. To explore further on gaming experiences with respect to presented related works, this study compared three gaming platforms using two video games with different levels of comfort.

We addressed the use case of more frequently used HMDs platforms, meaning mobile HMDs [Statista2017 2017]. The aim of this study is devoted to several QoE dimensions such as presence and controls intuitiveness. As such, gaming questionnaires were examined to determine whether they are valuable tools for the assessment of QoE in VR gaming. Finally, the possibility to reduce the number of questionnaire items during the assessment process was explored.

5.2.2.1 Experiment design

Here is a detailed description of the experiment design. The test questionnaires are first introduced, followed by the justification and description of equipment and video games.

Equipment The selected typical gaming platforms were mobile phone and computer. Game consoles were not considered here as it is unlikely to find the same game on all platforms. Besides, it would be questionable to generalize the results of a single console and apply them to all game consoles.

The test equipment consisted of three gaming devices (computer, mobile, and HMD), as well as headset and game controller (gamepad). A Samsung Galaxy S6 Edge smartphone[3] represented the mobile platform. The mobile display was 5.1-inch and Quad High Definition (QHD) (2560x1440). The Oculus Gear[4], which combined the above-mentioned smartphone and a portable HMD, was chosen as HMD platform. The computer platform was a Mac Book Pro, Retina, 15-inch (2880x1800), with a 2.8 GHz Intel Core i7 processor and an Intel Iris Pro 1536 MB graphic card. Specific attention was paid to keep the material clean and hygienic. To have no bias due to audio settings, the same sound system was used for the three gaming platforms. Stereo sound was provided by a professional Sennheiser Momentum On-Ear Headphone (frequency range from 16 to 22000 Hz). The Mad Catz micro CTRL R wireless (Bluetooth) gamepad[5] was used to control the HMD. The mobile phone was controlled by using the touch display, while a standard keyboard and mouse were used for the computer. The test room provided a calm and neutral environment, free from environmental distractions and noise, fulfilling audio-visual tests requirements given in Rec. ITU-T P.910 and P.911 [ITU 2008d, ITU-T Recommendation P.911 1998].

An example of a subject taking part to VR gaming sessions is presented in Figure 5.4.

Video games The experiment did not aim to compare the possibilities of each platform regarding storytelling, game design or graphics. Its purpose was to verify if the same experience provided on various gaming platform will result in different QoE. Accordingly, the selected games should be available on the three platforms. Also, games' platform-variants must implement the same game design, gameplay, and level design. The number of available video games fulfilling those constraints was highly limited. This explains why only two games were eligible.

Deer Hunter is a first-person shooting game which aims at hunting animals in the wild. Gunship Battle is a flight simulator game in which the player is assigned to missions to protect his military base or attack the enemy. Mobile versions of games are available on the Google Play platform. HMD versions of Deer Hunter and Gunship Battle are available on Oculus platform.

Concerning HMD, there is an internal classification provided by Oculus [6] regarding the extent of VR sickness provoked by the game, which classifies Gunship Battle into intense games and Deer Hunter into comfort games. Therefore, we expect more

[3] http://www.samsung.com/ch_fr/consumer/mobile-devices/smartphones/galaxy-s/SM-G928FZDAAUT

[4] http://www.samsung.com/us/explore/gear-vr/

[5] http://madcatz.us/gamepads/microctrlr.php

[6] https://support.oculus.com/help/oculus/918058048293446/

Figure 5.4: Test environment and illustration of VR gaming experience, with and without controller

(a) Deer Hunter (b) Gunship Battle

(c) Duck Hunter

Figure 5.5: Games graphics

sickness effects from Gunship Battle.

The computer version of Deer Hunter is a Facebook application, namely Deer Hunter 2014. The Facebook application prevents game reset and forces subjects to use their accounts. To avoid any privacy issue, all Facebook sessions were launched on a private browser page and the experimenter made sure that the game was not

publishing information to any other person but the participant. Gunship Battle version on the computer is a mobile application played on Andy Android emulator. This might cause users to have less intuitive control over the game.

Additionally, Duck Hunter was used in a training session. Duck Hunter is available on the Oculus platform. Assuming participants have used a computer or mobile phone for gaming before the experiment, only HMD was used during training. Training was done to familiarize players with controls and to reduce learning effect. It also permitted several participants to have their first experience with VR before the test session.

Views of computer games Deer Hunter and Gunship Battle are shown in Figure 5.5, along with a viewport of the VR Duck Hunter game.

Questionnaire The evaluation of subjects' gaming experiences was a three steps process. Prior to the test, participants provided their gamer profile with a pre-questionnaire. During the experiment, they filled a core questionnaire after each test stimulus. After the last stimulus, subjects were asked to complete a post-questionnaire. To further investigate the levels of satisfaction induced by gaming experiences, expectations of subjects towards their gaming experience on different devices were collected during pre- and post-questionnaires.

- Pre- and Post-questionnaires:

  The pre-questionnaire targeted the identification of participants' gaming expertise. Subjects' demographic information (age, gender), gaming experiences and habits (e.g., preferred genre of games and platform for gaming) were gathered.

  In details, we found important to be able to observe any correlation between gamers profile and their expectations or behaviors. We thought of several dimensions that could relate users profiles to ratings of the experiences.

  - Gamers experience may influence experiences. Experts gamers, for instance, developed skills such as learning controls quickly, understanding the underlying level design or noticing game cues more easily which possibly increase their enjoyment when playing. Also, gaming habits not fitting with controls design may prevent such users to have a great experience.
  - Related to the first item, the familiarity with specific games genre may influence ratings. We went for the most typical games genre, including those of selected test games.
  - Previous works of my DT partner hinted that game consumer habits might influence our results. Accordingly, we investigated how participants buy their video games, how often do they play computer games, what is the usual duration of a game session, what screen size is their usual platform display system and where do they usually play.

How do you rate your game experience?						
1 Novice/ Casual gamer	2	3	4	5 Expert/ Core gamer		
Have you ever played game of the following genre? (check if yes)						
Shooting	Sports	Flight simulator	Racing	Puzzle	None of them	Others
Rank the following consuming games habit.						
Download on steam	Buy in a store	"Free" version	Only play free games	Other:		
Which one of the following games have you played before?						
Dear hunter	Gunship battle	Duck hunter	None of them			
How often do you play computer games in a month?						
What is the average duration of your gaming session?						
Rank the gaming platform you are using (Put "None" in a place you are not using the remaining platforms).						
Mobile	Console (Playstation, Wii, etc.)	Computer	HMD (VR)	Other:		
Rank the following display size in function of the frequency of usage (Put "None" in a place you are not using the remaining display sizes).						
Less than 5 inches (regular smart phone)	Between 5 and 8 inches (phablet)	Between 8 and 12 inches (tablet)	Between 12 and 18 inches (laptop)	More than 18 inches (PC display)		
How frequently do you play games at the following places? (daily, weekly, monthly, less than once a month, etc.)						
At home (including a home office)	At work office/ At school	At a public terminal (e.g. library, cybercafe, etc.)	On the way to work/ home	At game-net places (friends place, e-bar, etc.)	At other places	

Table 5.6: Demographic questionnaire

  - Encountering games for the first time is a different experience than playing a game for several months (or even years). Hence, we asked if subjects had already played to the games included in the test.
  - Platform comparison may be influenced by the platform that subjects prefer. It thus seemed interesting to gather subjects' ranking of platforms.

  The questions asked to subjects are reported in Table 5.6.

  Apart from the questions mentioned above, pre- and post-questionnaire were there to collect participants' feedback on expectations towards their gaming experience on different devices.

- Core questionnaire:

  The conducted study aimed at the comparison between gaming experiences on HMD and conventional devices. In the state of the art, it has been stressed out that presence and controls should be evaluated during the cross-platform comparison. However, most widely used presence and gaming questionnaires do not assess both dimensions. The adequate questionnaire concerning these constraints is the PENS.

  Regarding the extension of the work of [Jennett 2008] for the reduction of the number of questionnaire items to specific aspects of VR, questions were added to the PENS. These questions were appraising overall quality, enjoyment, presence, and immersion. In addition, assessing VR Sickness was essential when conducting experiments using such devices [Munafo 2017].

  These dimensions were defined as follows.

  - *Quality* is "the outcome of an individual's comparison and judgment process. It includes perception, reflection about the perception, and the description of the outcome" [Brunnström 2013].
  - *Enjoyment* refers to the resulting combination of satisfaction, pleasure, fun, and pride.
  - Virtual reality *Sickness* is explained by discrepancies between sensory inputs (equilibrioception and vision) [Ko 2011] and its symptoms are similar to motion sickness [LaViola Jr 2000] (e.g., general discomfort, headache, nausea, fatigue and disorientation [Kolasinski 1995]).

  23 questions constitute the final used questionnaire, 18 from the PENS questionnaire and five additional self-designed questions. A detailed description of the PENS questionnaire can be found in [Lau 2015, Ryan 2006]. As examples, among other questions, "I experienced a lot of freedom in the game." assesses autonomy, "I feel very capable and effective when playing." evaluates competency, "When moving through the game world I feel as if I am actually there." appraises presence and immersion, and "The game controls are intuitive." identifies controls intuitiveness. All questions are evaluated

Dimension	Questions
Overall quality	How do you rate the overall quality of the game?
Autonomy	The game provides me with interesting options and choices. The game lets you do interesting things. I experienced a lot of freedom in the game.
Competency	I feel competent at the game. I feel very capable and effective when playing. My ability to play the game is well matched with the game's challenges.
Presence and immersion	When playing the game, I feel transported to another time and place. Exploring the game world feels like taking an actual trip to a new place. When moving through the game world I feel as if I am actually there. I am not impacted emotionally by events in the game. The game was emotionally engaging. I experience feelings as deeply in the game as I have in real life. When playing the game I feel as if I was part of the story. When I accomplished something in the game I experienced genuine pride. I had reactions to events and characters in the game as if they were real.
Intuitive controls	Learning the game controls was easy. The game controls are intuitive. When I wanted to do something in the game, it was easy to remember the corresponding control.
Self-designed questions for sickness, enjoyment, presence, and immersion	I have experienced sickness/ discomfort while playing. Overall, I have enjoyed playing this game. Overall, I felt present in the game. Overall, the device and the game provided me with immersive experience.

Table 5.7: Game questionnaire

using a 7-point Likert Scale.

Regarding self-designed questions, the overall quality was assessed with the question "How do you rate the overall quality of the game", using a 7-point scale from "Extremely bad" to "Ideal".

Remaining dimensions were assessed on a 7-point Likert scale from "Not at all true" to "Very true". "I have experienced sickness/discomfort while playing", "Overall, I have enjoyed playing this game", "Overall, I felt present in the game", and "Overall, the device and the game provided me with immersive experience", were used to evaluate sickness, enjoyment, self-designed presence and immersion, respectively.

Table 5.7 reports all the items of the core questionnaire.

- Expectations evaluation:

What are your expectations regarding the following devices as gaming platforms?
Dimension
Quality (fps, resolution, bitrate)
Controls (intuitive controls, interface)
Immersion (physical)
Presence (feeling present in the content)
Competency
Novelty
After-effects
Duration of a game session
Emotion/ Fun/ Pride
What this platform brings you more compared to others

Table 5.8: Expectations questionnaire, asked for the three gaming platforms mobile, computer and VR

In this study, we have decided to focus on two types of expectations: *ideal* expectations represent the expected performance of an ideal service [Parasuraman 2002], and *minimum tolerable* expectations describe an acceptable consumption of a certain service in a given situation and context [Zeithaml 1993]. Indeed, conclusions of [Zeithaml 1993] indicated that these types of expectations are of importance in assessments.

Ideal expectations are assumed to be constant, while adequate expectations are time-varying due to context adaptation. The latter type of expectations was investigated here, as it was more compliant with a comparison across gaming platforms and it allowed the analysis of expectations variation. We follow the

E-P gap model theory [Parasuraman 1985], which characterized satisfaction as the difference between perceived quality and user's expectations about a service.

To obtain further insights into users' satisfaction, the questions about expectations were included in pre- and post-questionnaire.

The explored dimensions of expectations are quality, immersion, presence, sickness, enjoyment, control, competency and novelty. They cover the range of questions asked in the core questionnaire.

The novelty scale was defined from 1 to 5 as follows: "Completely new experience", "Pretty new experience", "usual experience", "boring experience" and "Not interesting experience". Remaining dimensions were assessed using a usual 5-point ACR scale.

The question formulation for all expectations dimensions is presented in Table 5.8.

Stimuli and Subjective Evaluation Methodology The combination of two games and three platforms gave six 8-minute long stimuli. These were assessed individually in a within-participants design, using the core questionnaire.

Stimulus duration should be long enough to let subjects get engaged in the gaming experience, while it should also be short to limit in-game variations across subjects. Stimuli started with the tutorial of controls followed by the very first level of the game. On request, help was provided if subjects did not thoroughly understand the controls. The main uncontrollable variation of stimuli was subjects' ability and style of play.

The experiment consisted of a training session before two test sessions. The 3-minutes training stimulus ensured that subjects comprehend the evaluation process. The HMD platform was evaluated during this training session to prevent players from having their first VR experience during a test session.

Each test session presented one game stimuli on three platforms. The first game to be investigated was Deer Hunter, as it is a comfort game. In every test session, devices order was randomized to remove any platform order bias. An exception was made for the HMD Gunship Battle stimulus, as experiencing it was likely to result in VR sickness. To be independent of the occurrence of VR sickness, this stimulus was the last stimulus to be assessed.

Subjects were required to fill the pre-questionnaire before the experiment. An explanation sheet defining technical terms, presented in Table 5.9, was available during all the assessment process.

After every training and test stimuli, subjects filled the core questionnaire. At the end of the test, participants completed the post-questionnaire. The experiment took about two hours, without considering pre-questionnaire filling duration. Each session lasted for about 45 minutes. To prevent participants from boredom and lack of attention, a 10-minute break was enforced between two test sessions. This experimental procedure is presented in Figure 5.6.

Quality	Quality is the outcome of an individual's comparison and judgment process. It includes perception, reflection about the perception, and the description of the outcome. [Brunnström 2013] In our context, quality can be seen as a function of frame per seconds(fps), resolution and bitrate, among others.
Control	Control is the set of commands allowing the conversion of gamers intentions into actions. inspired from [Rigby 2007]
Immersion	The extent to which the computer displays are capable of delivering an illusion of reality to the senses of a human participant. [Slater 1997]
Presence	The concept of presence means that players feel they are truly "in the game"- not just in terms of it holding attention, but involving the player emotionally and drawing them into the world it creates. Physical presence is a primary measure of immersion and measures the extent to which the player feels that have been physically transported into the game environment during play. Emotional presence measures the extent to which the game action elicits emotion that feels authentic to the player, much like they may feel in response to real life events. Narrative presence looks at the involvement of the player in the story or narrative of the game. [Rigby 2007]
Competency	Competence can be defined as the intrinsic need to feel a sense of mastery or effectance in what one is doing. [Rigby 2007]
Novelty	The novelty effect, in the context of human performance, is the tendency for performance to initially improve when new technology is instituted, not because of any actual improvement in learning or achievement, but in response to increased interest in the new technology. [Gravetter 2018]
After-effects	Virtual reality sickness (also known as cybersickness) occurs when exposure to a virtual environment causes symptoms that are similar to motion sickness symptoms.[LaViola Jr 2000] The most common symptoms are general discomfort, headache, stomach awareness, nausea, vomiting, pallor, sweating, fatigue, drowsiness, disorientation, and apathy. [Kolasinski 1995]
Duration of a game session	Game session is the period you are playing a game from starting playing until stop playing. It is assumed you stop a game when you plan not to play for more than an hour. A game session duration is the average number of minutes or hours spent playing during such a session.

Table 5.9: Definition of important terms

Among the 18 participants that took part in this experiment, one was female. The subjects sample is 25.6 years old on average with a median age of 25.5 years old. Participants can be approximately categorized into casual (22%), expert (22%) and moderate gamers (55%), based on their judgment of expertise assessed in pre-questionnaire. Despite most of them usually play shooting games and half of them flight simulator, none of them had already played the evaluated games before taking part in the test.

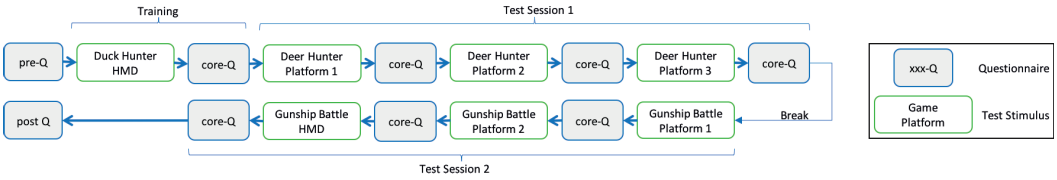


Figure 5.6: Diagram of the experiment procedure

5.2.3 Results

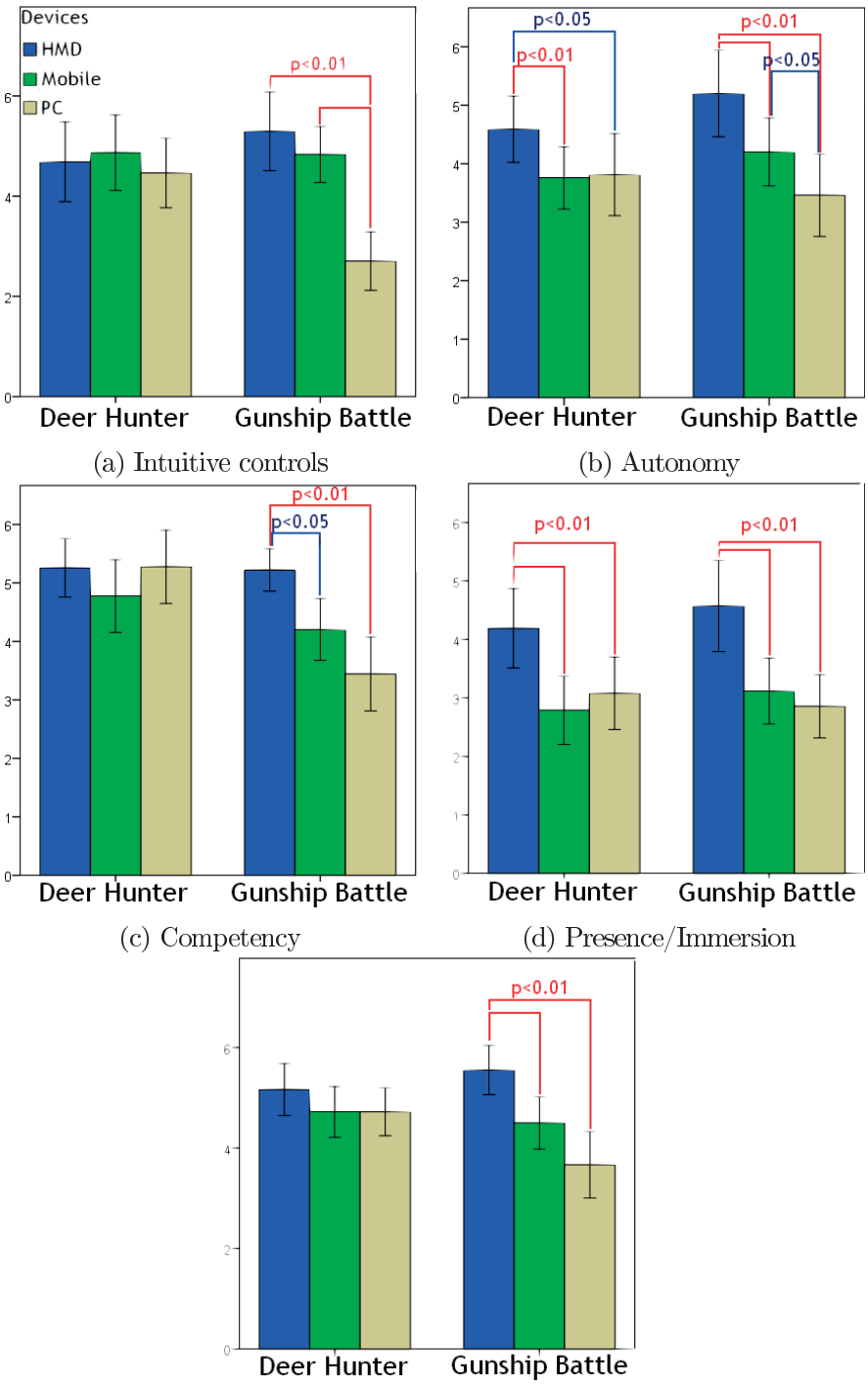
This section introduces the conducted analysis of gathered subjective ratings. First, differences between platforms were analyzed using MOSs, CIs, and one-way repeated measures ANOVA for assessed dimensions. Furthermore, we processed our data to verify whether it was possible to reduce the number of questionnaires items for presence and immersion. With respects to validity concerns, the influence of VR sickness has been investigated. Then, presence ratings were analyzed in more details to understand what VR gaming really changed compared to typical platforms in terms of presence. We dug deeper into subjects ratings to observe subjects' satisfaction level and expectations variation. Finally, several feedbacks of subjects were reported and thus reported here.

5.2.3.1 Gaming platforms comparison

There were two kinds of variables. Independent variables were HMD, mobile and computer platforms. Dependent variables were the dimensions of the PENS questionnaire and overall quality.

A one-way repeated measure ANOVA was performed to determine whether there were some statistically significant differences between these two types of variables. Mauchly's test [Mauchly 1940] indicated that the assumption of sphericity had been violated for some dimensions. Therefore, the degrees of freedom were corrected using Greenhouse-Geisser estimate of sphericity [Field 2013].

ANOVA results are presented in Table 5.10. MOSs and associated 95% CIs of PENS dimensions and overall quality of each stimulus as well as results of post-hoc tests using the Bonferroni correction are graphically illustrated in Figure 5.7. Since both games provided different results on several dimensions, a separate analysis was carried out for each of them.



(a) Intuitive controls

(b) Autonomy

(c) Competency

(d) Presence/Immersion

(e) Overall Quality

Figure 5.7: 95% CIs and Mean of PENS dimensions and MOSs of Overall quality. Indication of significant differences across platforms based on post-hoc tests.

Results of Deer Hunter showed that the HMD presented experiences with significantly higher presence and autonomy compared to other platforms. As already mentioned, presence was viewed as a crucial characteristic of VR. Based on our results, autonomy also influenced VR gaming experiences. However, we witness no impact of the increase in these dimensions on overall quality results.

Results of Gunship Battle differ significantly between computer and HMD, for all dimensions. HMD and mobile platforms are generating significantly different experiences regarding all aspects but intuitive controls. It has to be recalled that an Android emulator provided the computer platform for this game, which mainly resulted in weak game controls for this platform. This explains the poor results of this platform.

Mobile and computer platforms exhibit differences in autonomy and intuitive controls. It appears that a violation of intuitive control and autonomy does not necessarily affect overall quality.

With respect to our previous findings, it appears that overall quality is likely to be related to competency and presence. The change of platform always led to a similar effect on competency, presence and overall quality, even when the influence on other dimensions is different. This finding strengthens the possible causal effect of presence and competency on overall quality for this game.

When merging all our results, we concluded that variations in intuitive controls, autonomy, and presence had a relatively low effect on overall quality. Thus, we related significant increase of overall quality to competency. This key finding clearly indicates to gaming developer on which game dimension they should pay attention.

	Deer Hunter					Gunship Battle				
	F	df1	df2	p	$\eta_p{}^2$	F	df1	df2	p	$\eta_p{}^2$
Autonomy	10.1	1.5	25.6	0.001	0.37	18.2	2	34	0.01	0.52
Competency	2.2	2	34	0.12	0.12	14.9	2	34	0.01	0.47
Presence	30.5	2	34	0.01	0.64	34.8	1.4	24.3	0.01	0.67
Intuitive Control	0.4	2	34	0.65	0.02	36.3	2	34	0.01	0.68
Overall Quality	2.2	1.5	25.7	0.13	0.11	18.4	1.4	23.6	0.01	0.52
Sickness Experience	2.7	1.3	21.6	0.11	0.14	13.9	1.4	23.3	0.01	0.45
Overall Enjoyment	5.9	2	34	0.01	0.26	10.9	1.3	22.9	0.01	0.39
Overall Immersion	22.9	2	34	0.01	0.57	31.9	1.5	25.6	0.01	0.65
Overall Presence	17.7	2	34	0.01	0.51	18.5	2	34	0.01	0.52

Table 5.10: One-way repeated measures ANOVA of the assessed dimensions

A few more points needed to be looked at carefully. First, since two games were utilized in this study, results were not generalizable for all games. Second, other HMD devices such as HTC Vive may provide a higher feeling of presence, which usually results in better overall quality. However, it is recalled that we addressed the use-case of affordable mobile display HMD. Thirdly, some factors might influence each other, as it was observed for Gunship Battle using the computer. Here, low

intuitive controls most likely decreased the ratings of other dimensions. Last but not least, other aspects (e.g., flow and aesthetic), which were not evaluated in this study, might have an impact on obtained results.

5.2.3.2 Presence sub-dimensions

The presence dimension of PENS consisted of three sub-dimensions: (1) **physical presence**: extent to which the player feels to have been physically transported into the game environment, (2) **emotional presence**: extent to which game action elicits emotion that feels authentic to the player similar to experiences in real life and (3) **narrative presence**: player involvement in the story or game narrative.

To investigate differences between these presence sub-dimensions, an ANOVA was applied. It revealed that there is a difference between HMD and two other devices for all sub-dimensions of presence ($p < .001$). While for HMD physical presence was the highest among the three presence sub-dimensions, it was the lowest for the other two devices.

However, it was not expected to have a significant difference in emotional presence and narrative presence for different devices, since same games scenarios were used for all devices.

Besides, as it was expected, no significant difference was observed for presence sub-dimension between mobile and computer devices. We can conclude that HMD platforms increases presence sub-dimensions, especially the physical dimension.

5.2.3.3 Presence and immersion assessment

With the aim to reduce the number of items within questionnaires for future studies as shown. Following Jennett's process [Jennett 2008], *presence* and *immersion* were assessed with a single question.

From now on, presence and immersion refer to these self-designed questions, while PENS presence relates to presence dimension of the PENS questionnaire.

A Pearson correlation analysis revealed that there is a very strong, positive correlation between PENS presence items ($r = 0.84$). These results demonstrate that the number of questions for PENS presence dimension could be reduced to a single question with almost the same accuracy. Further studies are required to confirm this finding.

Furthermore, we investigated the correlation between PENS presence and immersion, since in PENS dimensions the concept of presence and immersion is merged to some extent [Johnson 2014]. Results showed a strong, positive correlation between PENS presence and immersion, ($r = 0.80$). Additionally, presence and immersion themselves are highly correlated ($r = 0.86$). Even though the correlation between PENS presence and immersion is slightly weaker compared to the correlation of PENS presence and presence, it appears that participants cannot distinguish between both concepts based on provided definitions.

We recall that decreasing the number of items is only applicable in specific contexts as a self-designed single question does not distinguish between possible

sub-dimensions. This process is only to apply when the required level of insights of the evaluated dimension is low.

5.2.3.4 Influence of VR sickness

As reported in many studies, VR sickness is a severe problem when using HMD [Munafo 2017]. We can confirm VR sickness occurrence for a specific game (Gunship Battle), as several participants reported sickness during the conducted study, while no sickness was reported for Deer Hunter.

ANOVA revealed that for discomfort game Gunship Battle, HMD usage led to significantly increased virtual reality sickness ($p < 0.01$). This is in line with comfort classification of these two games. Fast motion in the game Gunship Battle increased the likelihood of sickness.

Pearson correlation between sickness and overall quality revealed no significant correlation neither for Deer Hunter ($r = -0.317$), nor for Gunship Battle ($r = 0.187$).

Participants reported in interviews that even when experiencing VR sickness, they had a rich gaming experience due to the high presence provided by the game.

It appears that even though sickness is a problem when using VR technology and should be carefully considered when conducting experiments, it does not necessarily affect overall quality.

5.2.3.5 Expectations qualitative analysis

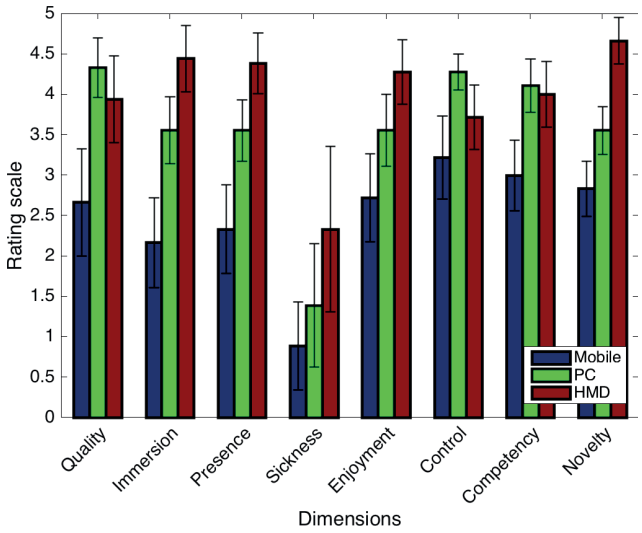


Figure 5.8: MOSs and 95% CIs of expectations

In this section, we inspect the collected results regarding expectations. Expectations collected in pre-questionnaire are called *expectations* while those collected in post-questionnaire are referred to as *revised expectations* within this section.

Expectations MOSs and CIs are presented in Figure 5.8. It can be observed that expectations about mobile platform performance are always the lowest compared

to two other gaming platforms. For quality, control and competency dimensions, computer platform expectations exceed that of HMD. The opposite behavior is observed for remaining dimensions.

Higher expectations for computer compared to mobile phone can be explained by the difference between typical games designed for these platforms. Since computer has higher processing power and larger screen size, more complex games can be created for this platform. Also, large companies spend more money on the development of high-quality computer games, compared to mobile games. Computer games are thus more expensive.

Expectations towards HMD are reasonable since this technology is new and advertised with an increased level of presence and immersion. Overall, expectations regarding the three platforms are pretty high, as ranging from fair to excellent. An exception is raised for sickness dimension, which is assessed lower than medium. It should be noted that, according to CIs, a significant difference between the three platforms is shown for immersion, presence, and novelty.

We can observe a significant difference between mobile and HMD platforms for quality, competency end enjoyment, between mobile and computer platforms for quality and competency as well as between computer and HMD platforms for controls. These results emphasize dimensions which will impact cross platforms comparison: immersion, presence, novelty effect as well as quality, competency and enjoyment.

These results strengthen our previous findings in which we have observed that overall quality, presence, and competency were related for Gunship Battle.

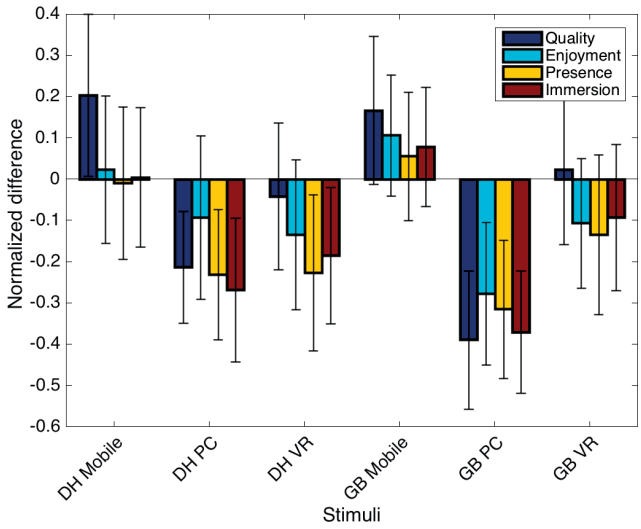


Figure 5.9: Gap model results

The gap model resulting from the difference between expectations and stimuli scores for quality, enjoyment, presence and immersion dimensions are presented in Figure 5.9. Scores were normalized assuming the linearity of the rating scale.

Overall, subjects were satisfied by the mobile gaming experience, especially for the game Gunship Battle, while they were disappointed by computer and HMD platforms. Stimulus Gunship Battle on computer platform was notably less appreciated than other experiences. This can be because this stimulus results from the use of a mobile game version played on an emulator.

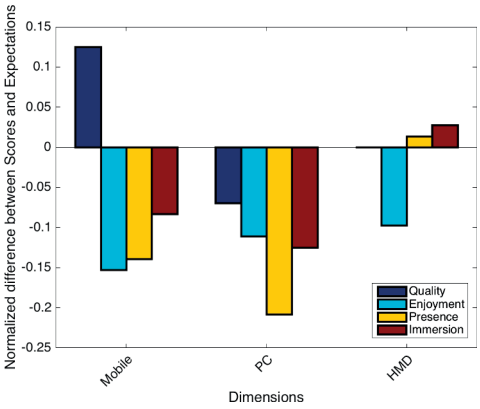
Expectations variation is the difference between revised expectations and expectations. They are depicted in Figure 5.10, overall and for the sample of populations sorted by platform preference.

Regarding expectations variations, overall, expectations about mobile quality increased, while presence and immersion expectations of the HMD platform increased slightly. The remaining expectations decreased moderately.

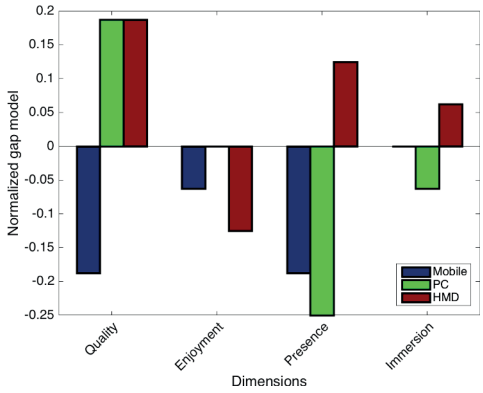
It is interesting to note that the subset of 10 participants preferring computer platform in Figure 5.10c(assessed by using a preference ranking in pre-questionnaire) have significantly decreased their expectations for all dimensions about computer platform.

The subset of subjects having no preferred platforms (2 participants) decreased their expectations significantly about the HMD platform, especially considering the enjoyment dimension. Their results are reported in Figure 5.10e. These results emphasized participants' overall dissatisfaction regarding games performance, except for the quality level provided by the mobile, and presence and immersion experiences offered by the HMD.

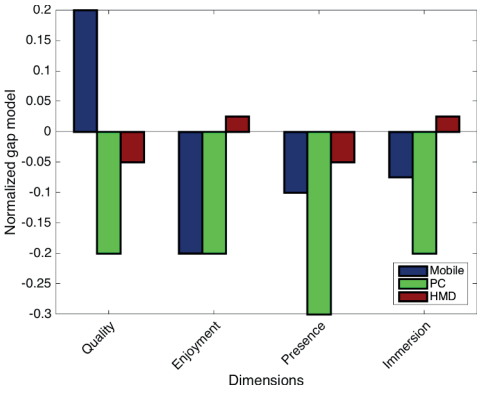
It is surprising that presence and immersion expectations increased, while the gap models showed slight dissatisfaction for these dimensions considering both games. An explanation could be that participant expectations are now based on a new estimation of HMD potential, considering their experience acquired during the test, knowing that 72% of the participants had never played with HMDs before the experiment. This would mean that minimum acceptable expectations can be formed based on the foreseen capabilities of a service or device after several experiences, rather than on the experiences themselves.



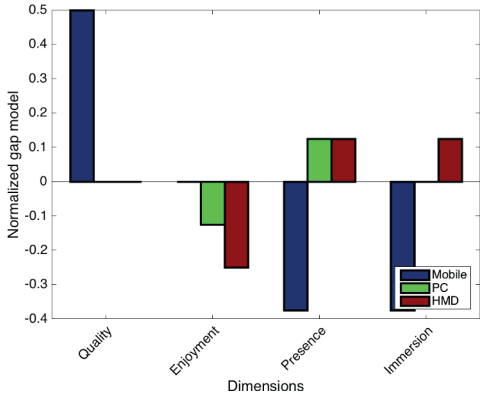
(a) Overall expectations variation



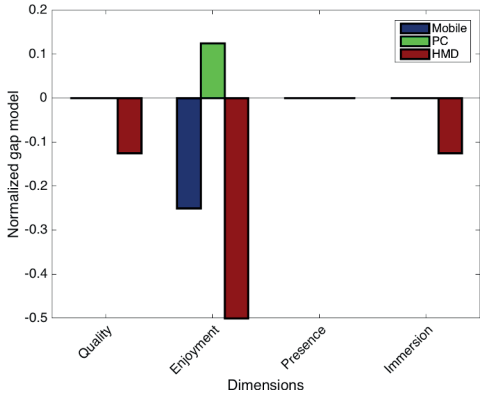
(b) Expectations variation from gamers preferring the mobile platform



(c) Expectations variation from gamers preferring the computer platform



(d) Expectations variation from gamers preferring the HMD platform



(e) Expectations variation from gamers having no preferred platform

Figure 5.10: Expectations variations overall and for population samples having or not a preferred platform

5.2.4 Discussion

During the subjective assessment, PENS dimensions (autonomy, competency, intuitive control, and presence) and self-designed questions for presence, immersion, sickness, and enjoyment, were investigated. However, this study is constrained as other potentially influencing dimensions could have been examined. As reported earlier, we wanted to use a questionnaire including both presence and gaming dimension. PENS was the most reliable questionnaire fulfilling these requirements.

Because of the HMD Gunship Battle gaming experience, which provokes VR Sickness, it has been decided not to run randomization over the games (but constrained randomization was done for the used platforms). This limitation appears to be reasonable as the experiment deals with a cross-platform evaluation.

Results of computer platform should be carefully considered for the game Gunship battle. The game, and especially its controls, were not designed for computer platform and resulted in difficulties in interacting with the game interface. This might bias the results, especially for control intuitiveness and competency dimensions. However, our analysis has been carried out by taking this consideration thoroughly into account, and no disputable conclusion has been drawn.

Besides, playing time was set for 8 minutes. The selection of this duration results from a trade-off between proposing a realistic gaming experience and limiting in-game variations across subjects.

The studied population is mostly composed of male subjects. An analysis comprising more females should be performed to get more gender-balanced results. We can argue that this population is, to a certain extent, representative of the proportion of male and female players in the gaming market.

Several participants reported that they tend to play higher quality games on specific platforms only, and that particular game types are just suited for a particular platform. This fact is due to the limited choice of available and above all similar games on all three platforms.

The conducted study showed that PENS questions on presence could be aggregated into a single question. However, an appropriate definition of presence has to be set before conducting the test, as in PENS immersion and presence is considered as a single dimension. We believe that this method could be extended to other dimensions to reduce the number of questions, while more studies are required to confirm such a result.

5.2.5 Conclusion

This study is a contribution to ongoing research on the benefits of VR experience on new and affordable HMDs in gaming. To the best of our knowledge, this is the first study that provided a comprehensive investigation of three different gaming platforms involving VR gaming.

This study is the result of a collaboration between the MMSPG of EPFL and the QUL of DT. EPFL was responsible for the test design, from questionnaires

creation and selection of experiment procedure, and for data analysis, especially expectations results. DT was in charge of preparing content and material (e.g., game and platform selection and settings) and perform data statistical processing. All along the process, fruitful discussions about all study aspects were conducted. Subjective evaluations were performed jointly.

The platforms involved in the study were a mobile platform, through the use of a Samsung Galaxy S6 Edge, and a computer platform, precisely a Mac Book Pro Retina. An Oculus Gear coupled with the mobile platform device constituted our VR gaming platform. 18 people participated in our study and were 25.6 years old on average.

The added value of three gaming platforms was explored by comparing the experiences of two video games with regards to influencing factors such as presence, autonomy, and competency.

From the outcome of our investigation, it is possible to conclude that HMDs provided higher presence than the two other platforms. Although experiencing presence may have a positive effect on overall quality, its extent varies as other influence factors, such as game preference, have an impact on overall quality.

Results emphasized the probable causal effect of competency on overall quality, making it an essential QoE dimension.

Another interesting finding of this study is the high rating of overall quality for Gunship Battle, despite being an intense game that introduced VR sickness.

Future experiments should also be conducted to confirm the highlighted possibility to reduce the number of questions in gaming assessment for specific dimensions, such as immersion and presence. Additionally, future studies should consider a longer stimuli duration when investigating VR sickness.

Regarding expectations analysis, the gap model between expectations and stimuli scores as well as expectations variations (revised expectations - expectations) present insights that were not visible in usual analyses.

First, we observe significant discrepancies between initial constructs of importance for platforms. For instance presence, immersion and novelty must be evaluated for VR gaming while quality, control, and competency are influence factors most related to computer platforms assessment. This result stresses which features are important for the comparison.

Second, from gap analysis, we went more thoroughly into differences between platforms. It is pretty clear with the gap model that Gunship battle on computer platform was the least satisfying experience. This finding pointed at our test design and the fact that we used an Android emulator to play Gunship Battle on a computer.

It showed that controls are crucial for creating enjoyable gaming experiences. Moreover, it is relevant to play games on the platform they are designed for.

From expectations variation, subjects behaviors were also easier to investigate, overall and within population samples. The gathered knowledge with expectations

is meaningful as it is easier to understand what happened during the evaluation. Also, most findings were confirmed by the interviews conducted in pre- and post-questionnaires and orally after the experiment.

Last but not least, the expectations analysis provided insights on how expectations are created. We were able to observe that ideal expectations may more be based on the foreseen potential of technology than the actual quality of previous experiences. This finding is highly valuable.

Our results showed the interest to include expectations analyses in QoE evaluations.

5.3 Conclusion

This chapter dealt with research on aspects of QoE which are user-centric and possibly mutual to all immersive technologies.

The novelty effect appears when a new product, technology or service is issued. This effect threatens the validity of subjective evaluations performed on new or emerging multimedia technologies. Indeed, ratings are biased by several factors such as enjoyment, excitation or caution, and wariness.

However, no standard definition for the multimedia field was provided in the literature. We proposed the following definition. **The novelty effect is the transitory and unconscious tendency of subjects to evaluate their experience differently the first times they are using a product, not because of the product intrinsic features, but because the user perceives some of those features as unfamiliar.**

We related novelty effect behaviors with one's representations of product features, predictions of relative advantage, perception of the degree of newness. To our mind, user's present state of mind, former experiences, expectations, and interest about a product or service also impact and are impacted by the novelty effect.

We thus tried to observe the novelty effect generated by VR gaming experiences. We compared the results of two different sets of a population. The first one is free from of any real experience with VR. The second has been trained through familiarization sessions, expected that the novelty effect wears off for this population. During each of the four familiarization sessions, subjects played to two games for a quarter of hour.

The novelty effect has been observed through expectations variation. Any other evaluated dimensions were not statistically different from the two population sets. We thus claim that it is better to include expectations in QoE evaluations than appraising novelty.

Accordingly, we conducted an experiment which investigated expectations. We needed to examine if expectations bring their share of information about subjects behaviors and QoE.

The study was a cross gaming platform comparison to understand the difference between VR gaming and common platforms such as mobile phones and computers.

We have defined several purposes for this study. The first was the identification of VR gaming advantages over other platforms. Providing which influence factor is to develop for a particular platform is of definite interest for game developers. Our results indicated that competency is an aspect to expand game satisfaction. Presence was always higher in VR experiences but did not necessarily improved the game quality. Also, violation of intuitive controls and autonomy did not strongly impact game quality.

The second was to understand precisely in what extent presence is leverage for VR gaming. To emphasize physical presence in VR gaming is the key outcome of our study. What was surprising is that although games share the same game design, graphics and level design, emotional and narrative presence were higher for HMD games.

Thirdly, presence/immersion items of the PENS strongly correlated with presence and immersion single questions. Should the context allow it, it would be possible to wonder about immersion or presence with a unique question. This is an important finding regarding the extensive amount of items in gaming questionnaires, possibly creating boredom and lack of attention in subjects during evaluation.

A fourth interest was to verify if VR sickness effects prevent the enjoyment of a gaming experience. It has been shown that in the case of Gunship Battle, known to generate a lot of sickness, sickness was counterbalanced by presence, resulting in rich experiences.

Finally, we discovered the numerous advantages to include expectations in QoE analysis.

Indeed, in MOSs and CIs analysis, we could identify if an influence factor is mattering, and for which platform this aspect is meaningful. Without surprise, immersion, presence, and novelty were influence factors related to the assessment of the VR platform. We learned that control and competency are factors related most to the computer platform.

Expectations gap model (*expectations - quality*) proved to be a great deal. Results highlighted the satisfaction or not regarding the gaming experiences. Indeed, if mobile experiences were fine, remaining experiences were not satisfactory. But most of all, results highlighted the defect of using an Android emulator for Gunship Battle on the computer platform. In addition to being a satisfaction indicator, the gap model analysis helped to identify what to improve in future tests.

Regarding expectations variation (*revised expectations - expectations*), insights were deep and explained subjects behavior. We reached the capability of understanding subjects answers based on their gamer profile. For instance, subjects who preferred computer platform to play video games were disappointed by the computer experiences of the test.

We witnessed the increase in expectations concerning mobile games quality and VR games presence and immersion. This finding is amazing as previously, we have seen that presence and immersion of VR games were not perceived as high.

It means that ideal expectations may be revised based on the foreseen potential of the technology. This opens the door to the understanding of subject's expectations revision and represents a step further along the path to predict QoE.

From what has been learnt in these two works, we recommend paying close attention to the diversity in the population sample under test. Besides, sampling population may introduce biases in data, so it is recommended not to take subjects interest in future experiments as a criterion for selection. Expectations regarding a system technical features are a good indicator of the diversity of the population.

Should the context allow it, that is no refinement is necessary about presence and immersion, these aspects may be assessed through a single question.

We support the use of expectations analysis in pre- and post-questionnaires for QoE evaluations.

- The type of expectations (forecast, normative, ideal or minimum tolerable) must be selected prior to the test based on the study context.
- If one wants to observe expectations variation, we recommend the evaluations of ideal expectations.
- Expectations aspects must be precise, measurable and tangible. We advise to identify the influence factors of the system under test, and then to ask subjects their expectations about these factors.
- A pilot test can validate influence factor selection. In this test, one must verify that expectations about influence factors discriminate evaluated systems.
- Expectations gap model is a convenient satisfaction indicator. We encourage the use of this processing in future studies.
- The gap model has also shown the ability to identify defects in subjective experiment methodologies. We advocate for always improving assessment methodologies, and this is a relevant method, relying on subjects ratings.
- Expectations variation combined with user profile information enable in-depth analyses of behaviors. We advocate for the inclusion of expectations variation analyses in QoE tests.
- Understanding cognitive processes, such as the creation or evolution of expectations, is of huge interest towards QoE prediction. It is thus highly recommended to study expectations variation when measuring QoE.
- We also favor the study of expectations instead of novelty effect studies.

Based on these recommendations, more knowledge will be extracted from collected scores and will make researchers progress in the QoE field.

5.4 Future work

First of all, our results on expectations must be replicated for validation. The presented evidences of interest in including expectations in QoE are undeniable. However, another study confirming those results will strengthen our outcomes.

After proving the benefits of expectations analysis, the evaluation of expectations must be clearly defined and described (e.g., evaluation in pre- and post-questionnaire only). The designed methodology must be validated in view of being standardized.

In my opinion, we can dig deeper to understand the strong relation between expectations and novelty effect. I have the intuition that novelty effect happens when ideal expectations are unstable, the first times one encountered an experience. Of course, ideal expectations are always evolving, but I would expect them to change more during this "honeymoon phase". This study would bring a lot of knowledge on expectations formation and could ultimately help to predict the bias due to novelty effect. This final outcome is valuable for all media providers and multimedia technology manufacturers.

Expectations analysis have a real potential towards QoE evaluations and possibly prediction. They also have the advantage to provide insights on subjects behaviors. The most important future work is to use our works outcomes and to use expectations in future subjective evaluations.

Chapter 6

Prediction of QoE: physiological signals

Contents

Experts on appraisal of expectations clearly claimed that "one should believe only what people do, not what they say."[Manski 2004].

In our studies related to expectations and quality, we asked for explicit self-rating. It is pretty clear that asking subjects to report explicitly the perception about their experience is not optimal. Any explicit subjective assessment by human subjects is influenced by emotional, cultural, educational, and environmental

differences across subjects [Forgas 1999, Geng 2010]. More importantly, explicit assessments can have an impact on the experience itself, and interfere with it. From the formation of a subject's opinion to its conversion on provided rating scales, or even subjects awareness of the coming rating period, various biases may appear. This has been previously discussed in Chapter 2.3.9, when considerations about subjective test validity demonstrated that many biases should be avoided during an individual experiment design.

We thus searched for a way to assess experiences in a more tangible way. In recent years, researchers came up with an alternative. In addition to individual scores, experimenters gathered observers' physiological responses. The mental states and peripheral signals of observers reflected their sensations, perception processes and decision making. Hence, this fulfills the need of researchers to go beyond expression of judgment, and investigate deeper perceptual and cognitive processes. In addition, physiological signals may lead to the creation of an objective modeling of QoE through physiological response analysis. This perspective may seem far-fetched, however, with current progress made in this field, this is a likely and valuable future application.

First, the motivation for including physiological signals in subjective experiments is considered. As those signals bring much more advantages than drawbacks, we will describe them in more details. Then, a review introduces the usual procedures for processing physiological signals. A focus is made on QoE- and presence-based studies.

6.1 Physiological signals

Although explicit subjective ratings provide accurate and reliable results, momentum is given to physiological signals for several reasons.

- The first one resides in the fact that individual ratings are the result of a conscious process. Explicit ratings are prone to numerous biases and many subjective factors (e.g., expectations and current mood) [Bosse 2016]. They provide only limited insight into underlying perceptual and cognitive processes [Engelke 2017].

  Collecting physiological responses to multimedia stimuli makes possible to analyze information that is not explicitly supplied by subjects and could be more genuine [Arndt 2016]. This prevents subjects' difficulties in identifying or constructing their opinion and converting this judgment on the experiment rating scale [Moon 2017, Insko 2003]. In physiological analyses, there is no need to define a rating scale. Moreover, results are less dependent on the task.

- Physiological signals are representative of the transmission and processing of emotional reactions, arousal, cognition processes and sensory cues [Bosse 2016]. In particular, peripheral signals are reliable indicators of emotional responses [Moon 2017], and they allow the capture of high-level QoE

concepts such as engagement, immersion, and SoP [Engelke 2017].

For instance, eye blink rate is known to be associated with visual fatigue [Bang 2014] and pupil dilation relates to cognitive load [Beatty 1982]. Analyses of physiological signals open the door to understanding cognitive processes and predicting human reactions to stimuli.

- Furthermore, the time-continuous dimension of such signals carry information unavailable from individual ratings [Insko 2003]. Subjects are grading experiences globally, while physiological signals are monitoring the physical responses continuously over the entire stimulation. Continuous subjective assessments exist, e.g., employing a sliding grading rule all along a stimulus or by examining the opinion at regular time-periods. However, such disruptive approaches considerably affect the experience.

  As long as the recording equipment is non-invasive, the collection of physiological responses is non-disruptive and offers the opportunity for real-time assessment. It is worth mentioning that the high temporal resolution of physiological signals potentially exhibits instantaneous or high responsiveness to events.

  New QoE-based applications can be derived from real-time assessment. For instance, in [Moon 2017] authors foresee QoE-aware content delivery, personalized recommendation systems, content filtering, implicit tagging through metadata, QoE-based Brain-Computer Interface (BCI) and new VR and video games controls.

- Insko [Insko 2003] compared state-of-the-art subjective, behavioral and physiological analyses. He concluded that, overall, the latter produces results which are reliable (e.g., repeatable), valid (internal and construct validity), objective (external validity) and sensitive (capable of discriminating levels of what is measured). Subjective analyses usually fail to provide objectivity and sometimes sensitivity, while behavioral investigations often do not ensure any of the four aspects.

Even though the advantage of including physiological signals in subjective evaluation is clear, several negative aspects exist as well.

- We have already mentioned the huge impact of previous experiences, knowledge and expectation on individuals' experiences. These factors modify the physiological perception and response of viewers to a stimulus.

  In physiological studies including functional Magnetic Resonance Imaging (fMRI), subject-wise difference in physiological signals is such that observers identification is possible [Wang 2015]. This variance of perception among people generates systematic errors [Engelke 2017]. This is an obstacle to form a general system for identifying high-level cognitive processes such as quality perception. This does not mean that overcoming subjectivity is impossible.

An effort has been made to find solutions reducing this variability within subjects. A widely used option is to measure baseline levels and analyze signals relative to the baseline instead of absolute values [Insko 2003]. To further reduce the impact of variability across individuals, [Moon 2017] stressed the necessity to discover physiological features that are common to viewers.

- Even with non-invasive equipment and despite the non-disruptive characteristic of the acquisition of physiological signals, this type of evaluation process is intrusive. Discomfort and dullness may appear during and after placing the measurement material on participants. Both responses are likely to change subjects' behavior. This potentially threatens the construct validity of tests, depending on what is intended to be measured [Engelke 2017]. Moreover, outfitting subjects lead to lengthy experiment duration. This directly impacts the availability of participants and the experiment design.

- Subjective evaluations which comprise physiological responses have a drastically increased complexity, regarding design and implementation of experiments, signal-to-noise concerns, and advanced signal analyses [Engelke 2017].

  Such assessments are time-consuming, labor intensive and require specific skills from the experimenter. For instance, attaching EEG sensors to subjects is time-consuming, requires prior knowledge on how to correctly place electrodes and to ensure the connectivity between the scalp surface and the electrodes. The use of highly conductive gels or electrolyte solution, the limitation of having a wired network of electrodes and the need to amplify the captured electric potentials are a few examples of other constraints to overcome [Moon 2017]. Experimenters are forced to adapt the experiment design to its duration, setup, and the foreseen discomfort and boredom of subjects. They also need to carefully measure the additional time dedicated to equipment preparation, hygienic practices and preparation of subjects, when compared to usual subjective tests.

  Characteristics of physiological responses typically correspond to those of noisy and distorted signals [Engelke 2017]. For example, only small electric potentials are measured on the scalp surface with EEG. Equipment sensitivity and amplification, conversion and filtering processes applied to signals explain the noise introduced or captured by the recording system [Moon 2017].

  Besides, undesired information is present in EEG recordings, such as eye and body movement artifacts and other unrelated activities [Arndt 2016]. An appropriate and sophisticated processing is then required for noise-cancellation and removal of artifacts.

  Pre-processing the signals also includes filtering, bandwidth extraction, dimension reduction, etc. The main processing practices for physiological signals are described later in this section. A strong background in signal processing and understanding of physiological signals is required to process and interpret physiological responses correctly.

The construct validity of subjective evaluation hinges on highly controlled experimental conditions. The experiment must present experiences that are interpreted as similarly as possible by all observers. The test aspects of concern are multimedia stimuli, the environment, test conditions and interactions between the experimenter and subjects. Such requirements are magnified in case of physiological measures.

The connection between a test condition and observed responses should be as indisputable as possible. Indeed, non significant additional activities (e.g., eye and body movements) are contained in physiological signals. Results should be free from these irrelevant body activities.

All aspects of the experimental conditions should be similar and must only differ with respect to influence variables under test [Insko 2003]. In [Bosse 2016] regarding BCI applications, authors strongly suggest to give particular attention to three aspects of experiment design: (1) variability of conditions per subject (number of trials per subject), (2) variance of conditions over all subjects (number of subjects) and (3) sensitivity (what kind of distortions can be differentiated).

- The monitoring equipment impacts the performed analyses, as already brought up in the previous point. High-quality equipment is expensive [Insko 2003]. It includes numerous specific devices such as monitoring equipment (e.g., electrode network, lead-wire, amplifier and box integrating multiple physiological entries), side material (e.g., a set of disposable and spare electrodes, conductive gel and components for the electrolyte) and a powerful computer with high storage capacity.

  Even though this equipment provides signals of high quality and resolution, it brings out discomfort to subjects. To set the equipment and place sensors on subjects takes more time, which ultimately increases the experiment duration. More convenient and user-friendly systems are now available thanks to the current development of wearable devices (Emotiv, OpenBCI, NeuroSky, Mitsar portable EEG system, Avatar EEG, and Melomind) [Moon 2017]. For obvious reasons, these systems provide signals of inferior quality (e.g., less spatial resolution, noisier signals, and no synchronization during signal acquisition). Hence, a trade-off between cost, quality of signals and comfort level of subjects is made when selecting the recording system for physiological signals [Engelke 2017].

- Many challenges inherent to the fast growth of media technologies are not compliant with physiological signals. Several multimedia experiences go beyond passive content in developing interactive content (VR and AR) or in enabling interaction between users while viewing a content (social media [Tian 2010]). Physiological responses to such experiences cannot be measured [Engelke 2017]. The additional physiological response to movement and interaction increases drastically the complexity of signal analysis. Nowadays,

researchers focus on current issues and developments. Here, we refer mainly to the identification of QoE high-level cognitive processes.

To summarize, despite having drawbacks, physiological assessments are highly promising. They provide implicit assessment free from subjective test biases. The recorded signals are continuous, non-invasive, non-disruptive and highlight internal perceptual and cognitive processes. Results from physiological analyses have a high likelihood to be reliable, valid, objective and sensitive. However, several considerations should be kept in mind when dealing with physiological signals. Physiological analyses require extensive efforts in their implementation. The experiment design, its execution, and the processing of signals are complex and need to be adjusted to physiological monitoring constraints. The equipment is selected based on a trade-off between quality, comfort of subjects and cost of expenses. Lastly, such an assessment is not applicable to some multimedia applications, such as interactive content experiences.

6.1.1 Definition

Physiological assessments have been introduced in Chapter 2.3.8. The distinction between CNS and PNS have been presented, along with the most recent modalities to observe physiological responses. More details are given below regarding the physiological signals used in this thesis. The motivation of using them and their characteristics are indeed crucial information.

6.1.1.1 The Central Nervous System (CNS) activity

There are different alternatives to observe the CNS activity. The most popular non-invasive modalities are EEG, Magnetic Resonance Imaging (MRI), fMRI, and Positron Emission Tomography (PET).

MRI is a magnetic resonance imaging technique. The nuclear magnetic resonance of protons indicates the density of protons in a specific area, enabling to recreate 2D and 3D brain mapping with high spatial resolution, but low temporal resolution. Such a method is valuable for medical applications but not for subjective evaluations as we are interested in brain activation.

fMRI is also a magnetic resonance imaging modality. This approach detects changes in blood flow, an indicator of active areas in the brain. It is based on the energy consumption process of activated neurons. This energy comes from the glucose and oxygen brought by blood. The exchange of oxygen between blood cells and neurons causes a change in the magnetic properties of blood. This effect is compounded by the fact that the blood flow is increased in active brain regions to deliver more oxygen and glucose. The modification in magnetic resonance, due to variation in blood flow, is captured by fMRI. This is an indirect way to observe the electrical activity in the brain.

PET detects the gamma-ray emission of glucose substitute (tracer), introduced in the body beforehand. This emission is the consequence of glucose uptake by

the brain. Images of tracer concentration in 3D or 4D space (inclusion of time dimension) are then computer-generated. The PET allows one to determine the brain regions using a lot or little sugar, indicating brain active areas.

EEG records electrical impulses from the brain through numerous electrodes placed on an individual's scalp. The activity of sets of neurons is the cause of observed electrical responses. This technique has a high temporal but low spatial resolution. Indeed, temporal information can be recorded with millisecond precision while the number of electrodes placed on an individual's scalp is restricted (from 4 up to 512 electrodes). This is a major difference to the CNS measures described above. fMRI and PET have a temporal resolution in the second or minute time range. Another distinction is that EEG directly records the brain electrical activity, whereas fMRI and PET indirectly measure brain electrical activity through changes in blood flow or metabolic activity.

The reason why most physiological analyses in the multimedia field are conducted with EEG is the fact that spatial resolution is less crucial than the temporal one. Also, most studies focus on spectral analyses or relate physiological responses to cortical areas. In this regard, there are four main regions in the cerebral cortex which are activated by specific stimulation [Bonmatí Coll 2016]:

- the frontal lobe. Located at the front of the brain, this cortex controls motor skills, high-level cognition (e.g., reasoning, problem-solving and planning) and expressive language.
- the temporal lobe. This bottom section of the brain is associated with hearing, the formation of memories and the interpretation of sensory cues.
- the parietal lobe. Located in the middle of the brain, this lobe processes tactile sensory information (pressure, touch, pain and temperature).
- the occipital lobe. At the back of the brain, this lobe is associated with visual processing as it receives and interprets information from eye retina.

In the literature, quality and QoE are conceptualized based on various hypotheses regarding internal cognitive processes and experiences. Establishing a connection between the activity of cortical regions and QoE yields new information or confirms previously made assumptions. We can thus conclude that EEG, and physiological signals in general, help to improve our knowledge and to validate hypotheses about the internal responses of viewers during multimedia consumption.

6.1.1.2 The Peripheral Nervous System (PNS) activity

The PNS is divided into two main branches:

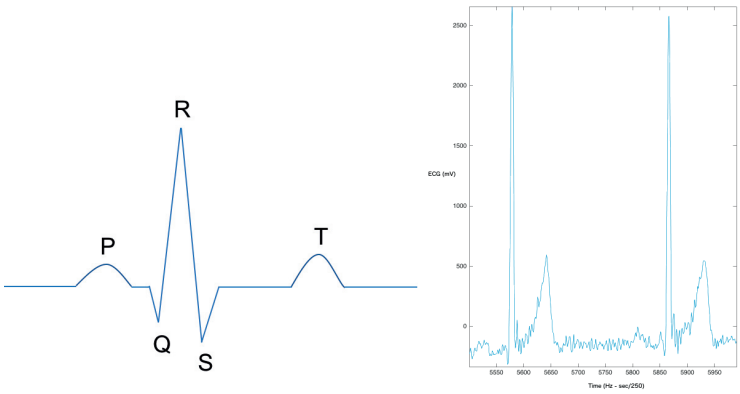
The first one, the SNS, relates to the voluntary control of organs and muscles. It includes the sensory nervous system, which collects sense information (e.g., touch and taste) and transmits it to the brain and spinal cord.

The second is the ANS, which regulates the physiological involuntary responses. This system has two functioning states, activating either the sympathetic or the parasympathetic nervous system.

1. The sympathetic nervous system is activated when encountering a "fight or flight" situation, also called hyperarousal or acute stress response. This system releases neurotransmitters which increase heart rate and blood flow in certain regions while decreasing irrelevant and non-critical functions.

2. The parasympathetic nervous system activation corresponds to a "rest and digest" period.

The interest in PNS lies in the opportunity to study the sensory nervous system responses, especially those related to touch, sight and hearing. Also, the sympathetic or parasympathetic activation of the ANS appears to indicate the physiological response to events and to bear witness to individuals' arousal and stress. Several peripheral signals can be recorded, as mentioned in the Chapter 2.3.8. We focus here on ECG and respiration as those modalities were used in our studies.



(a) Schematic representation of a PQRST complex in normal EEG

(b) Two PQRST complex in real ECG signals

Figure 6.1: Schematic and raw signals depicting the PQRST complex in ECG.

ECG

We have seen previously that an increase in heart rate is directly indicative of the ANS state. This justifies the wide use of ECG as a peripheral signal in physiological analyses.

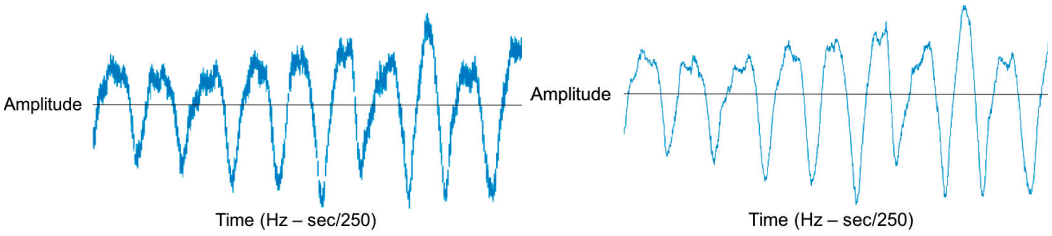
To understand how to record and process the cardiac activity, one should first know how the human heart operates. The human heart is composed of two atria (small cavities on top of the heart where the blood enters) and two ventricles (lower chambers which expel blood). The sinoatrial node, a cluster of cells considered as the heart's natural pacemaker, is located in the right atrium. Another cluster of cells, the atrioventricular, is located at the center of the heart, between atria and ventricles.

A depolarization process triggers heartbeats. An impulse is initiated in the sinoatrial node. The excitation wave follows specialized conduction channels, which induce a contraction of the atria. Consequently, the blood in the atria is pushed down in ventricles. The impulse reaches the atrioventricular node. The functionality of this node is to delay (by approximately 0.09 seconds) the excitation before its transmission to ventricles. This delay is important as it ensures that the ventricular contractions take place after the atrial contractions. These ventricular contractions force blood out of the heart, to the lungs and body. A new sinoatrial impulse triggers the next heartbeat.

ECG monitoring captures this polarization pattern through currents generated by the depolarization and repolarization of cardiac cells. Depolarization refers to the process of cells losing their negativity while repolarization stands for cells returning to resting polarity. This complex is depicted in Figure 6.1 where one can observe that a heartbeat pattern is composed of:

1. a P wave, indicative of the atrial depolarization,
2. a QRS complex which represents the ventricular depolarization, and
3. a T wave, corresponding to ventricular repolarization.

In the context of multimedia QoE, ECG activity is thought to relate to arousal, tiredness and discomfort [Kroupi 2014c].



(a) Real raw respiration signal (b) Pre-processed respiration signal

Figure 6.2: Pre-processed and raw respiration signals.

Respiration

The function of respiration is to collect the oxygen in the air and bring it to the body. It is also responsible for carbon dioxide rejection.

During inhalation, diaphragm and intercostal muscles contract and move downwards or aside. These movements increase the chest cavity space into which lungs expand. The expansion of lungs sucks in air containing the oxygen needed by the body. The transfer of gas is operated in alveoli rich in capillaries (blood vessels), enabling the exchange of oxygen and carbon dioxide. The relaxation of the diaphragm and intercostal muscles reduces the thoracic cavity, forcing out the air that lungs contain, rich in carbon dioxide. This process is called exhalation.

Respiration is usually an involuntary process satisfying humans' basic needs. Yet, it can be controlled, such as when holding breath, sneezing, singing, etc. Hence, respiration is related to both ANS and SNS. Similarly, both sympathetic and parasympathetic activities affect respiration.

Respiration is the measure of fluctuation in the amount of air contained in the lungs. It can be monitored in several ways. Thermal infrared imaging is a passive non-contact modality that can be used for breathing flow analysis [Fei 2006]. However, a more common and reliable modality for respiration analysis is a set of plethysmograph belts [Wu 2009]. They capture the change in volume within the thoracic cavity and the abdomen. Figure 6.2 shows recorded and smoothed signals from a thoracic belt.

Combination of physiological signals

Current and future multimedia technologies provide rich and diverse applications and experiences. Hence, one can expect that QoE evaluation with one physiological modality does not capture the entire physiological response of all applications. Besides, it suffers from the limitations of the selected modality (e.g., low spatial or temporal resolution). The combination of several modalities is a promising option to tackle this issue. A multimodal collection is likely to provide information from various nervous systems and to overcome limitations of a single modality. Additionally, signals are complementary and tend to cover the physiological response of the whole body to a stimulus.

Insights into a wide range of cognitive processes make possible more precise inferences about users' experiences. Also, they may capture high-level QoE concepts such as immersion, engagement and SoP [Engelke 2017].

6.1.2 Processing of signals

This section introduces state-of-the-art processing of biosignals, from pre-processing (noise and artifact removal) to best practices to extract meaningful information from signals.

6.1.2.1 EEG

A first challenge to overcome in EEG use is pre-processing. It aims to identify and remove artifacts from brain signals. The most common artifacts are Electrooculography (EOG), Electromyography (EMG) and ECG generated by ocular, muscular and cardiac activities, respectively. Other distortions may be introduced by various influencing factors (e.g., the equipment or the environment).

Removal of artifacts

Numerous artifacts can be introduced in EEG signals which are subject- or equipment-related [Teplan 2002]. Subjects-related artifacts can be caused by any minor body movement, EMG, ECG, EOG and sweating. Technical limitations may

appear due to impedance fluctuations, cable movements, broken wire contacts, low connectivity between electrodes and scalp or 50/60 Hz power line interference signal and its harmonics.

Some technical limitations can be addressed by applying a notch filter to remove power line interferences and rejecting malfunctioning or high impedance electrodes. This last point influences the electrode system selection. For instance, most studies are conducted using the 10-20 electrode placement system, assuming a sufficient spatial resolution. If possible or necessary, standardized systems with up to 512 electrodes (32, 64, 126 systems) exist.

Considering subject-related artifacts, a first bottleneck is that involuntary eye-movements and blinking artifacts strongly impact EEG signals, especially in the frontal and central regions. Indeed, the magnitude of such interference signals is more significant than that of brain-generated electrical potentials. This necessitates the removal of blinking and EOG artifacts.

A first approach is the rejection of the contaminated signal segments. An expert manually scans the recorded signals and excludes the parts containing artifacts. Rejection can be a cumbersome procedure. Moreover, the validity of this approach is questionable as its accuracy is hardly predictable. The expert can easily neglect non-obvious artifacts, especially when dealing with numerous or lengthy signals. The second drawback in rejection is the reduction and loss of data.

The second approach is the correction of signals. One may apply correction by using calibration data or may use the signals to estimate correction coefficients. Calibration trials, recorded before test sessions, extract eye movement and blinking responses from subjects. During these trials, an individual makes large eye-movements in a controlled, calm and neutral environment. The recorded patterns form the baseline representations of artifacts. Then, artifact correction consists in a least square regression, based on the baseline of artifacts [Croft 2000].

Regarding blinking artifacts, it has been shown that eye blink rate is related to visual fatigue [Bang 2014]. However, these artifacts cause data loss in terms of cerebral activity. As a matter of fact, during blinking EEG signals are not observable anymore.

Current practices for manual or automatic removal of eye-related activity from EEG project the signals in a sub-space. Independent Component Analyses (ICA) or Principal Component Analysis (PCA) algorithms are widely used [Joyce 2004, He 2004], since they decompose the signal into independent and uncorrelated components, respectively. Experts or specialized algorithms identify and remove the components related to EOG located in the frontal and central regions.

Figure 6.3 presents the ICA decomposition in 32 components of an EEG signal, through the EEGLAB open-source Matlab toolbox dedicated to physiological research [1]. An expert eye would spot component 3 as an eye artifact. The strong far-frontal projection is typical of eye artifacts. The spacial location observed in

[1] https://sccn.ucsd.edu/~scott/ica.html, http://cognitrn.psych.indiana.edu/busey/temp/eeglabtutorial4.301/maintut/ICA_decomposition.html

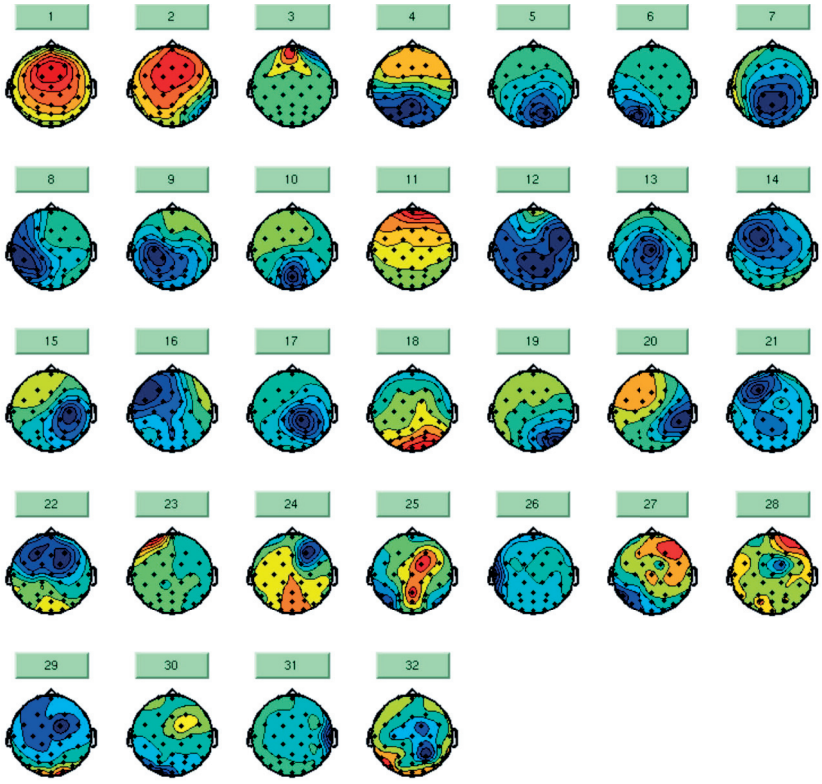


Figure 6.3: Example of EEG decomposition using ICA. Removal of components 3, 23 and 31 responsible for EOG, noise and EMG artifacts, respectively.

component 31 shows muscle activity. Finally, when a single channel is active, like component 23, it indicates a noise component. Those three components should be rejected.

When dealing with EMG, artifacts can be avoided, or at least minimized, when recording signals. Subjects can be trained not to move their head and keep a neutral facial expression. Though, in the context of multimedia experiences and especially QoE, we cannot prevent subjects to laugh for instance. Hence, in the same manner as with eye movements, facial muscle activity and voluntary breath activity (e.g., snort, laugh) can interfere with EEG. Calibration trials are again the practice indicated for the removal of these artifacts.

It is strongly recommended, though not mandatory, to record calibration trials following the steps below. To record baseline signals, subjects have to:

- close eyes for 5-7 seconds,
- keep eyes open for 5-7 seconds,
- look up/down/left/right and move eyes back to the center, five times each,
- blink, five times,

- clench their teeth, five times,
- if possible, move ears, three times, and
- snort, twice.

Pre-processing should strike a delicate balance between artifact removal and loss of essential information. Some researchers choose not to perform artifact removal in order not to impact EEG data, which are difficult to discriminate from unrelated artifacts. Otherwise, one should limit pre-processing to simpler artifact removal methods, such as bandpass filtering or replacement of the artifacts using interpolation. Both approaches were adopted in this thesis.

Extraction of EEG features

Literature surveys [Engelke 2017, Arndt 2016] revealed that most studies rely on well-established data analysis techniques in the field of physiological measurements. Spectral analysis, functional and effective connectivity, and machine learning approaches are widely used and are described below. However, the most widely used method, Event-Related Potentials (ERP), is not examined in this thesis. Considered as one of the most informative and dynamic approaches for brain monitoring, they reflect the reception and processing of sensory information in response to an event. Time precision (e.g., event timestamps or perfect synchronization between the stimuli presentation and the event marker) and high resolution (variation analyses in milliseconds) are mandatory. Extracted features are event response polarity, latency and scalp distribution [Romero 2015]. ERP are specifically indicated for stimuli of short duration [Arndt 2016]. This aspect and the type of extracted features are not fully in line with QoE or high-level QoE evaluations of concepts. Indeed, QoE, engagement, immersion and presence appear in long-duration stimuli and are not necessarily triggered by an event onset.

Oscillatory system	Frequency band (Hz)	Related brain activities
Delta	0.5 - 4	Deep sleep
Theta	4 - 8	Information encoding and retrieval
Alpha	8 - 12	Relaxation phases, especially with eyes closed
Beta	13 - 30	Increased alertness and focused attention
Gamma	>30	Not well understood yet
Low beta	13 - 16	
Middle beta	17 - 20	
High beta	21 - 30	

Table 6.1: Brain waves frequencies and brain activities they indicate [Sawyer 2011]

Spectral analysis

Extensive work in psychophysiology showed a connection between oscillatory phe-

nomena and functional EEG. Brain scientists recognize that brain's natural oscillations govern brain activities. In particular, alpha, theta, delta and gamma waves are responsible for sensory and cognitive brain functions [Başar 2001]. We here refer to functions of perception, movement, attention, learning, and memory. Table 6.1 presents the frequency bands for the alpha, theta, delta and gamma frequency bands and the activity they are related to.

- Audio and visual stimulation elicits alpha activity, which has been proven to be strongly correlated with working memory and probably long-term memory engrams [Başar 2001]. In relaxation or drowsiness, the alpha activity rises.
- Beta waves dominate signals in normal states of wakefulness with open eyes or information processing (e.g., thinking and calculating) [Başar 2001]. Beta waves are associated with active thinking, active attention, focusing on the outside world or solving concrete problems [Nidal 2014]. A panic state may also induce a high level of beta waves.
- Concerning the gamma frequency band, it has been shown that this oscillatory system is a component of the human auditory and visual response and that 40 Hz responses indicate attention-related responses in humans [Başar 2001].
- The delta and theta systems correlate in a similar manner to cognitive processing: signal detection, decision making, selective attention. For instance, delta oscillations have been observed in response to stimuli at the hearing threshold [Başar 2001].

Cognitive neuroscientists typically focus on the study of alpha, beta, and gamma waves [Sawyer 2011].

From these oscillatory systems, several features can be extracted: signal power, mean spectral power, peak frequency, peak frequency magnitude, mean phase angle, mean sample value, zero-crossing rate, number of samples above zero, and mean spectral power difference between two input channels are also calculated in every frequency band [Nidal 2014].

Welch's method [Welch 1967] is commonly used to estimate the Power Spectral Density (PSD), as suggested in [Kroupi 2014a]. Welch's method involves the averaging operation of overlapping periodogram estimates. To do so, Hamming windowing is applied to EEG signals. The interest behind Welch's method is that the averaging operation reduces the periodogram variability. In the context of EEG, before each stimulus, a baseline period (a ten-second long presentation of a centered white cross over a black background) is recorded. The logarithm of the mean baseline power is then subtracted from the logarithm of the mean trial power to remove variations that are not related to the stimulus.

Functional and effective connectivity

Brain activity can be examined from two perspectives, functional or effective connectivity. Effective connectivity, in hypothesis testing, aims to model the brain

activity. Functional connectivity is based on empirical data and seeks to establish significant correlations between two areas or to find predominant patterns of correlation. This approach offers the exciting possibility to predict or classify physiological responses [Friston 2011]. This approach is more compliant with the need for QoE objective measures.

Current practice in functional connectivity is to apply Dynamic Causal Modeling (DCM) or Granger causality and hold to basic neuroscience. A particularly interesting technique is Granger causality applied to small-world brain networks. The high clustering and short path length of small-world networks make it a promising asset to model brain connectivity. Empirical and technical knowledge supports this approach for several reasons:

1. "The brain is a complex spatial and temporal network, on multiple spatial and temporal scales.

2. The brain supports both segregated and distributed information processing. Sensory and cognitive processing may be localized in specialized regions or distributed in large-scale areas.

3. The brain appears to maximize the efficiency, and minimize the cost of information processing " [Bassett 2006].

Previous studies have shown that EEG features can be derived from small-world network characteristics [Rubinov 2010], that are extracted from the functional connectivity map estimated by Granger causality [Liao 2011]. The main features extracted from the estimated functional connectivity maps are the characteristic path length [Watts 1998], the global efficiency [Latora 2001], the clustering coefficient [Watts 1998], and the local efficiency [Latora 2001].

For more information on functional and effective connectivity as well as on DCM and Granger causality, please refer to [Friston 2011], which reviews these approaches in a highly comprehensive manner and presents a detailed review on brain connectivity.

Deep learning Machine learning-based physiological processing has become a new trend among researchers. This is a powerful tool for learning features from experimental data, especially for multimodal data. Such processing opens the door to a whole field that is out of the scope of this thesis. Hence, no further information is provided here.

6.1.2.2 ECG

ECG pre-processing

A wide range of variables may be extracted from ECG measurements. Among these, the most common are HR and Heart Rate Variability (HRV). The HR is a measure of contraction frequency, i.e. the number of heartbeats per minute (bpm). This measure can indicate physical exercise, sleep, anxiety, stress or illness. The HRV measures the variation in time intervals between consecutive heartbeats. Also named RR variability, it is computed by measuring the time difference between the R peaks of two successive QRS complexes. We used the HRV in this thesis, to explore heart rate activity beyond HR.

One first needs to detect the R peaks, compute the R-R intervals and resample the non-uniform R-R intervals in a time-series. The localization of QRS is usually performed following the real-time algorithm developed by Pan and Tompkins [Pan 1985]. The latter consists in applying band-pass and five-point differentiator operators to the signal, in this order. Those operations are followed by a squaring process which intensifies the derivative responses and brings more accuracy to the entire algorithm. A moving window integration operation extracts the slope and amplitude, and width information of QRS complexes. Figure 6.4 shows the detected R peaks in raw signals.

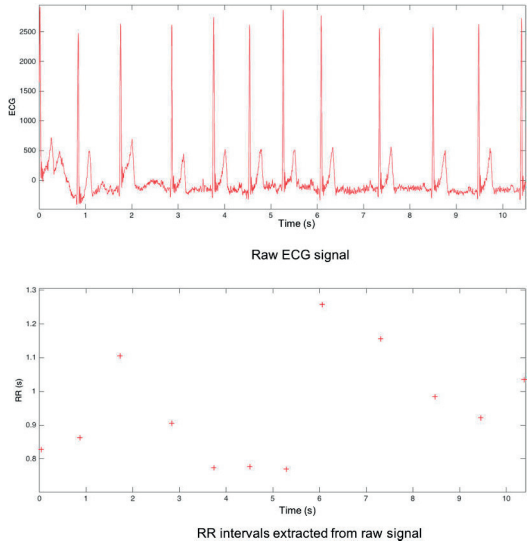


Figure 6.4: Processing of ECG signals: extraction of R-R intervals.

Extraction of ECG features

Mean and standard deviation are extracted from HRV signals. The mean and std of raw signals $X_n, n \in [1, N]$ are defined in equations 6.1 and 6.2, respectively.

$$\mu_x = \frac{1}{N} \sum_{n=1}^{N} X_n \tag{6.1}$$

$$\sigma_x = \sqrt{\frac{\sum_{n=1}^{N}(X_n - \mu_x)^2}{N-1}} \tag{6.2}$$

Regarding the frequency domain of HRV, the power in the Low Frequency (LF) (0.05-0.15 Hz), High Frequency (HF) (0.15-0.5 Hz), and the LF/HF ratio are also extracted [Bilchick 2006, Jang 2015]. The LF/HF ratio describes the sympathetic versus parasympathetic balance. The signal duration impacts the reliability of the processing of HF and LF bands. According to the committee report of the Society for Psychophysiological Research [Berntson 1997], one-minute long epochs are reliable to compute the HF component while two-minute long epochs are suggested to calculate the LF component. Those durations correspond to approximately 6-10 times the period of the wave with the lowest frequency (0.15 Hz for HF and 0.05 Hz for LF). However, it is still possible to determine the energy in both bands from epochs of shorter duration, although less reliably [Chanel 2009].

6.1.2.3 Respiration

Respiration pre-processing

Respiratory signals do not necessarily need pre-processing. However, a noise canceling operation through a wavelet multivariate de-noising [Aminghafari 2006] can be applied to both respiratory signals (abdomen and thoracic). Figure 6.2b shows processed signals.

Respiration features

The features extracted from respiration signals are its mean and std, low (0.05-0.25 Hz) and high (0.25-5 Hz) spectral power and the ratio of the two spectral powers [Koelstra 2012]. More general respiration features consist of respiration rate and average power across the 0.1 to 0.4 Hz frequency band [Kroupi 2014a].

6.1.3 Feature fusion

It has already been explained why multimodal analyses are better than unimodal ones. The fusion operation implied by multimodal analyses is either feature-based or decision-based.

In the first case, features of various modalities are combined prior to classification. In the second, a classifier is learnt for each modality and results of all classifiers are combined afterwards for final decision making.

In order to implement feature-based fusion, the features of each modality must be pooled. This operation is named fusion and usually consists in the concatenation of the feature vectors.

Decision fusion schemes are numerous: the product rule, min rule, max rule, sum rule or median rule, to name but a few. All rules depend on the posterior probabilities of the classifiers, which refer to the confidence that is assigned by each classifier as a label to an instance.

For further details on physiological analysis and feature extraction, the reader can refer to the work performed in [Kroupi 2014a] by which we have mainly been inspired.

6.1.3.1 Classification

After having extracted and possibly merged features, machine learning techniques are used for classification purposes. Classification refers to the prediction of the stimulus class through the observations performed (features). Supervised learning corresponds to the case when a ground truth is available to train the model. This involves creating two sets of observations: a training set used to learn the model and a test set, independent from the training set, which assesses the performance of the model and verifies whether there is data overfitting. Data overfit indicates that the learnt model has derived training-set specific correlations instead of general inferences. Such a behavior should be prevented. In the context of physiological analyses, subjective ratings are used as ground-truth to conduct a supervised classification.

Additionally, the number of features should be limited to prevent the curse of dimensionality: classification errors increase when the number of features becomes too large [Hua 2004].

Widely used classification schemes are briefly described below.

The k-Nearest Neighbor (kNN) classifier is a non-parametric method, which considers the assigned classes of the k training instances closest to a new observation. The decision is made by majority voting. Selecting k, the number of neighbors to consider, is usually done heuristically. The selection of a proper distance measure can influence the results. However, the distance between instances is usually the Euclidean distance metric [Hastie 2001]. The k-NN classifier is not able to capture complex structures in data because of its local nature, which indicates the need for more advanced classifiers.

A Support Vector Machine (SVM) classifier is a non-probabilistic binary classifier. A set of labeled instances are projected to a high dimensional space. Instances are then separated by a hyperplane, created by maximizing the margin between different classes. A new instance is categorized according to which side of the hyperplane it belongs to. Typically, SVMs perform linear classification, but they can perform non-linear classification using kernels. The Radial Basis Function (RBF) kernel is generally used to consider the non-linearity of the relationship between class labels and features. It also requires fewer hyper-parameters than other kernels, reducing the complexity of model selection as well as parameter tuning.

The efficiency of a classifier is evaluated through precise performance metrics. An extensively adopted visualization tool is the confusion matrix (or contingency table). An example of a confusion matrix is presented in Table 6.2.

Sensitivity (or recall, see Equation 6.3) , specificity (or TP, see Equation 6.4) and accuracy (see Equation 6.5) are statistical measures of performance of a binary classifier (two class discriminant). Sensitivity represents the proportion of instances correctly classified as positive; specificity refers to the proportion of correctly clas-

	Actual positive	Actual negative
Predicted positive	True Positive (TP)	False Positive (FP)
Predicted negative	False Negative (FN)	True Negative (TN)

Table 6.2: Example of a confusion matrix

sified negatives. Accuracy indicates the proportion of well-classified samples. A perfect predictor would achieve 100% sensitivity, specificity and accuracy. A random classifier (with balanced classes) would achieve sensitivity and specificity of 50%.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{6.3}$$

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{6.4}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.5}$$

To assess further the performance of a classifier, especially in the case of scarce data, cross-validation is an accurate model validation [Hastie 2001]. In cross-validation, one splits the data in K folds, roughly equal-sized, uses $K - 1$ folds as the training set and the remaining one as the validation set. Performance metrics are applied to the validation fold. This process is repeated until each fold has been the validation fold. The final performance score of the classifier is the average performance across all validation-folds. This process is referred to as K-fold cross-validation. Typical choices of K are 5 or 10 [Hastie 2001]. When the validation set is composed of one instance ($K = 1$), the model validation is named leave-one-out cross-validation. The advantage of this last method is that the classifier is learnt on the maximum number of possible training sets, and hence is more likely to find the best fitting model. However, the variance of the classification performances may be increased, when compared to K-folds cross-validation.

6.1.4 QoE-related studies

Following the processing described previously, one can derive classifiers from physiological signals or examine the cognitive processes under the QoE. A rather limited number of work has been performed on the evaluation of QoE and its high-level aspects through physiological signals. However, publication trends show that more and more research is dedicated to the analysis of biosignals. For instance, three reviews on physiological signals analysis regarding QoE [Engelke 2017, Arndt 2016] and multimedia experiences (QoE, emotion, aesthetic satisfaction, preference, fatigue, attention, etc.) [Moon 2017] have been published in the last two years. They provide a comprehensive and complementary information on the development of physiological signals analysis in quality and QoE evaluation. In this section, the

works conducted on QoE and the high-level QoE concept, the SoP, are briefly summarized.

6.1.4.1 QoE

There are two ways to investigate multimedia experiences. The first approach aims at designing evaluations reproducing realistic multimedia consumption scenarios. This implies research on the impact of long stimuli impairments on observers' state. The second approach uses rather short stimuli and examines the difference in cognitive responses created by different media modalities (e.g., comparison of 2D and 3D contents) [Arndt 2016].

The instantaneous change in physiological responses to multimedia presentation has been the most extensively examined aspect in the analysis of biosignals [Antons 2012, Scholler 2012, Mustafa 2012, Lindemann 2011]. The central message conveyed by such studies is that abrupt quality variation in images, audio and videos are recognizable in the brain activity [Moon 2017]. As already motivated, QoE does not appear at the very millisecond following the stimuli presentation. Hence, these studies were not considered when designing our evaluations.

The QoE has been assessed on speech quality, video quality and a comparison between cross-modal and uni-modal perception of audiovisual stimuli [Antons 2013, Arndt 2014, Arndt 2013, Belardinelli 2004, Lassalle 2011]. Conclusions consistently point to a higher neural activity when participants are presented with lower quality or large degradations. When dealing with audio-visual stimuli, the memory and visual processing related to brain areas are more activated. It also seems that the type of multimedia may also have an impact on physiological responses [Moon 2017].

Several works explored the perceptual experience in specific emerging multimedia technologies (3D, VR and HDR). A comparison between 2D and 3D stimuli demonstrated that frontal asymmetry patterns in the alpha band are related to perceived quality, based on the correlation between subjective scores and power spectral density of alpha, beta (low, middle and high), gamma and theta frequency bands [Kroupi 2014b].

The sensation of reality in 3D videos has been evaluated [Kroupi 2014c] through EEG, ECG, and respiration signals. If a high correlation has been observed between ratings and EEG statistical features (same brain oscillations as above), a level slightly above chance has been reached when using peripheral features only.

In [Egan 2016], a comparison between VR and non-VR environments was conducted through subjective ratings, HR and EDA recordings. Conclusions indicated a strong negative correlation between EDA and arousal in a VR environment. However, no correlation was observed between HR and ratings.

When comparing SDR with tone-mapped HDR stimuli, [Moon 2015] measured implicitly QoE through EEG and peripheral physiological signals (GSR, respiration, heart rate, and skin temperature). Spectral statistical features were extracted from peripheral signals and the EEG theta, alpha, beta, and gamma frequency bands. It turned out that the power of the gamma frequency band was highly correlated with

the perception of tone-mapped HDR videos.

From those works, we can see that there is room for exploration concerning the measure of QoE through physiological signals. Apart from quality and a limited number of emerging technologies, various influence factors such as resolution, device used and environmental conditions are still to be investigated. A specific attention should also be given to more typical consumption scenarios.

It should also be mentioned that this research field is nowadays in an exploratory phase, focusing on providing proofs of concept for QoE measurements through analyses of physiological signals [Bosse 2016].

6.1.4.2 SoP

Numerous researches on QoE focused on measuring the intended effect of technology, such as sense of reality and presence. Such high-level QoE aspects are induced after a viewer consumes a content for a certain duration, through long-duration evaluations [Moon 2017].

Presence is the human response to immersion according to [Slater 2009]. Indeed, given different people, the same immersive system may elicit various levels of presence, and different immersive systems may give rise to an equal level of presence. In their review on presence and its measurements, the authors emphasized the strong link between presence and, to name a few, attention, emotions, involvement and engagement. There has been extensive research on physiological signals for emotion recognition [Jerritta 2011]. This motivates the increasing interest in using physiological signals to measure presence.

A first physiological analysis for presence was conducted on a stressful VE with ECG, skin conductance and skin temperature [Meehan 2002]. The authors proved the reliability, validity, and sensitivity of those peripheral signals, especially ECG, to subjectively and objectively measure presence in stressful VEs.

An event-related approach for presence evaluation in VR has shown that frontal negative slow waves in EEG are accurate predictors for presence experience [Kober 2012]. While in immersion in a VR environment, subjects were occasionally presented with task-irrelevant audio, supposedly decreasing their presence level.

To the best of our knowledge, no physiological analysis measuring presence was conducted in real environments (meaning non VE). However, SoP is an effect of every emerging technology and content on a viewer and contributes to its QoE. There is an apparent lack of extension of SoP studies in more typical multimedia technologies.

6.1.4.3 Standardization and databases

Regarding the standardization of acquisition and processing of physiological signals, standardization committees launched the P.PHYSIO ITU-T SG12 work program for speech processing (see Chapter 2.3.8) and for audiovisual quality Rec. ITU-T COM12 [ITU 2013b, ITU 2013d, ITU 2013c, ITU 2014b] and [Arndt 2016].

Concerning databases making available subjective ratings and physiological signals, the Database for Emotion Analysis using Physiological signals (DEAP) is one of the most widely used, and is dedicated to the affective response to music videos [Koelstra 2012]. In [Moon 2017], authors listed 7 datasets, mostly on emotion recognition. The datasets enabling QoE evaluations are PsySyQX [Gupta 2015] for speech analysis, providing EEG and functional Near-Infrared Spectroscopy (fNIRS) (an EEG-like brain waves signals acquisition method), and Yonsei-tHDRv [Moon 2015], with EEG, GSR, respiration, and skin temperature for tone-mapped HDR video experiences. We see here a clear lack of databases on more contemporary contents for TV.

6.1.5 Conclusion

Physiological analyses for QoE analysis and measurements are highly promising. Even though they require a specific and precise knowledge, they are a powerful indicator of subjects response to multimedia stimuli.

We have identified several deficiencies in the evaluation performed towards the prediction of QoE. For instance, most studies focus on quality or technology used. Several other parameters influence QoE and could be detected through biosignal processing. Additionally, most works do not consider typical multimedia consumer scenarios. Finally, new datasets with broadcasting-like contents are needed to extend researches on QoE with physiological signals.

We addressed these deficiencies, in the context of SoP prediction, in the two studies presented below.

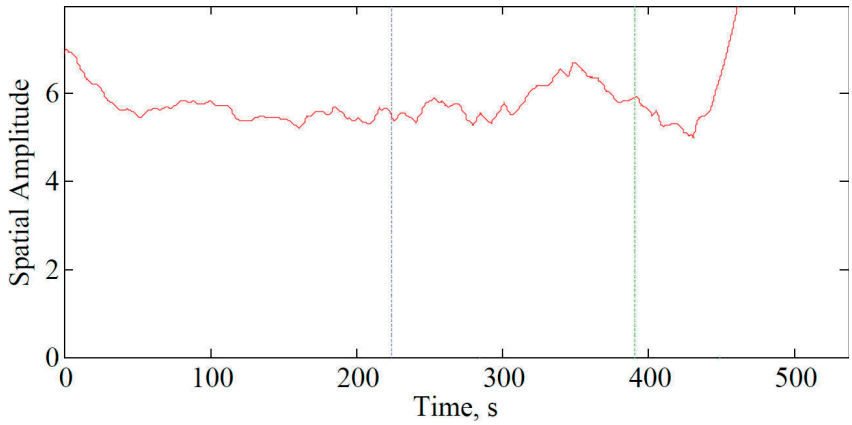
In both of them, we defined three different levels of presence (low, middle and high). In the first one, technological properties (such as quality, resolution and sound system) define the SoP levels. In the second, more typical consumption scenarios are considered and three presentation devices set the SoP levels.

The independent variables in both experiments are evaluated for the first time through physiological signals to the best of our knowledge. Additionally, publicly available datasets have been created during these works, contributing to the development of physiological analyses for QoE evaluations.

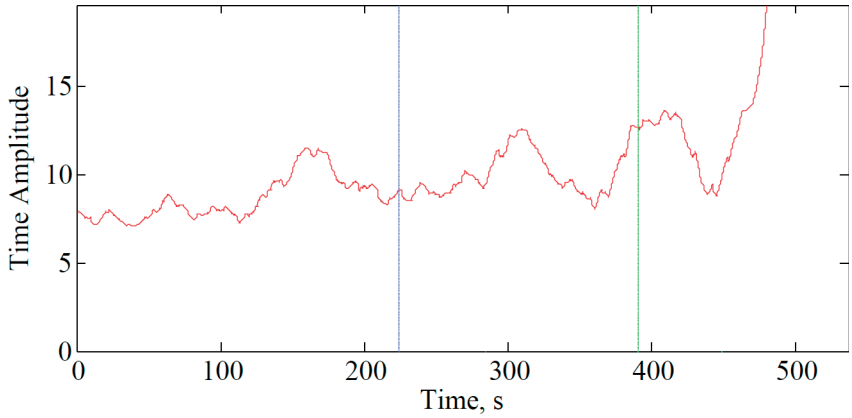
But before describing these two studies, information applicable to both pieces of research are presented below (i.e., contents and equipment).

6.2 Equipment and content

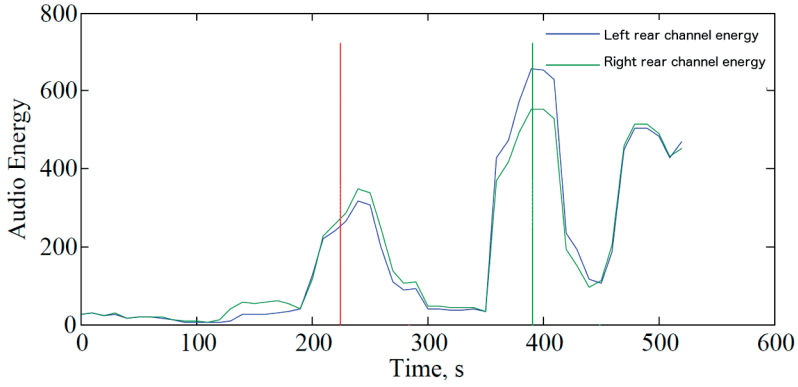
Before introducing our works, mutual information to both tests are presented in this section. Indeed, same test sequences were used in both studies, so was the equipment to acquire physiological signals.



(a) Spatial amplitude



(b) Temporal amplitude



(c) Audio signal energy

Figure 6.5: Selection of Big Buck Bunny test sequences through audio rear channels energy, SIs and TIs

6.2.1 Content preparation

We stated previously that more broadcasting and TV contents should be used in physiological analyses. Hence, we have created a test dataset composed of a collection of one-minute test sequences extracted from four Blender open source movies[2].

Uncompressed original movies were in YUV 4:2:0 8-bit format at 24 fps, either in HD (Big Buck Bunny and Elephant Dream) or in UHD resolution (Sintel and Tears of Steel). Audio signals were available in stereo and 5.1 surround and were represented in FLAC format.

A total of ten test sequences were produced for the data set. Out of these sequences, nine (C1-C9) were used for test stimuli production, and the remaining one (C10) was allocated to training stimuli creation.

Sequences duration of one minute resulted from a trade-off. This duration must be sufficiently long to lead to presence effects. However, stimulus duration is limited because of constraints in subjective evaluations, especially concerns about subjects boredom and lack of attention.

Test sequences were selected based on a careful analysis of the entire original movies. More specifically, the temporal and spatial characteristics of the luminance component along with the energy level in surround audio channels of each original movie were considered. In particular, rear audio channels were used during the analysis to discriminate surround sound to stereo sound in sequences better.

We computed SIs, TIs and audio energy levels of left and right surround channels for each second of movies. Windowing is applied to signals. That is we summed the measured quantity over a time window of the targeted stimuli duration.

Selected test sequences generally correspond to the highest values of the properties mentioned above (TI, SI, and audio energy), whereas the scene cuts of the original movie were also taken into account. In particular, the selection was made to prevent abrupt changes at the beginning and the end of each test sequence.

Figure 6.5 shows the spatial, temporal and audio energy analysis for the original movie Big Buck Bunny. The blue line (red in the sound analysis) represents the beginning of the first sequence while the green line shows that of the second.

An example frame of each test sequence used in experiments is presented in Figure 6.6. Additional information for each test sequence including its original movie, spatial resolution, and the position of the first frame within its original film is reported in Table 6.3.

An analysis of the TI, SI, and audio energy was subsequently conducted on the selected test sequences to better understand their properties, namely, spatial complexity, amount of motion, and impact of the rear audio signals. We verified the variety in selected contents.

The distribution of test sequences along their SI and TI values are presented in Figure 6.7.

It is observed that test sequences extracted from the movie *Sintel* have large TI and small SI values. Thus, the C5, C6 and C10 test sequences contain much motion

[2] `http://media.xiph.org/`, `http://www.blender.org/foundation`

(a) C1 (b) C2 (c) C3

(d) C4 (e) C5 (f) C6

(g) C7 (h) C8 (i) C9

(j) C10

Figure 6.6: Typical frame of each test sequence used in experiments. Sequences C1 - C9 (a-i) are used for testing and sequence C10 (j) is used for training.

and a low level of detail.

The test sequences originating from the movies *Elephant Dream* and *Big Buck Bunny* have small TI and SI values. This means that there are few motions and a low level of detail in C1, C2, C3, C4, and C9 test sequences.

The test sequences extracted from *Tears of Steal* present high SI and fair TI values. Hence, C7 and C8 test sequences contain few motions with detailed spatial information. This can be explained by the fact that the three first movies are computer-generated, whereas the last film contains real-world environment with additional computer-generated information.

The analysis of audio channels is presented in Table 6.4. Values E_L and E_R express the level of audio energy in percentage as a ratio between sound energy of left and right channels (both front and rear) and total audio energy (sum of all channels). Moreover, values E_{SL} and E_{SR} represent the ratio between the level of audio energy of the rear channel and side channels (both front and rear) in percentage for each side.

The general conclusion from these numbers is that, on average, a relative balance

Sequence	Original movie	Original resolution	Start frame
C1	Big Buck Bunny	1920x1080	5388
C2	Big Buck Bunny	1920x1080	9379
C3	Elephant Dream	1920x1080	2160
C4	Elephant Dream	1920x1080	6516
C5	Sintel	4096x1714	6900
C6	Sintel	4096x1714	12607
C7	Tears of Steal	4096x1744	4152
C8	Tears of Steal	4096x1744	10392
C9	Elephant Dream	1920x1080	11280
C10	Sintel	4096x1714	360

Table 6.3: Original movie, original resolution and start frames of test sequences

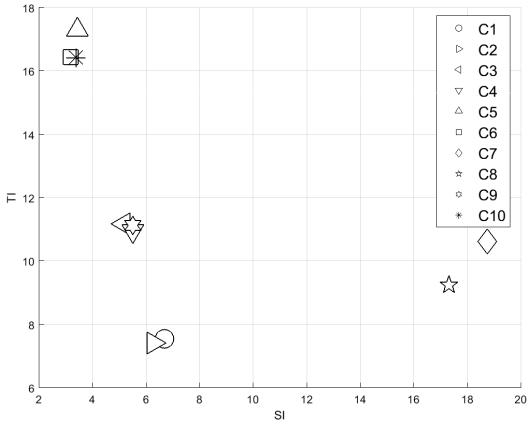


Figure 6.7: Values of SI and TI of test sequences

is observed between left and right audio signals. Test sequences having the highest amount of information in surround channels are C8, C5, and C10. Test sequences presenting the lowest amount of data in the rear left, and right channels are C1, C2, and C4.

After the extraction of sequences, stimuli are processed to correspond to the three SoP levels, also referred to as ILs[3].

Different processing is applied to generate test and training stimuli to match the design and purpose of experiments.

- In the first study, experts defined low, middle, and high ILs based on the used audio sound system (no audio, stereo, and 5.1), the video quality/level of compression (high and low QP), and the resolution (UHD, HD, and SD). Table 6.5 reports the settings of the three determined ILs.

[3]This naming convention is suitable as presence is the response to immersivness level, according to [Slater 2009].

Sequence	E_L[%]	E_R[%]	E_{SL}[%]	E_{SR}[%]
C1	38.12	48.16	1.06	0.91
C2	34.87	34.88	4.97	4.19
C3	45.51	42.48	15.29	15.06
C4	40.21	39.45	4.42	4.38
C5	37.84	38.11	23.01	24.17
C6	28.15	26.89	11.67	14.78
C7	47.36	44.99	12.94	13.49
C8	45.87	51.49	20.52	19.79
C9	36.35	32.61	6.22	6.35
C10	29.92	30.56	26.11	23.41

Table 6.4: Audio ratios of test sequences in percentage: E_L and E_R are ratios between audio energy of both front and rear channels and total audio energy. E_{SL} and E_{SR} are ratios between rear and side (both front and rear) channels for each side

Parameter Sets	Immersiveness Levels (ILs)		
	Low	Middle	High
Audio	No Audio	Stereo	Surround
Quality (QP)	36	20	20
Resolution	SD	HD	UHD

Table 6.5: Set of parameters for different ILs

- In the second test, all audiovisual stimuli were compressed to achieve the best quality regarding their corresponding device rendering capabilities. Audiovisual stimuli generated from the C3 test sequence were encoded with a QP of 25, whereas the remaining audiovisual stimuli were encoded with a QP of 20. A specific compression parameter was set for C3 as its demand in memory exceeded hand-held devices capacities. Expert viewing sessions were conducted to confirm the transparent quality of all audiovisual stimuli.

Regarding stimuli processing itself, sequences were compressed with an AVC/H.264 encoder (using x264 encoding in FFmpeg), and were scaled using bicubic filtering. No sound processing was required as stereo and surround signals were available. Nevertheless, in the first test, low presence level stimuli had no audio, due to unavailable audio mono signals.

An audio mono signal can be recreated from stereo or surround signals. However, one do not retrieve the same mono audio signal from stereo and surround information. To the best of our knowledge, no previous study investigated what the extent of differences between recovered signals is, or if these differences significantly impact audio perception. In this regard, we cannot guarantee the construct and external

validity of our test should mono audio be used for the lowest level of immersion.

As the research focused on SoP, ILs had to be discriminatory. Having highly different ILs was important. It is indeed sure that having no audio or stereo audio generates two different levels of immersiveness. Consequently, no audio signal was chosen for low IL stimuli.

Three ILs variants of nine video sequences make 27 video stimuli, presented to each subject during the experiment. Three audiovisual training stimuli were generated from C10, the training sequence.

6.2.2 Equipment and acquisition of physiological signals

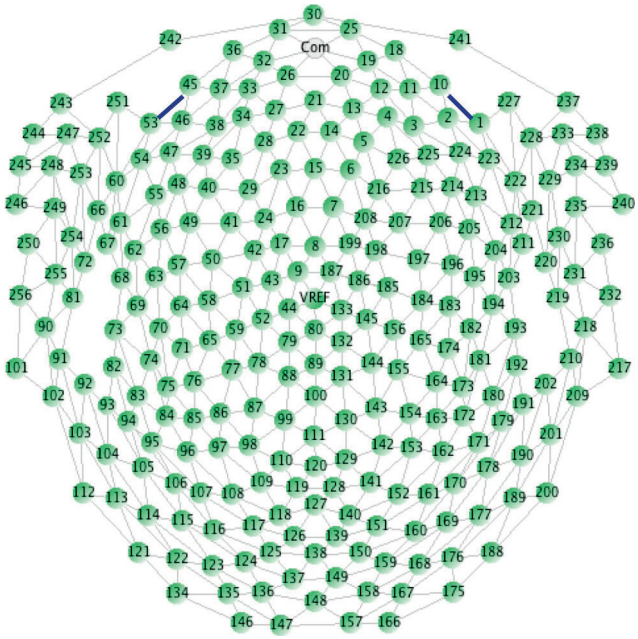


Figure 6.8: 256-channel GSN 200

The acquisition system EGI's Geodesic EEG System (GES) 300 of the company Electrical Geodesics [4] was used to record all physiological signals. EGI is the fourth company of EEG hardware worldwide. The benefit of using such equipment is the fact that it is easy to find GES 300-compliant supplies. Besides, GSE 300 is used in lots of laboratories around the world for physiological signals studies.

This equipment presents no risk of electrocution because the system of electrodes is galvanically isolated from the rest of the acquisition system and is supplied by power battery.

The particularity of this system is that it uses active electrodes. It means that a very low quantity of current is diffused at the surface of the electrode. According to the technical specifications, with this very low quantity of current, there is no

[4] http://www.egi.com

danger or risk for the subject's health. In the same manner, we are not aware of contraindications or risks when using this equipment.

GES is a complete package for working with biosignals. It supports the acquisition and processing of biosignals, particularly EEG. In addition to a Geodesic Sensor Net (GSN), the Net Station software, a data-acquisition computer, and an amplifier were available to us.

Dense-array GSN of 256 channels recorded the brain activity. This high-quality network records high spatial resolution signals when compared to usual EEG networks. This array of sensors tessellates the surface of the head as presented in figure 6.8.

The conduction between one's head and electrodes is made through a potassium chloride electrolyte solution. Specific attention must be paid to place the central electrode (REF/CZ) at the standard positions on the scalp. This particular spot is the pre-auricular midpoint and the nasion-inion midpoint. Besides, we had three sizes GSN to match the subject's head sizes. The precise way to use the GSN is exhaustively described in the EGI documentation of this system [5].

The amplifier filtered and measured EEG signals recorded by electrodes and sampled them at millisecond intervals. The frequency of acquisition was 250 Hz.

Net Station software collected, displayed and stored digitized EEG samples. This software specified how to filter data to the amplifier. In our case, a notch filter set to 50 Hz was used to remove the power line interference signal. Also, this software enabled the display and recording of electrodes impedances. It prevented to start any recording if more than 10 % of electrodes had an impedance above a certain threshold. This impedance limit has been set to 5 KOhms as recommended by EGI. The documentation to use the software describes how to configure our recording pipeline and to interact with stored data [6].

The entire GES system is really sensitive to any electric or magnetic fields. We made sure that any electronic devices not included in tests were turned off as well as verified our experiment equipment did not create interference with the EGI system.

In the GES, it is possible to add signals such as heart activity and respiration. As our aim is also to study those peripheral signals, we had additional equipment. Two standard electrodes, located on the lower left rib cage and the upper right clavicle of subjects, acquired the heart activity. This is the standard equipment and procedure for ECG recordings. Respiration signals were captured through thoracic and abdomen respiratory inductive plethysmography belts. Both pieces of equipment were GES-compliant and were wired to the amplifier. The acquisition system of these signals also had a temporal resolution of 250 Hz.

[5] https://f-origin.hypotheses.org/wp-content/blogs.dir/2484/files/2015/05/geodesic-sensor-net.pdf

[6] http://cb3.unl.edu/dbrainlab/wp-content/uploads/sites/2/2013/12/Acquisition_Manual.pdf

Figure 6.9: Fully equipped subject

Figure 6.9 shows a fully equipped subject. We can see the 256-electrodes network on his heads, the two plethysmography belts and wires of ECG electrodes.

The laboratory environment was calm and quiet. The ambient light setup guaranteed a satisfying level of comfort regarding the high or low brightness of the display, as well as during voting phases and breaks.

Figure 6.10 presents the environment setup during experiments. We can see that EEG, ECG, and belts were wired to a GSE 300 system and amplifier. Just as side note, the visible light on the picture was not visible from subjects position. This light sets the absolute indirect lighting of the test environment. The surround sound speakers positions, visible on the image, were compliant with the Rec. ITU-R BS.775 [ITU 2012a].

6.3 SoP prediction in respect with quality, resolution and sound system

This section introduces our first study performed with physiological signals. As stressed in section 6.1, there were several advantages to assess perceptual quality through physiological signals over explicit ratings. Also, it has been shown that it

Figure 6.10: Test setup during visualization of 4K stimuli

is possible to predict quality through the analysis of biosignals.

Our primary interest was to move forward towards the prediction of QoE. As already justified and explained, QoE is more than quality of contents. It results from a set of contents characteristics processed by a user's perceptual and cognitive functions.

Physiological signals were expressing the perception of a stimulus and contain brain, heart and respiration activities related to cognitive processes.

It thus remained to observe several levels of QoE through biosignals. However, it was at the time not possible to predict for sure what settings of contents characteristics will generate high or low QoE.

Instead, we investigated high-level aspects of QoE, the SoP. Presence is highly related to immersion (i.e., the extent to which technology systems are capable of delivering an illusion of reality to the senses of a human) that can be expressed as a function of content features. We assumed that most impacting contents variations of features were quality, resolution and sound system.

- By quality, we meant the inverse of compression level. Indeed, the more a content is compressed, the more artifacts and impairments are found in the content. If a high level of details improves the perceptual quality, the perception of artifacts cut it. Compression is an essential component of multimedia delivery and consumption, this factor had to be included as a content feature.

- Resolution represents the precision and refinement of visual information transmitted to the viewer. Recent and current development of pixel resolution legitimated our decision to include this content characteristic in our test. Also, going from SD to 4K is to increase the field of view covered by the displayed content. Consequently, resolution variations impact viewer immersion.

- Similarly, we wanted to include variations in sound systems that impact the

perception of contents. Channels added when going from no audio to surround sound recreate 3D sound. Immersion is thus also impacted by the variation of content sound system.

Variations of audio systems, quality, and resolution formed what is named ILs.

This study deals with the prediction of SoP through the analysis of physiological signals. The evaluated scenarios present variations of quality, resolution and sound systems. The experiment material and physiological sugnals equipment were previously presented in 6.2. In the following, test design, analyses and conclusions are presented.

6.3.1 Experimental design

Equipment: To render stimuli, the same display was used. Test stimuli were all visualized on a professional high-performance 4K LCD reference 56-inch monitor Sony Trimaster SRM-L560[7]. The viewing distance was set accordingly to Rec. ITU-R BT.1769 [ITU 2008b] at 1.6 H.

The Lansing 5.1 THX speaker system with super subwoofer, served as stereo and surround audio system.

Experiment methodology: The experiment implemented a SS with an ACR scale assessment methodology. Indeed, we believed that it is not possible to (sequentially or spatially) compare felt SoP.

Regarding the assessment, we decided to include several criteria.

- Of course, the first criteria was about presence. However, we had the feeling that subjects will understand better the concept of presence if we called it immersion. Indeed, both concepts are really close, such that even some researchers do not make the difference between both concepts. During the test, we gave to immersion the definition of presence.
- Presence is the psychological belief of being in an environment different from the physical one. We wondered if feeling presence cuts any awareness of our physical environment. We thus asked about surrounding awareness.
- Quality is the next criterion. Indeed, this is the only content characteristic that has not been selected for its immersion impact. It is thus an important dimension to evaluate. We only investigated video quality as a third of stimuli do not contain audio.
- We have seen in our test on HDR 360° the interest in content influences its perception. Besides, dislike of content may prevent people to feel presence during the media visualization. This fact is especially true for one-minute stimuli. Interest in audio and video content have been included in this study.

[7] *http://pro.sony.com/bbsccms/assets/files/cat/mondisp/brochures/di0195_srm1560.pdf*

Overall, self-assessment of subjects was designed to evaluate five criteria related to the audiovisual experience and SoP. The criteria were interest in video content, perceived video quality, interest in audio content, level of immersiveness (aka presence), and surrounding awareness.

Each criterion assessment followed a 9-point rating scale. Typically, 1 denoted the lowest, and 9 the highest value, corresponding to "low" and "high" for video quality, interest in video and audio content. We asked subjects to first consider this scale as a 3-point scale (low, medium and high) and then to refine their assessment. We will then obtain three classes of ILs, while having refined ratings. Regarding the level of immersion and awareness of the surrounding, "no immersion" and "full immersion", and "no awareness of my environment" and "full awareness of my environment" were the values of the two extremes, respectively.

Experiment protocol: Before the experiment, examples of each level of SoP were presented to subjects in a training session. The experiment was divided into three test sessions, interrupted by 10-minute breaks, to avoid tiredness and lack of attention of subjects.

The order of audiovisual stimuli in test sessions was set to low-middle-high, middle-low-high, and high-middle-low ILs, in this order. This pseudo-random stimuli order aimed at reducing the learning effect. It is likely that this effect appears when subjects are presented with successive stimuli having different resolutions.

Specific attention was paid not to repeat the same sequence during a session, to prevent subjects' boredom or fatigue. The combination of the nine test sequences with the ILs led to 27 audiovisual stimuli, forming 27 trials; nine of which were presented in each test session. Accordingly, subjects took part in three sessions.

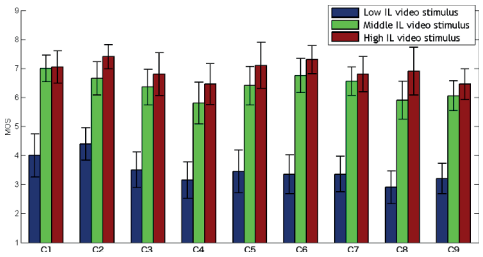
A trial comprised a ten-second baseline prior to a one-minute stimulus, followed by the rating of the stimulus.

The baseline period recorded signals of a resting period, during which subjects were instructed to breathe calmly and to concentrate on a 2D white cross on a black background, displayed on the screen for this purpose. This phase is important as unrelated variations in stimulus periods were removed based on baseline signals.

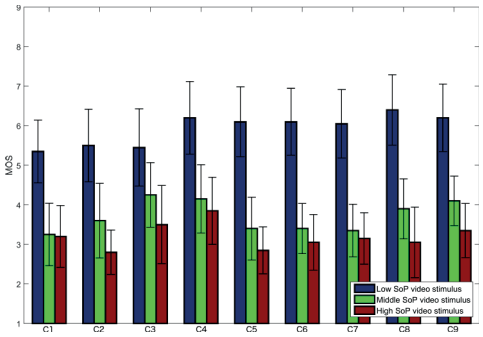
After the end of a stimulus, the observer provided his/her ratings, under no time constraints.

The end of the voting period indicated the end of the trial and triggered the start of the next trial. The experiment ended when the 27 video stimuli had been presented and graded. Despite an experiment duration of about two hours, subjects did not report fatigue.

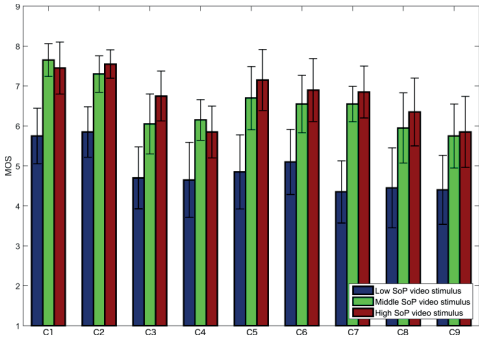
Participants: Twenty subjects, including eight females, participated in the study after being screened for correct visual acuity and color vision. They were 23 years old on average, with an std of 3.03 (age range from 18 to 30 years old).



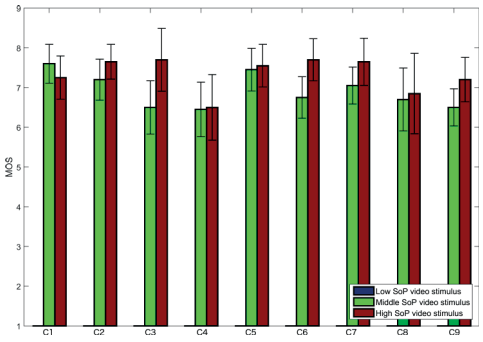
(a) MOSs and CIs for the experienced IL.



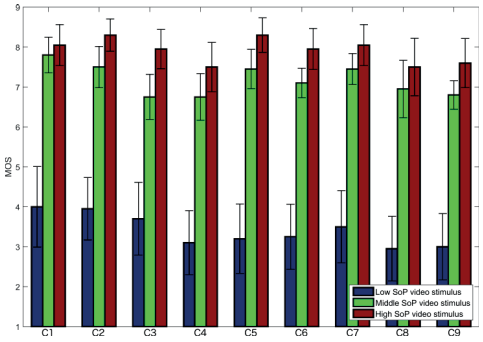
(b) MOSs and CIs for surrounding awareness.



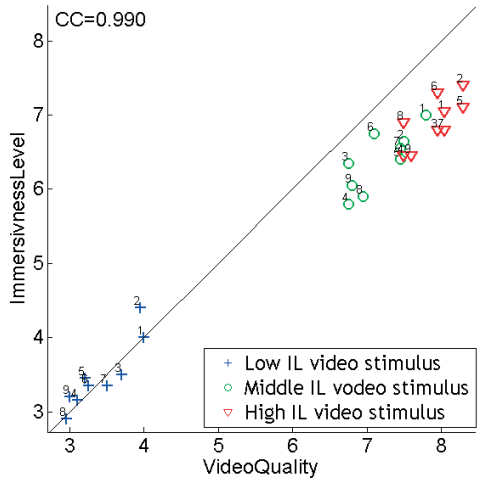
(c) MOSs and CIs for interest in video content.



(d) MOSs and CIs for interest in audio content.



(e) MOSs and CIs for video quality.



(f) Correlation between assessed SoP and video quality

Figure 6.11: Subjective ratings analysis.

6.3.2 Analysis of subjective scores

Before conducting the physiological signals analysis, we processed subjective ratings. Usual scores analyses have been performed such as MOSs and their associated 95% CIs, ANOVA and correlations between evaluated criteria.

Before any processing, we tested subjects reliability with an outliers analysis based on ILs. No outliers were detected and hence eliminated.

MOSs and CIs: Figure 6.11e presents MOSs and CIs of every stimuli for all criteria. We can observe that three ILs were experienced. Overall, the defined ILs correspond to subjective ratings, with an average MOS of 4, 6.5 and 7 to low, middle and high ILs, respectively. Considerations about CIs overlap confirm that middle and high ILs are similar and are different from the low IL. Therefore, the prediction of SoP based on this database is possible.

In more details, we attributed the clear distinction between low and other ILs to the fact that low IL stimuli had no audio. It is also possible that the huge difference in ILs when having audio and no audio impacted the perceived difference of IL between middle and high stimuli. As a result, even if we observed that middle and high ILs were different, CIs indicated that this difference was not significant.

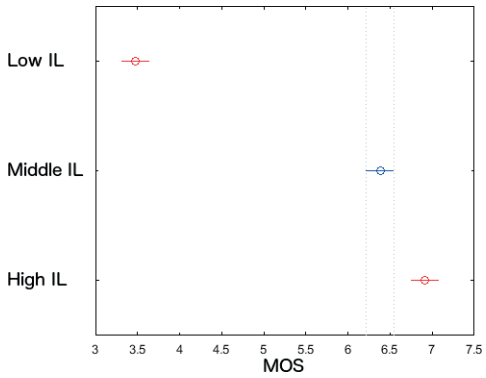
Regarding surrounding awareness results, we observed the opposite behavior of ILs MOSs. It seems that feeling presence during multimedia visualization decreased one's awareness of the physical environment. In a similar way than previously, only low IL stimuli were assessed significantly different from other ILs.

The observation of IL MOSs applies to interest in video content. No observation is possible for interest in audio content as there is no sound for low IL stimuli. Although, interest in both video and audio content C1 were higher for middle IL when compared to high IL.

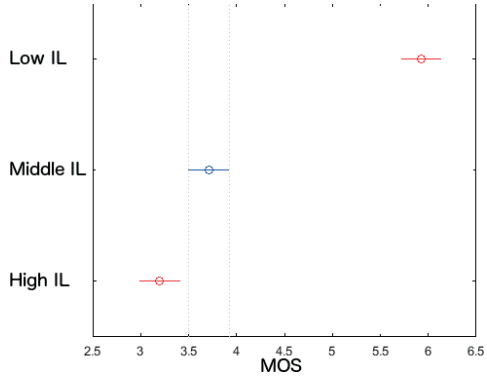
Criteria	p-values			Post-hoc test multicomparison
	ILs	Contents	Contents x ILs	
Immersiveness	0	0.0001	0.88	All ILs significantly different
Surrounding awareness	0	0.1617	0.64	All ILs significantly different
Video quality	0	0.001	0.96	All ILs significantly different
Interest in video	0	0	0.96	Low IL different from two others
Interest in audio content	0	0.005	0.04	All ILs significantly different

Table 6.6: One-way repeated measures ANOVA of assessed dimensions

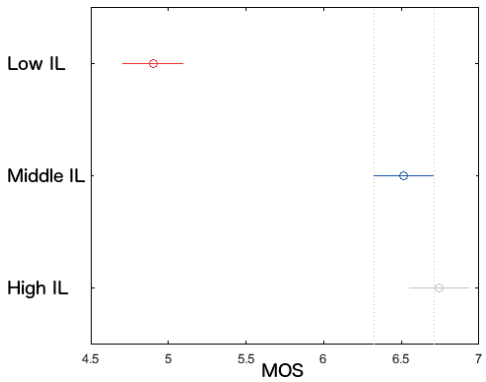
Statistical analysis: We have run two-way ANOVA and multicomparison posthoc tests in order to verify if ILs were significantly different from each other for all criteria. It turned out that it was the case, except for the interest in video content for medium and high ILs, as it can be read from Table 6.6 and Figure 6.12. However, when combining contents and ILs, no significant difference was observed, except for surrounding awareness.



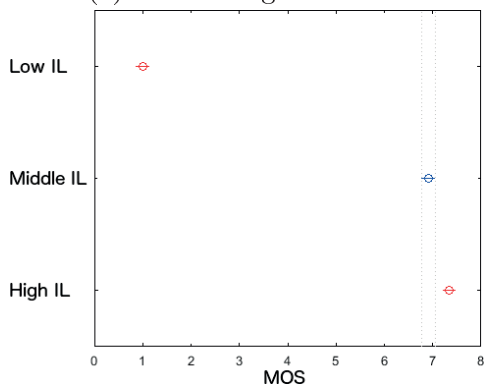
(a) IL.



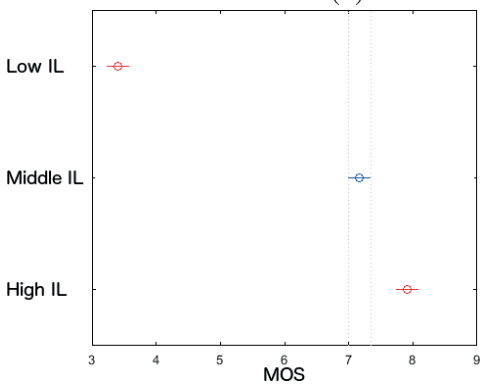
(b) Surrounding awareness.



(c) Interest in video content.



(d) Interest in audio content.



(e) Video quality.

Figure 6.12: Results of multicomparison post-hoc tests.

Even though middle and high ILs seemed really close (less than one scale unit of difference), the ANOVA and posthoc test showed their statistically significant difference. Accordingly, we can run the physiological signal analysis to predict three ILs.

Correlation: It was important to understand the interrelation between the five selected criteria. Besides, the assumption that the loss of physical environment awareness is directly due to the degree of immersion experienced was reviewed. To

	Video Quality	Video Content Interest	Audio Content Interest	Surrounding Awareness
Immersiveness Level	**0.990**	**0.914**	**0.974**	**-0.986**
Video Quality	-	0.892	0.988	-0.987
Video Content Interest	-	-	0.857	-0.903
Audio Content Interest	-	-	-	-0.965

Table 6.7: Pearson correlation coefficients between the ratings of different perceptual criteria

this end, the Pearson correlation was studied between all criteria of the experiment.

The correlation coefficients between criteria are detailed in Table 6.7 and Figure 6.11f illustrates the correlation between video quality and SoP.

In Figure 6.11f, numbers indicate the content from which stimuli originate. As already observed through MOSs analysis, there was a huge difference between the low IL class and the two other. We also saw the possibility to predict the SoP generated by experiences.

It must be stressed out that, for each content, an increase in defined IL resulted in a higher presence experience.

We observed in the table and figure the high correlation ($LPCC = 0.99$) between SoP and perceived quality. The assumption regarding surrounding awareness and immersivness has been verified by their strong inverse correlation ($LPCC = 0.99$).

Also, immersion and sound are highly related ($LPCC = 0.97$). This result was assumed to be due to our experiment design, particularly the choice to define low IL as visual stimuli only.

Overall, video quality and the inverse of surrounding awareness proved to be good candidates for SoP influence factors. Ultimately, these aspects may be included in QoE models.

6.3.3 Analysis of physiological signals

This section introduces the physiological processing from the pre-processing to the classification. Analyses of results are then described and commented.

6.3.3.1 Pre-processing, feature extraction and classification

ECG: As described in section 6.1.2.2, HR and HRV are the most common information extracted from ECG signals.

Presence is a time-varying effect, for instance, one does not feel in the content the first seconds of a movie. Based on presence specificities, we thought it was more appropriate to study time variations between heartbeats, meaning HRV.

We derived the HRV from ECG signals, from which typical features were extracted. Features include mean and variance, heart rate, low-frequency band power (0.03 - 0.12 Hz) and high-frequency band power (0.12 - 0.49Hz).

Respiration: Respiration signals do not necessarily require specific pre-processing. However, because the equipment used an amplifier during signals acquisition, a fair amount of noise is present in digitized signals due to the intensification of capture-related noise.

We decided to run a denoising operator, similar to PCA, getting rid of insignificant wavelet bases. Abdomen and thoracic respiratory signals were filtered using a wavelet multivariate de-noising [Aminghafari 2006].

We kept usual respiration features, knowingly respiration rate and average power across the 0.1 to 0.4 Hz frequency bands.

EEG: First of all, data were pre-processed following current practices in this field. A fourth-order Butterworth filtered the EEG signals between 3 and 47 Hz to reduce the EOG- and EMG-related artifacts.

From the total of 256 channels, we selected 19-electrode signals, based on the international 10-20 system configuration.

Electrodes impedance had been verified beforehand, meaning that they all were below the threshold of 5 KOhms.

An ICA was manually performed to remove eye-movement and blinking artifacts. By manually is meant that two experts selected which component(s) were to be removed as they were representative of artifacts such as eye artifact, muscle activity or noise component. The EEGLAB open-source Matlab toolbox was used for this purpose.

To have features representative of the dependencies between sub-regions of the brain, we explored the functional connectivity of signals [Fingelkurts 2005]. Functional connectivity seeks to find patterns of correlation.

A simple but widely accepted and used functionality technique is the Granger causality combined with small-world representations of electrodes network.

Thus, the Granger causality estimated the functional connectivity of the pre-processed EEG data. Conventional small-brain network features that have been extracted from the estimated functional connectivity maps are the characteristic path length [Watts 1998], global efficiency [Latora 2001], clustering coefficient [Watts 1998], and local efficiency [Latora 2001].

Features fusion: Two possibilities were offered to us for fusion. One approach is to combine features from various modalities (feature-based fusion); the second is to combine the outcomes of classifiers (decision-based fusion).

[Kroupi 2014a] showed that in several cases fusion between peripheral and EEG features was found to improve the predictor performance. The redundancies in signals, transmitted through extracted features, happened to improve classification accuracy.

Accordingly, a feature-based function was applied to the three physiological modalities. Pooling consisted in a concatenation of features in a single vector.

Classification: In the analysis of subjective scores we have shown that the perception of presence of the three ILs were significantly different. Thus, we can conduct a classification discriminating each of the ILs.

Supervised learning was possible thanks to subjective ratings, forming the ground truth to train the model.

The classifier was predicting the degree of SoP based on physiological signals hinged on a 3-class SVM.

Regarding the SVM kernel, [Maaoui 2008] supports the use of a linear kernel for emotion recognition. Also, polynomial kernels are used for emotion recognition when investigating galvanic skin responses [Liu 2016]. When investigating EEG signals, it is recommended to use polynomials or RBF kernels due to the high likelihood that the problem requires a non-linear model. In [Li 2014], multiple kernels learning support vector machine, implementing a mixture of polynomial and RBF kernels, was used for prediction.

Polynomial and RBF function have been tested as kernels. Highest performances of predictor were achieved with the RBF.

To validate our prediction model of SoP we split the features set into ten folds to perform a leave-one-out cross-validation.

6.3.3.2 Results and analysis

We report here the results of the ten folds leave-one-out cross-validation on the RBF-kernel SVM model learned on the fusion of EEG, ECG and respiration fusion.

Table 6.8 reports the confusion matrix of the predictor. It must be noted that random classification accuracy is 33% as the three IL classes were equally balanced (i.e., 180 instances each). "Actual IL" were ILs provided by subjects when scoring.

The classifier demonstrated its ability to correctly predict low and high ILs with an accuracy of 61% and 94% accuracy, respectively. However, it poorly predicts a medium sense of presence as an accuracy of 11% is reached for middle IL.

		Predicted IL			
		Low	Middle	High	Total
Actual IL	Low	**110**	0	70	180
	Middle	70	**20**	90	180
	High	11	0	**169**	180
	Total	191	20	329	540

Table 6.8: Confusion matrix of classification results between ILs. Numbers in the confusion matrix represent the resulting number of trials that are classified into each class.

Even though differences between middle and other ILs were significant statistically, the predictor was not able to correctly discriminate them. This finding

strengthened what we felt about defining classes of immersion: they must be highly different. In that sense the use of no audio for low IL was judicious.

The very high accuracy regarding the prediction of high IL, and the fact that no high IL trial was sorted as medium SoP is highly promising for multimedia fields.

6.3.4 Conclusion

When designing a product, service or creating brand new experiences, one needs to predict how users will perceive the experience of the developed entity. This explains the decades of research dedicated to the prediction of quality and QoS. This work position itself in making a step forward on the way to predict QoE.

Physiological signals are well known to contain implicit responses of an individual when presented with an experience. Biosignals have been proved to enable emotion recognition. Besides, more and more works relate physiological signals analyses to quality predictors.

These signals capture perceptual and cognitive processes responsible for judgment formation and decision making. Moreover, the implicit and time continuous aspects of such signals are valuable as no subjective self-assessment methodology efficiently implements implicit or continuous assessment. For these reasons, we figured out that such signals may be capital to predict QoE.

To predict QoE means having the ability to define to what extent certain stimuli will generate low, middle or high QoE. For now, we are not able to determine content characteristics that will for sure cause a specific level of QoE.

We thought about investigating a high-level aspect of QoE instead, namely the SoP. SoP is tightly linked to immersion, that can be defined as technical characteristics. Additionally, more and more emerging and new multimedia give priority to achieve high levels of immersion and presence (e.g., VR, 4K and volumetric sound). SoP is gaining momentum and attention from academicians and industrials.

Thus, to dig deeper into cognitive and perceptual processes, physiological signals were analyzed to verify if implicit body responses are indicators of SoP and ultimately QoE.

In this study, we investigated if the use of physiological signals helps to predict the SoP. The immersion dimensions that variated to create low, middle and high degrees of immersion were quality (aka inverse of compression rate), resolution and sound system. To the best of our knowledge, this is the first time that SoP is physiologically evaluated, particularly based on these variables.

Low IL stimuli presented no audio, a high compression rate and are rendered in SD. Middle IL corresponded to most current multimedia streams. That is, stimuli were in stereo, HD and compressed transparently. High IL were defined as surround sound, UHD stimuli compressed transparently.

We first created our content dataset as we observed that more broadcasting-like materials must be appraised in biosignal analyses. We extracted ten sequences from

four Blender open source movies. One of them was saved for training purposes. Sequences were selected based on audio rear channels energy, spatial and temporal complexity. We assumed that the more pieces of information are present in the three components, the more stimuli are likely to generate presence to subjects.

Sequences were set to last one minute to satisfy two constraints. The first one is the concern that stimuli must be long enough for creating presence effects. The second limit comes from subjective tests and restricted duration of sessions to prevent boredom and tiredness of subjects.

Stimuli were presented in a SS fashion and rated accordingly to a 9-point ACR scale. Five criteria were selected for evaluation, namely presence, surrounding awareness, video quality and interest in video and audio content. To increase subjects understanding of what they were asked to report, we referred to presence as immersion in questionnaires. Besides, the question on surrounding awareness has been added to verify if this measure is inversely proportionate to the level of presence. Interest in visual and audio cues may influence subjects liking of contents. We thus added questions about these two aspects.

Analyses of subjects' ratings demonstrated that subjects had experienced three statistically different ILs. We thus could proceed and learn a three-classes predictor based on physiological signals.

During the analysis, we observed that inverse surrounding awareness as well as video quality are influence factors that are to be considered for SoP evaluations.

Regarding physiological signals, EEG, ECG and respiration signals were recorded. Signals were pre-processed, and specific features were extracted. EEG features are based on functional connectivity, ECG on HRV and respiration on denoised signals. After the fusion of features, a classifier was learnt to predict the level of presence of stimuli. A SVM with a RBF kernel was used as learning model.

High, middle and low ILs were predicted with accuracies of 94, 11 and 61 % respectively, knowing that random classification was 33 %. This fact highlights that SoP and QoE may be investigated through physiological signals and can lead to accurate predictors for multimedia experiences. This result is excellent news for prediction of SoP and so ultimately for QoE.

Subjective scores, test material, and physiological signals form a dataset that is now publicly available [8]. This dataset is valuable for the research community as numerous study may be conducted on explicit rating and on physiological signals.

[8] http://mmspg.epfl.ch/SoPMD

6.4 SoP prediction for typical multimedia consumption scenario

If middle IL in the previous experiment was representative of multimedia viewers' consumption habits, the low IL wasn't. It is nowadays rare to watch videos without sound. In consequence, we focused on the second study on typical scenarios of multimedia consumption.

As highlighted in the previous analysis, sufficiently different experiences must be evaluated so that we feed the supervised learning model with data of three distinct classes of experiences. Again, we relied on differences of immersion, especially the coverage of viewers field of view, to design the ILs.

Nowadays, people consume multimedia in queue lines or buses when traveling from one point to the other. Hand-held devices, also referred to as mobile, are mostly used in such contexts. Then, when coming back home, people favor the consumption of media on desktop computers and TV, reliable and high-quality media platforms. Actually, in 2015, over young populations, time spent on media platforms is mostly shared between mobile and TV platforms [9]. Momentum is thus still improving concerning the use of mobile devices for multimedia consumption.

We believed that multimedia experiences on a mobile phone, a tablet or a 4K screen were dissimilar enough to form three ILs. They also are highly representative of today's and future media consumption habits. That is ILs were defined to match these devices best quality in terms of rendering capabilities. To mimic uncompromised transmission and to satisfy constraints of device memory, transparent compression was applied to stimuli.

Also, it would be very interesting to be able to discriminate experiences physiologically on different devices. One could observe what type of perceptual and cognitive processes were activated with specific devices. It is an implicit way to evaluates forces and weaknesses of a device and improving services accordingly.

This section describes our test design, presents the obtained outcomes and concludes about the prediction of SoP caused by typical scenarios of consumption, through physiological signals analysis.

6.4.1 Experiment design

We designed the test under the assumption that devices, used to render audiovisual stimuli, define low, middle and high ILs.

Test equipment: The iPhone5[10] and iPad4[11] were used to render stimuli corresponding to low and middle ILs. The luminosity of the iPhone and iPad was set to 75% of their maximum brightness. This setting pledge for visual comfort.

[9] https://www.themediabriefing.com/analysis/nine-trends-in-us-media-consumption-in-charts/
[10] *https* : *//support.apple.com/kb/SP*655?
[11] *https* : *//support.apple.com/kb/SP*662?

The professional Sony Trimaster SRM-L560 monitor, introduced in Section 6.3.1 was used to render high IL stimuli. The Table 6.9 illustrates characteristics of the three devices.

As recommended in [ITU 2012b], the DVD of the UHD TV was set to 1.6H. To prevent additional noise in recorded EEG signals and to reduce the influence of movements of the portable devices, subjects did not hold the iPhone or the iPad. Instead, they were fixed in front of them at a PVD of about 6H (30cm) and 4H (60cm) respectively. H corresponds to the height of the display area of each device.

The estimated foveal area of the video stimuli on each device was therefore around 15%, 10% and 3% of total video stimuli, on iPhone, iPad, and UHD TV, respectively. We verified here the coverage of the field of view achieved by ILs.

To match realistic scenarios the test design comprehended two sound systems. Stereo audio signals were used for iPad and iPhone, and 5.1 surround audio signals were used for UHD TV. The stereo sound was provided by a professional headset Sennheiser HD 280 Pro[12] (accurate for linear sound reproduction in critical monitoring applications and attenuating up to 32 dB of ambient noise). The Altec Lansing 5.1 THX speaker system super subwoofer was used as 5.1 surround sound system.

Parameters Setting	ILs		
	Low	Middle	High
Device	iPhone5	iPad4	UHD TV
Audio	Stereo	Stereo	5.1 Surround
Native Device Resolution	1136x640	2048x1536	3840x2160
Video Resolution	1280x720	1920x1080	3840x2160
Foveal area[px]	171	202	121
Foveal area[%]	15	10	3
Viewing distance	6H	4H	1.6H
Viewing distance[cm]	30	60	110

Table 6.9: Immersiveness Level (IL) settings based on characteristics of devices

Experiment Methodology: The experiment included three sessions. During each session, nine stimuli were visualized on the same device. This process has been selected for the convenience of running the test, but more importantly to reduce the learning bias introduced by the change of device.

A minimum of one day separated every two sessions to avoid any statistical bias between the evaluations of two different devices and tiredness of subjects. Each session lasted approximately one hour, including the training phase and setup of devices for the acquisition of physiological signals.

[12] $http://en-us.sennheiser.com/professional-dj-headphones-noise-cancelling-hd-280-pro$

A training session was inserted at the beginning of each test session to remind subjects about the test procedure and show on stimuli of each IL. After the training, subjects were asked to calm down to record the calibration trials described in Section 6.1.2.1.

In the same way that in the previous experiment, a session presented nine stimuli which led to a total of 27 audiovisual stimuli forming 27 trials per subject. Each trial consisted of a ten-second baseline phase, a stimulus period and a voting phase. After the audiovisual stimulus was over, subjects were asked to provide their ratings in 30 seconds (about 5 seconds per question). The next trial followed the voting phase until completion of the entire session. Figure 6.13 illustrates an example of one test trial including baseline, audiovisual stimuli, and voting. The order of trials was randomized in every session for each subject.

Regarding ratings, subjects evaluated audiovisual stimuli according to six different criteria, namely, Interest in Audio content (IA), perceived Audio Quality (AQ), Interest in Video content (IV), perceived Overall Quality (OQ), IL, and Surrounding Awareness (SA). As audio information highly impacted our previous test, we added a question on audio quality. Also, video quality has been changed to overall quality to assess both perceptual streams simultaneously.

The evaluation methodology is an SS with 9-point ACR scale for each criterion. The extremes 1 and 9 represent the lowest and highest values, respectively. In particular, the correspond to "low" and "high" for IV, IA and OQ, "no immersion" and "full immersion" for IL, and "not aware" and "fully aware" for SA.

The 9-point rating scale was presented with clear separation lines between low (1-3), middle (4-6), and high (7-9) ratings creating three classes of grades. Subjects were instructed to evaluate the stimuli in the 3-class ratings (low, middle, and high) and then further refine their assessment.

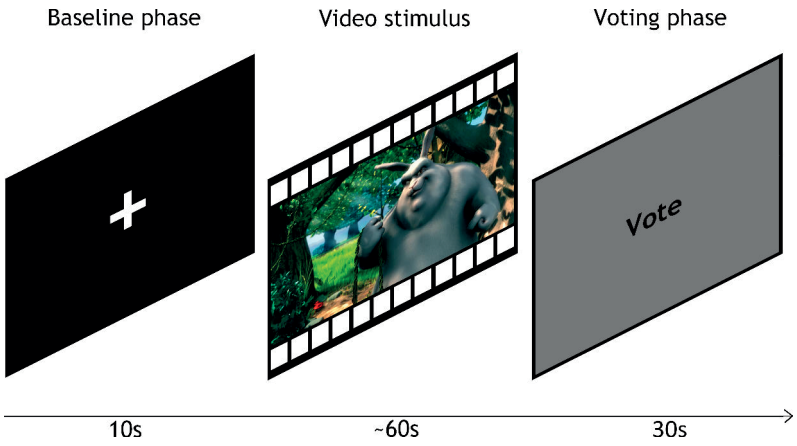


Figure 6.13: Example of a trial

Participants: Twenty subjects participated in this study (ten females and ten males). They were from 18 to 25 years old (21 on average with a 2.17 standard

deviation). Participants were screened for correct visual acuity and color vision.

Written and oral instructions were given to participants before they signed a consent form.

6.4.2 Analysis of subjective scores

This section describes the analysis carried out on explicit ratings to explore subjects' perceptual responses to different ILs and to investigate the interrelation between the various evaluation criteria.

First, we confirmed Student's t-distributions of the individual rates.

Detection of outliers was performed according to guidelines described in Section 2.3.1 of Annex 2 of [ITU 2012c]. Based on ILs ratings, no outliers were detected. Hence, the non-significant deviation across subjects ratings is ensured.

MOSs and 95 % CIs: Figure 6.14 depicts MOS and CI values for each content and device, corresponding to evaluated criteria.

Regarding ratings, for all criteria but SA, iPhone generally induced the lowest level, whereas UHD TV caused the highest level.

Middle and low ILs appeared to be quite similar. For contents $C3$ and $C9$, low presence MOS was even above that of the middle. $C3$ and $C9$ test sequences can be compared to the test sequence $C4$, which originated from the same movie Elephant Dream and exhibited very similar audiovisual characteristics (SI, TI, foveal area, aspect ratio, and audio energy). Therefore, the observed behavior of $C3$ and $C9$ test sequences is most likely due to statistical differences between subjective ratings.

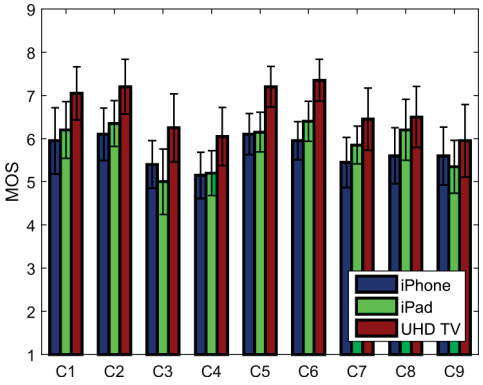
Concurrently, we observed a difference between the two lowest and highest ILs. However, this difference variates from content to content. We noticed that the sequences which video contents were assessed as interesting reached higher presence levels (e.g., for $C1$, $C2$, $C5$ and $C6$). Correlation between both aspects must be verified to include interest in video in future experiments on SoP.

Statistical analysis must confirm it, but it seems that subjects have experienced only two levels of IL. This scenario is particularly likely due to the small range of MOSs, that spans between about 5 to 7,5, on a 9-point scale. Actually, presented audiovisual stimuli, in the laboratory environment, did not induce low IL. Moreover, audio and overall quality results supported the view of having two levels of experience in this experiment.

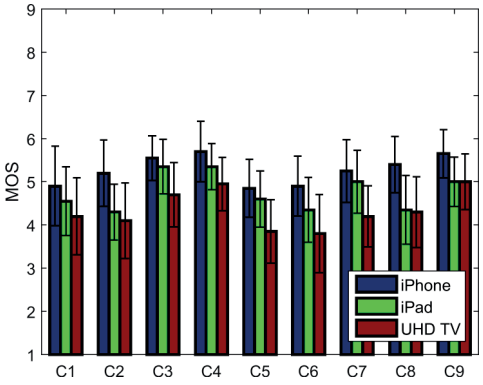
With respect to SA, results were opposed to those of presence levels, as expected. Interestingly, anomalies observed for $C3$ and $C9$ test sequences were not present. This fact supports our belief that these discrepancies were statistical.

Interestingly, middle and low presence levels were less similar for this dimension than for IL.

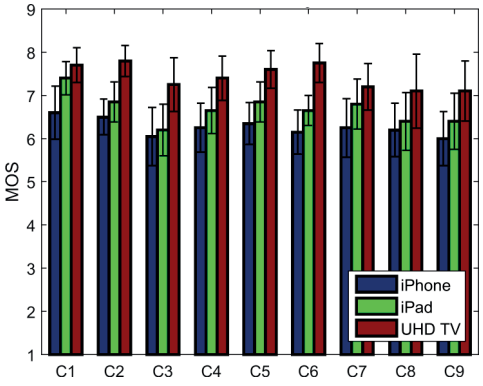
Last but not least, it seems that interest in audio and video content did not variate depending on ILs. However, in results about interest in video contents we



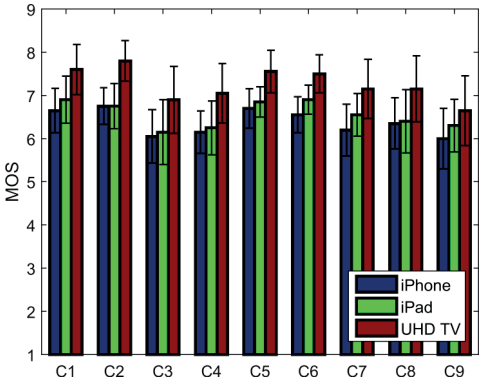
(a) MOSs and CIs for the experienced presence.



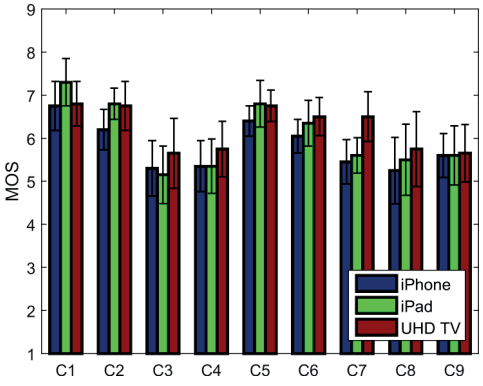
(b) MOSs and CIs for surrounding awareness.



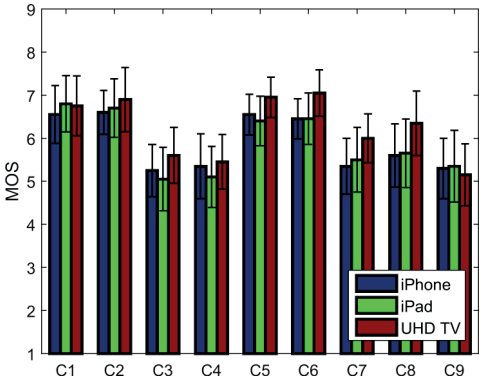
(c) MOSs and CIs for audio quality.



(d) MOSs and CIs for overall quality



(e) MOSs and CIs for interest in audio content.



(f) MOSs and CIs for interest in video content.

Figure 6.14: Subjective ratings analysis.

readily observed that subjects had a preference for test sequences extracted from Big Buck Bunny and Sintel movies ($C1$, $C2$, $C5$ and $C6$). As already mentioned, this preference impacted the felt presence.

Statistical analysis: We have run two-way ANOVA and multicomparison posthoc tests to verify if ILs were significantly different from each other for all criteria. Table 6.10 and Figure 6.15 present obtained results.

We verified that the interest in content is not related to the platform on which it is visualized ($p > 0.01$). However, we also witnessed that there is no statistical difference of interest across sequences. With regards to interest in audio, shown to present at least a statistical difference, $C3$ and $C5$ were the most apart but not sufficiently to be from different distributions. Thus, our previous comment on the preference of contents does not stand.

The statistical analysis confirmed our intuition that only two statistically different levels had been experienced. Indeed, if high presence was different from both other levels, low and middle levels were not proved to be dissimilar. The physiological analysis dealt thus about a binary classification of SoP.

Surprisingly, three ILs were experienced regarding sound quality. We expected a difference between lowest and high level due to stereo and surround sound systems used for corresponding stimuli. However, we have no explanation for the difference between low and middle levels. Based on Apple technical specifications both devices provide similar quality sound [13],[14].

Criteria	p-values			Post-hoc test multicomparison
	ILs	Contents	Contents x ILs	
Immersiveness	0	0.14	0.81	No difference between low and middle ILs
Surrounding awareness	0	0.12	0.49	No difference between low and middle ILs
Audio quality	0	0.39	0.29	All ILs significantly different
Overall quality	0	0.24	0.09	No difference between low and middle ILs
Interest in video	0.07	0.63	0.78	No difference between ILs
Interest in audio content	0.03	0.04	0.57	Only low and high ILs are different

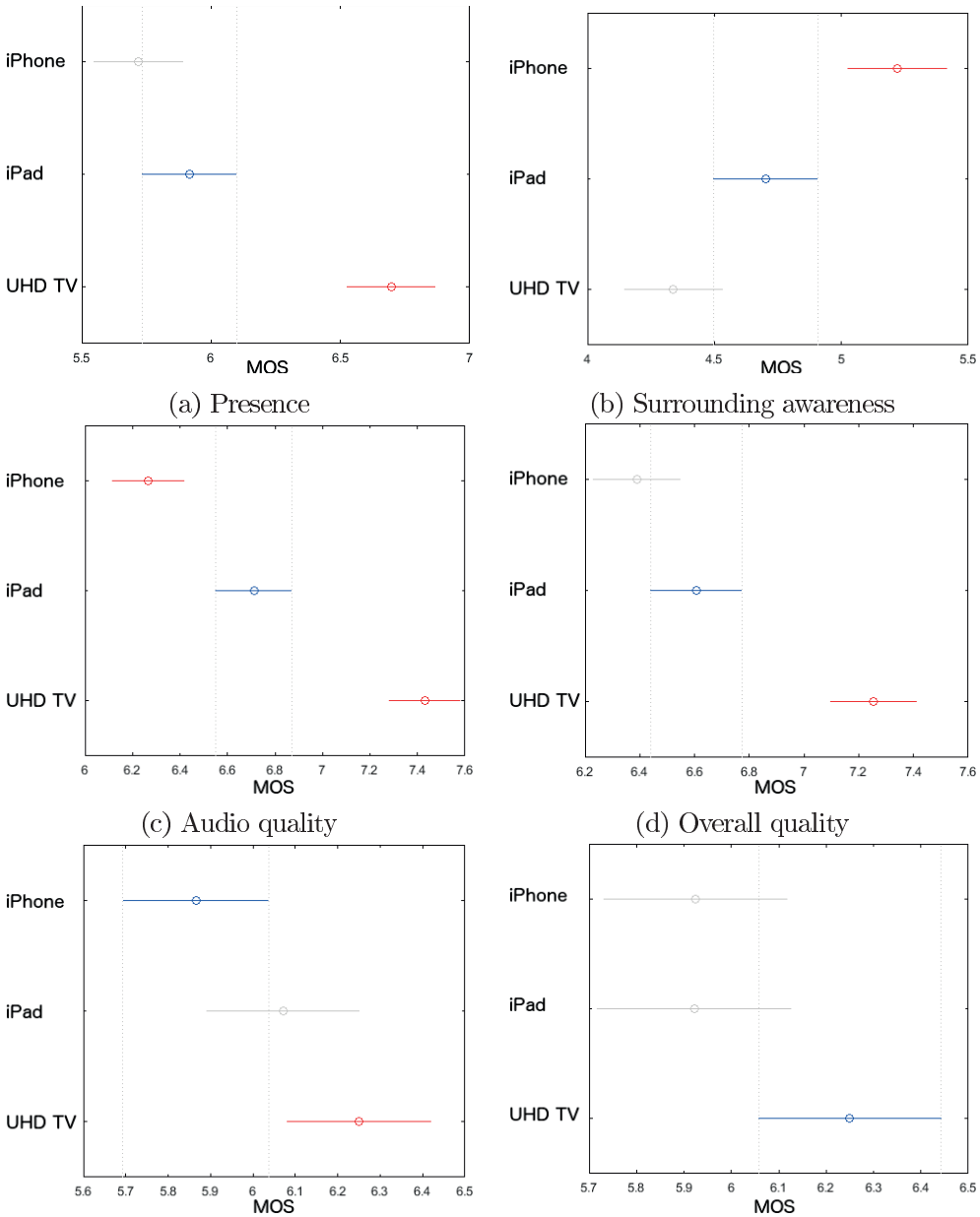
Table 6.10: One-way repeated measures ANOVA of assessed dimensions

Correlation: To understand mutual relations between evaluated criteria, Pearson's correlation was applied between MOS values of each pair of criteria. Table 6.11 summarizes the correlation results obtained.

A high correlation between SoP and OQ has already been observed in the previous study. Results in this test showed a similarly high correlation ($LPCC > 0.938$). It confirmed that the overall quality of audiovisual stimuli has an essential impact on immersive experiences.

Similarly, results showed a strong negative correlation between SA and IL ($LPCC < -0.913$). SA is thus a good indicator of SoP.

[13] https://support.apple.com/kb/SP655?locale=en_GB
[14] https://support.apple.com/kb/SP662?locale=en_GB



(a) Presence (b) Surrounding awareness

(c) Audio quality (d) Overall quality

(e) Interest in audio content (f) Interest in video content

Figure 6.15: Results of multicomparison posthoc tests

Moreover, OQ and AQ were strongly correlated ($LPCC = 0.914$) which hinted that high-quality audio plays an important role in the overall quality experienced by subjects.

There was also a good correlation between IL and both IV and AQ ($LPCC > 0.819$). If audio quality was expected to impact ILs, the extent of relatedness between ILs and IV is impressive. We thus highly recommend the inclusion of IV in future evaluations of SoP (and so for QoE), and strongly support the idea to evaluate sound quality in audiovisual experiments.

The inter-relation between IL and IA is weaker ($LPCC < 0.723$) therefore has less impact on SoP than previous criteria.

Figure 6.16 shows the correlations between IL and other evaluation criteria for each device and content in more detail. The color and shape of each marker indicate which device and test sequence were used for multimedia consumption, respectively.

In correlation graphs, the distinction between the three ILs is not always straightforward. Referring to results illustrated in Fig. 6.16a, UHD TV audiovisual stimuli form a quite separate cluster, whereas iPhone and iPad stimuli are more interleaved. It can be observed that the IL of each stimulus is often improved when an iPad is used instead of an iPhone, and when a UHDTV is used instead of an iPad or an iPhone.

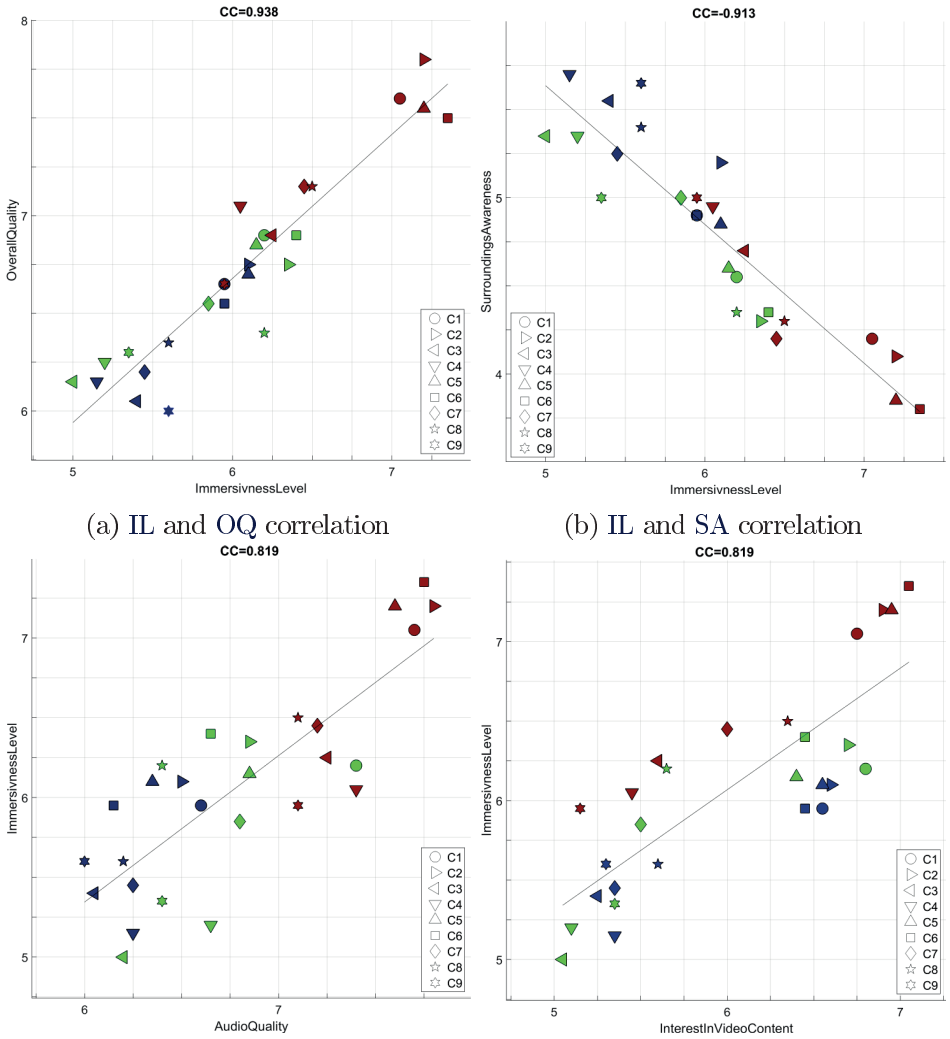
	Surrounding Awareness	Overall Quality	Video Content Interest	Audio Content Interest	Audio Quality
Immersiveness Level	**-0.913**	**0.938**	**0.819**	**0.723**	**0.819**
Surrounding awareness	-	-0.869	-0.756	-0.705	-0.781
Overall quality	-	-	0.764	0.712	0.914
Video Content Interest	-	-	-	0.889	0.557
Audio Content Interest	-	-	-	-	0.609

Table 6.11: Pearson's correlation $LPCC$ coefficients between the different evaluation criteria ratings

Classes for prediction: With regards to our previous conclusions, we performed a binary classification. Accordingly, we needed to select which classes were to be classified.

Table 6.12 shows the distribution of stimuli experiences in ILs. A differentiation is made base on defined ILs (i.e., stimuli watched on iPhone, iPad and UHD TV) and assessed ones (i.e., IL of ratings from 1 to 3, 4 to 6 and 7 to 9 are low, middle and high, respectively).

One can see the defined levels of presence were balanced, while this was not the case for assessed ones. Thus, we decided to set two phases in our classification learning. The first one is dedicated to the identification of defined ILs, meaning which device has been used for rendering. The second is related to the prediction



(a) IL and OQ correlation

(b) IL and SA correlation

(c) IL and AQ correlation

(d) IL and IV correlation

Figure 6.16: Correlation between the experienced IL and (a) perceived Overall Quality (OQ), (b) Surrounding Awareness (SA), (c) perceived Audio Quality (AQ) and (d) Interest in Video content (IV). Blue, green and red markers are iPhone, iPad and UHD TV stimuli, respectively

of SoP. Regarding the levels of SoP to predict, middle and high classes present a similar and sufficient number of trials for the learning process.

In both cases, the classifier was designed to make a binary decision.

6.4.3 Analysis of physiological signals

This section introduces the physiological processing from the pre-processing to the classification. Analyses of results are then described and commented.

	Number of trials		
	Low	Middle	High
Defined ILs	180	180	180
Assessed ILs	29	288	223

Table 6.12: Number of trials for defined and assessed ILs

6.4.3.1 Pre-processing, feature extraction and classification

ECG: For the same reasons as in the previous study we computed the HRV of ECG and same features.

As LF/HF ratio describes the sympathetic versus parasympathetic balance. This balance may be highly important to measure presence as it relates to calm or excitement. The mental state may indicate the experience of presence. We added this feature to our analysis.

Hence, the five features extracted were HRV mean and standard deviation, low and high HRV spectral powers and their ratio.

Respiration: In this test, we chose to denoise signals with the PCA. This decision was based on the representativeness of PCA components when compared to wavelet bases. The first PCA component was retained as it described respiration variations accurately.

For feature extraction, we wanted in this study to extract features similar to that of ECG. The five features extracted from respiration signals were the retained component mean and std, low and high variability spectral powers and their ratio.

EEG:

Pre-processing: Any electrode with impedance higher than 5 KOhms was rejected.

The remaining electrode signals were filtered with an eighth-order Butterworth filter with cutoff frequencies between 0.05 Hz and 30 Hz.

The pre-processed signals were then decomposed into five standard brainwaves (theta, alpha, and low, medium and high beta waves). Each feature extraction is applied on the five sub-bands signals.

Features extraction: In this study, we wanted to build our model based on a more learning-based methodology. After a review of the state of the art, a supervised dimension reduction method was selected.

The Common Spatial Pattern (CSP) is broadly used for features extraction of physiological signals when creating BCIs [Yang 2013, Zhang 2010], estimating emotions [Li 2009b] and predicting image quality [Acqualagna 2015]. This method

efficiently identifies spatial components maximizing the energy difference between two classes.

Given a signal trial E (N x T), with N the number of channels recorded and T the amount of samples.

One obtains the covariance matrix of the EEG from Equation 6.6.

$$C = \frac{EE^T}{trace(EE^T)} \tag{6.6}$$

The composite covariance matrix C_C is the sum of the covariance matrix of each class. It can be decomposed thanks to U and D, which denote the set of eigenvectors and the diagonal matrix of eigenvalues of C_C.

$$C_C = C_1 + C_2 = UDU^T \tag{6.7}$$

Signals in each class are filtered by the whitening transformation $P = D^{-\frac{1}{2}}U^T$. P decorrelates signals while normalizing their variance.

$$\hat{C}_1 = PC_1P^T \text{ and } \hat{C}_2 = PC_2P^T \tag{6.8}$$

A second eigen-decomposition is applied to filtered signals, leading to the orthogonal matrix V, the set of eigenvectors, and the diagonal matrix of eigenvalues λ.

$$\hat{C}_1 = V\lambda V^T \text{ and } \hat{C}_2 = V(1-\lambda)V^T \tag{6.9}$$

This decomposition maximizes the differentiation between the two classes signals.

From P and V is formed the CSP projection matrix W.

$$W = P^TV \tag{6.10}$$

First and last column vectors in W are the best spatial components discriminating both classes (provide maximum variance of classes). Those two vectors thus become our extracted features for EEG signals. Hence, we obtained ten features from EEG signals (two CSP features per frequency bands).

Electrodes system: The CSP is demanding in computing power, and its complexity is directly dependent on signals number of channels and samples. The number of samples is not varying as it is the entire set of recordings during a stimulus presentation. Hence, we investigated settings regarding the number of channels of EEG signals.

Considering EEG signals, due to over-fitting constraints, we thought about testing 20, 32, and 64 electrode systems.

Results motivated the selection of the 20-electrode system. We suspect that including too much information in CSP prevents the filter from finding the best discriminative spatial components.

Frequency bands study: Frequency bands are linked to mental activity, and it is most likely that beta bands are the most relevant frequency bands to analyze. Indeed, beta bands are responsible for focused attention and activation. These aspects seemed related to engagement and so to SoP.

We checked here whether considering a part of frequencies may lead to better prediction accuracy.

Our results showed that including all brainwaves or combining only theta and alpha waves leads to similar classification accuracy.

This finding provided valuable insights into SoP: it is linked to relaxation and information encoding.

EOG denoising: It is usually recommended to pre-process signals and remove artifacts related to blinking and eye movements. We compared the classification accuracies of denoised and not denoised signals.

Performed empirical tests showed that higher classification rates are achieved on EEG raw signals when compared to pre-processed signals. It could indicate that blinking artifacts are a SoP indicator. Indeed, it has been shown to be related to visual fatigue, possibly preventing subjects to experience presence.

We thus did not perform blinking-artifact removal in the following.

Delay: ERP studies are irrelevant for SoP analyses as no presence is experienced at the very beginning of a stimulus.

We went a step further regarding the fact that presence effect beginning is not simultaneous to stimulus start. We examined if the introduction of a delay can improve performance. By delay is meant that the first seconds of the trial signals are not taken into consideration.

We considered removing the 5, 10, 15 and 20 first seconds of every physiological signal. That represents a suppression of 8, 17, 25 and 33 % of signals sample, respectively.

No significant improvements were observed in the classification results. As our results were inconclusive, we removed any delay parameter in the remaining processing.

Fusion: A feature-based fusion has been performed. When combining the three physiological modalities, ECG and respiration signals features were concatenated with EEG features, leading to a 20-feature array.

Classification: The classification operator used is a SVM. Similarly to the first study, a RBF kernel has been used.

The classification efficiency was appraised by applying 20-fold cross-validation, on a leave-one-out cross-validation based on subjects. This process turned our analysis into a subject-independent classification.

6.4.3.2 Results

		Actual class		
		IPhone	TV	Total
Predicted class	IPhone	159	14	173
	TV	21	166	187
	Total	180	180	360

Sensitivity	0.84 ± 0.17
Specificity	0.92 ± 0.22
Overall accuracy	0.90 ± 0.17

		Actual class		
		IPad	TV	Total
Predicted class	IPad	134	65	199
	TV	46	115	161
	Total	180	180	360

Sensitivity	0.74 ± 0.39
Specificity	0.64 ± 0.37
Overall accuracy	0.69 ± 0.25

		Actual class		
		IPhone	IPad	Total
Predicted class	IPhone	123	96	219
	IPad	57	84	141
	Total	180	180	360

Sensitivity	0.68 ± 0.38
Specificity	0.47 ± 0.34
Overall accuracy	0.57 ± 0.25

Table 6.13: Rendering system identification confusion matrices

The cross-validation efficiency was analyzed by reporting averaged confusion matrices. Along with classification results, we indicated the sensitivity and the specificity, representing first and second class prediction accuracies, respectively, and the overall model accuracy. Confidence intervals were deduced from the 20-fold cross-validation. Our processing was applied to presentation device identification, followed by SoP prediction.

Devices identification: Results in Table 6.13 report a high accuracy in classification (90%) when discriminating iPhone and 4K monitor. This is in line with subjective ratings analysis, in which rating distribution centers of these classes are the most apart.

Surprisingly, the hand-held presentation device classifier achieved a fair efficiency (57%), when considering that the two classes have similar grades distributions and that an arbitrary classification would be 50%.

When classifying iPad and 4K stimuli, the classifier achieved an overall accuracy of about 70%, which corroborates with previous findings, the possibility to identify presentation devices using the proposed method.

SoP prediction: Table 6.14 summarizes the efficiency of the classifiers regarding SoP prediction.

Results discriminating low SoP levels are not reported. The unbalanced number of trials in SoP classes led to the classification of all low SoP trials in the second class due to over-fitting.

Regarding the distribution of trials in medium and high ILs, a random classifier would reach an overall accuracy of 56%. Therefore, the 61% classifier overall accuracy is not as promising as presentation device identification results indicated.

We can note the high classifier sensitivity and low specificity, probably due to the higher number of medium trials. This indicates that most trials were predicted as medium SoP levels.

This result may be interpreted in several ways.

1. Physiological responses of middle and high SoP stimuli are not different enough for performing accurate discrimination.

2. The second one questions the efficiency of CSP features for SoP prediction. The spatial characteristics may not be relevant to describe SoP.

Subjective ratings and ANOVA analyses advocate against the first point. Moreover, in [Abromavičius 2017], spectral features of physiological signals were successfully put in correspondence with the ground truth scores. This strongly advocates in favor of the second inference.

As the identification of the presentation device reached high accuracy, while SoP couldn't be predicted, we can conclude that this study design of ILs was correct for immersion but not for SoP.

6.4.4 Conclusion

This study implemented the investigation of physiological responses to typical multimedia consumption scenarios. Here again, we assumed that sufficiently different ILs will enable the observation of various SoP levels.

In the previous experiment, most representative functions of immersion were evaluated, namely resolution, compression rate and sound system. However, some

		Actual class		
		Medium SoP	High SoP	Total
Predicted class	Medium SoP	241	152	393
	High SoP	47	71	118
	Total	288	223	511
	Sensitivity		0.84 ± 0.17	
	Specificity		0.32 ± 0.22	
	Overall accuracy		0.61 ± 0.13	

Table 6.14: SoP prediction confusion matrix

scenarios did not suit the habits of today's users. This fact was tackled when designing the levels of immersion of stimuli.

ILs have been defined as best experiences that iPhone, iPad and UHD TV realize. In terms of details in the image, the foveal area perceives 15, 10 and 3 % of low, middle and high IL stimuli, respectively. Hence, a 4K covers more of the field of view than iPad, which in turn covers more of viewers sight than iPhone. For these reasons, these levels were assumed sufficiently different to conduct an experiment.

Same one-minute test sequences have been used in this test than that of the previous study. A perfectly balanced population of 20 individuals evaluated the generated stimuli.

Stimuli were presented in a SS fashion and rated accordingly to a 9-point ACR scale. Six criteria were selected for evaluation, namely presence, surrounding awareness, audio, and overall quality and interest in video and audio content. As observed in our previous study, the sound may be of importance when evaluating audiovisual sequences; we thus added a question on audio quality. We also changed video quality to overall quality to encompass both perceptual cues in this assessment.

MOSs showed that high, medium and low defined ILs performed from most to least, as expected. However, low and medium felt levels of presence were not significantly different. Thus, the classification had to be binary. When considering the distribution of scores, medium and high levels of SoP were to be discriminated.

We verified that SA is strongly inversely correlated to IL. Combined with other correlation results, we support the use of SA, overall quality as influence factors of SoP, and ultimately of QoE.

Our results hinted that interest in video favors the experience of presence. However, despite the high correlation between the two aspects, statistical analysis did not validate this result. We believe that the relation between these concepts must be further explored in future works for better evaluation of QoE.

Regarding physiological signals, EEG, ECG and respiration signals were recorded. Signals were pre-processed, and specific features were extracted. ECG and respiration features are frequency characteristics extracted from the HRV and

denoised signal.

Regarding EEG features, we got inspired from BCI and emotion recognition research. We proposed, implemented and tested a classifier based on CSP filter and RBF-kernel SVM, used for the first time in QoE evaluation. Instead of studying functional connectivity or conducting a spectral analysis, it relates signals to spatial patterns.

We defined the same classifier as in our previous study, meaning a SVM supervised learning technique with a RBF kernel. Leave-one-out cross-validation estimated the accuracy of prediction and turned our analysis into a subject-independent classification.

We investigated several settings for EEG features extractions to improve classification performance.

- A study on how many channels of EEG signals are to be processed resulted in having the 20-electrode system outperforming 32 and 64 systems. It seemed that CSP had hard time decorrelating signals should they have too many entries.

- As SoP is not an event related effect (e.g., it does not begin when a stimulus starts). We studied if removing first seconds of trials would improve classification results. Trials with 5, 10, 15, and 20 seconds less were envisioned.

  These results were inconclusive as no classifier with introduced delays achieved better prediction accuracies.

- EOG artifacts are known to be related to visual fatigue. This effects may potentially prevents subjects from experiencing SoP. We thus tried EEG features extracted from noisy or denoised signals. Classifier modeled on noisy-signals-based features performed better.

  Hence, no EOG removal have been performed. Also, this finding suggested that blinking and eye movements could be indicators of SoP.

- Frequency bands are related to mental states and level of activity. We thus performed classification while considering any combination of frequency sub-bands. It turned out that to combine theta and alpha features achieved same accuracies of prediction than that of all bands features coupled.

  This results indicated that SoP is linked to relaxation and information encoding.

Classification has been done in two phases.

1. On one hand balanced defined ILs trials were used to verify if devices can be identified through physiological signals analysis. It also gave us the opportunity to ascertain the relevance of using CSP to extract EEG features.

   The developed rendering device identifier successfully reached up to 90% accuracy with high sensitivity and specificity for discrimination between TV and iPhone.

This is an interesting finding for broad applications in the future such as retargeting and user-based delivery services.

2. On the other hand, we aimed to predict SoP, that is assessed ILs. We recall that due to the imbalance in classes, only middle and high levels were discriminated.

   The CSP-based RBF-SVM-classifier implemented here realized a classification accuracy of 61%, which is not satisfactory considering the random classification accuracy of 56%.

   This may indicate a difference in presence and immersion constructs which is interesting for future test designs exploring the SoP. However, further analyses strongly advocate for the irrelevance of CSP features for the study of SoP.

6.5 Data sets

For the two studies described above, physiological signals datasets were created. Datasets are about 40 and 50 GB each.

Each dataset includes collected subjective scores, recorded digitized physiological signals and stimuli.

These datasets are publicly available on the MMSPG website[15].

6.6 Conclusion

Subjective self-assessment is subject to cognitive and perceptual processes. First, an audiovisual stimulus is perceived through sight and hearing senses. Received signals are converted into electrical signals transmitted to the brain. Then, signals are processed cognitively in the brain. Cognition, such as emotional responses, memory or decision making, roots from former perceptual experiences but also impacts future ones. It is why such processes are individual-specific.

In subjective ratings, one collects scores that have been processed perceptually and cognitively by individuals. These results may be biased, consciously or even unconsciously, in many aspects. Experimental biases, learning effects, and subjects cognitive load are well know biased in experiments for instance.

Physiological signals are a collection of information which gives evidence of the processing of perceptual information. In that sense, perceptual and cognitive processes are observable through these signals analyses.

Moreover, these signals monitor physiological activities in time. This dimension is rarely efficiently observed through explicit rating. Lastly, data are gathered implicitly. If it is possible to influence what one says, it is not possible to control the way one's body reacts. As learnt during expectations analyses, "one should believe only what people do, not what they say."[Manski 2004].

[15] http://mmspg.epfl.ch/SoPMD

We thus conducted here studies which analyzed physiological signals as an indicator of QoE. We believe that ultimately these signals will be an efficient mean to predict the felt level of QoE accurately. Progress needs to be made in this regard, and we proposed to move forward by researching if the SoP, a high-level aspect of QoE is predictable.

The advantage of SoP is that it is a substantial component of QoE, mainly because more and more emerging technologies focus on developing the immersion and presence feeling during multimedia experiences. Several works also improve the natural and realistic rendering of contents to facilitate viewer's engagement and thus SoP.

By contrast with QoE, we can define levels of immersion that are likely to cause different levels of SoP, should they be different enough. We are presently not able determine contents specificities which will inevitably bring about a certain level of QoE.

Hence, physiological signals analysis were conducted to predict the experienced SoP. To the best of our knowledge, these are the first experiments which investigate SoP physiologically.

Regarding observable signals, there were many options opened to us.

First, brain activity, which relates to the CNS, is mainly recorded by means of EEG signals. Such signal time dimension is more precise than its spatial resolution, which is a definite advantage compared to other modalities (e.g., fMRI and PET).

Secondly, to observe peripheral activities (signals out of the brain and spinal cord) many different modalities may be used, such as ECG, GSR, or again skin temperature. ECG and respiration signals were included in our tests. They are widely used modalities for peripheral activities recording.

Concerning the conducted test, in the first one ILs were defined based on immersion parameters of contents, i.e., quality (aka compression rate), resolution and sound system. The second evaluated scenarios more in line with habits of multimedia consumptions.

Both tests included same original test sequences and equipment for signals acquisition.

Regarding test sequences, it had become apparent that contents more representative of broadcasting is needed. Ten one-minute sequences, among which one was dedicated to training sessions, were extracted from four open source Blender movies. This selection relied on the amount of energy in rear surround sound signals as well as on spatial and temporal complexities. Sequences duration is a trade-off between having a long length to cause presence and limiting the duration due to subjective tests constraints.

The acquisition system EGI's GES 300 of the company Electrical Geodesics enabled the acquisition of the three physiological modalities. A dense-array GSN of

256 electrodes recorded EEG, two standard electrodes captured ECG and thoracic and abdomen respiratory inductive plethysmography belts enabled the digitization of respirations waves. The Net Station software digitized, displayed and stored physiological signals at a temporal resolution of 250 Hz.

In the first study, stimuli of the low level of immersion presented no audio, a high compression rate and were rendered in SD. Middle IL stimuli were set to HD with stereo sound and were compressed transparently. High IL stimuli were defined as UHD surround sound contents compressed transparently.

Stimuli were rendered on a 4K professional screen and Sony Trimaster SRM-L5607 with the Lansing 5.1 THX speaker system.

The experiment implemented a SS with a 9-point ACR scale assessment methodology. This assessment enables a refined categorization of stimuli in three classes.

Five criteria were assessed per stimulus, namely interest in audio and video content, perceived video quality, IL (aka presence), and surrounding awareness.

Twenty subjects, including eight females, took part in our experiment.

Based on individuals ratings, subjects experienced three distinct levels of presence, which enabled the conducted three-classes classification. Our results also pledged for the use of surrounding awareness and video quality as influence factors in SoP evaluations. Accordingly, these factors must be considered in future tests about QoE.

Functional connectivity-based features were extracted from EEG signals while spectral analyses derived features from HRV and denoised respiration signals. A SVM with a RBF kernel was used as supervised learning tool. A leave-one-out cross-validation estimated its accuracy.

The classifier discriminating the three defined ILs reached a high accuracy level. 94 and 63 % of classification for high and low ILs were achieved, knowing that random classification is 33 %.

It demonstrates the capability of using physiological signals for SoP measurements, towards objective metrics and deep knowledge in sensory and cognitive processes of QoE.

The second work extended these findings by exploring more typical usage of multimedia technologies. ILs were defined as experiences provided by an iPhone, an iPad and a UHD screen at their maximum capabilities. Their ability to cover users field of view hinted that they cause sufficiently different experiences of presence.

The same 4K screen and surround sound system were used for high IL, while low and middle IL were rendered on iPhone 5 and iPad 4, respectively, with a professional stereo headset Sennheiser HD 280 Pro as sound system.

The 27 stimuli were assessed following a 9-point ACR during SS evaluations.

Six criteria were included in the assessment, namely interest in audio and video content, perceived audio and overall quality, IL, and surrounding awareness.

A balanced population of twenty subjects participated in the test.

MOSs showed that subjects experienced only two ILs. Low and middle presence stimuli were not perceived dissimilarly. In this context, the classification of stimuli was binary.

Again, our results showed the importance of evaluating surrounding awareness and overall quality as indicators of SoP.

Although our results hinted that interest in video favors the experience of presence, no statistical validation confirmed this finding. The relatedness of interest in content (aka users preferences) with SoP and QoE must be verified in future experiments.

CSP-based features were extracted from EEG signals while spectral analyses derived features from HRV and denoised respiration signals. A SVM with a RBF kernel was used as supervised learning tool. A leave-one-out cross-validation estimated its accuracy.

Several parameters settings such as electrode system, frequency sub-bands selection, EOG denoising, and delay were investigated in this work.

- CSP-based features lead to better predictors if the international 20-electrode system is used.
- Alpha and theta frequency bands showed to provide meaningful features. Hence, SoP is likely linked to relaxation and information encoding.
- Removing EOG artifacts lessen the prediction accuracy. It appeared that blinking or eyes movements could be related to SoP. It is a direction for future developments.
- We introduced a delay in considerations of EEG samples as SoP does not happen straight after a stimulus starts. This investigation was inconclusive.

Two different classifications were operated. The first one relied on defined ILs to define experiences classes. It verified if we can classify multimedia experiences as a function of rendering devices. The second relied on assessed ILs for classification. The model learnt to predict if the experience of presence was medium or high.

The first classifier can discriminate 4K TV from iPhone experiences with an accuracy of 90 %. This shows that CSP efficiently extracts features for this purpose. It also leads to the possibility to extend the analysis of physiological signals to other research fields in multimedia such as retargeting and broadcasting QoE-based applications.

The second classifier performed poorly. Middle and high levels of presence were predicted with an accuracy of 61 %, knowing that random classification is 56 %. This may indicate a difference in presence and immersion constructs which is a fact already known in the field. Being able to point at differences between these two concepts is a challenging though interesting future study. However, with regards to the accuracy of the previous classifier, this result also strongly advocate for the irrelevance of CSP features for the study of SoP. We thus recommend to use functional connectivity or spectral analyses for SoP prediction.

Overall, many outcomes of these studies are beneficial for the research community.

- We verified that immersion does not imply presence. Different levels of immersion can cause similar presence effect. The inverse is also true; similar immersive experiences may cause different levels of presence.

  Accordingly, we suggest designing ILs based on immersion, as long as one verifies that ILs are perceptually different.

- Both studies demonstrated a strong correlation between quality and presence. Similarly, there was a high inverse correlation between presence and surrounding awareness. We urge QoE researchers to include these influence factors in their studies.

- A strong correlation between presence and interest in video content has been observed. Our results did not validate this finding. However, we believe that interest in content is related to SoP, and even more to QoE. We thus strongly advise conducting works on this question.

- It is recommended not to use CSP for EEG features extraction in evaluations of SoP. However, this guideline does not stand for evaluations of other aspects of multimedia experiences.

- The most important outcome of these studies is that physiological signals analyses offer a great opportunity for studying QoE. We have seen that predictors can reach really high accuracies. Future objective metrics can be built on these predictors. Additionally, spectral analyses of EEG can provide new insights into perceptual and cognitive processes. Such knowledge is valueless as models of QoE could be more valid.

  We only want to specify that the quality of the predictor depends on the quality of the fed inputs. It is thus mandatory to continue the efforts towards evaluations of SoP and ultimately QoE to be able to predict them in the most precise way.

6.7 Future works

Future works have already been implemented. Our works and datasets have been used and referenced in the past years.

Regarding our contributions, we showed that more high-level concepts than quality could be evaluated and predicted through physiological signals. Researchers can build on top of these results to relate SoP to specific brain lobes activities and cognitive processes.

This research has been done in [Abromavičius 2017]. Authors conducted a spectral analysis of the EEG signals of our second dataset. As we did, they remarked

significant activities in theta and beta waves in all lobes. Additionally, they noticed significant changes in high beta waves in parietal and temporal lobes. They linked SoP to specific frequency bands and brain lobes while defining its relatedness to mechanisms of memory and learning, emotional involvement as well as arousal or cognitive fatigue.

Our datasets have also been listed in reviews on physiological signals analysis in multimedia fields. For instance, in [Engelke 2017, Moon 2017] and various other works on video quality and digital storytelling.

The first point that needs to be raised is the extension of this work to QoE. Reaching similar results for QoE would be tremendous progress for QoE modeling. Reliable predictions of QoE would be a strong basis for the development of emerging and new multimedia, especially immersive ones. Content creators would have more knowledge to create best experiences. Broadcasters could improve current services and developed user-centric new ones. The entire multimedia community could benefit from it.

Presently, one will have a hard time creating the database that is needed to learn the model. We believe that recording physiological signals when presenting a high number of different stimuli could be a successful methodology. However, it is very demanding in many aspects such as manpower, time and resources (even more than usual subjective tests).

To my mind, several aspects of physiological signals are not exploited. So far, and to the best of my knowledge, few studies took advantage of the high temporal definition of these signals.

A first application that appears to me is the prediction of SoP and QoE in real time. Variation in specific frequency bands and brain lobes could be detected in real time and fed to a simple model for prediction. Broadcasting services could adapt their services based on body responses of viewers. With the fast and extensive deployment of wearables, signals can be made available for such an application. Nowadays, it is already possible to access respiration and heartbeat signals through wearables.

A second application would be to go further in the study of what causes SoP and QoE. Temporal information could be split into presence phases or no presence phases. Then, physiological signals of presence phases would be analyzed to get more precise brain activities in specific lobes and frequency bands that are related to SoP and QoE. Based on such information, one could predict SoP or QoE of contents extracts. This would be a powerful tool for movie creators. The main challenge to tackle here is to be able to detect phases of presence. We have proposed a solution for this new direction towards a better prediction of SoP using physiological signals at the International Young Researcher Summit on Quality of Experience in Emerging Multimedia Services (QEEMS) in 2017 [Perrin 2017c]. We proposed to rely on relaxation states to identified phases of presences as our second study showed that

alpha-band variations are an indicator of SoP.

Let us go even further. Nowadays, many efforts are paid to define what are best practices to tell a story in VR environments. The main issue is that, most of the times, the attention of the viewer must always be drawn onto specific viewports. Today, it is hardly possible to discriminate two different storytelling approaches, especially on different contents. Having a tool able to predict QoE (in real time) based on physiological signals of a sequence would fasten and improve this process.

Our works leave a lot of room for improvements in this field but set a strong and reliable basis for the development of SoP and QoE models.

Chapter 7

Conclusion

Nowadays, multimedia technologies and services are prevalent. This powerful mean of communication is ubiquitous and presents itself in one's everyday life in a wide range of applications, from advertisements on the automaton on the fuel station to movie screening in cinema. We witnessed these last decades the exponential growth of multimedia services, including social networks and video on demand. Audiovisual content amount of consumption and sharing increased continuously.

The audiovisual industry is in constant evolution, current services are improved, and emerging, and new formats are designed. Multiple axes of development are under study, such as reaching efficient compression for large-size videos and images, improving contents properties such as level of details and definition (e.g., 4K and high frame rate), increased immersion (e.g., 360°, VR, and volumetric sound), and more realistic and vivid representations (HDR and WCG), to mention but a few.

These emerging or new experiences are rich and high-quality. They rarely violate quality requirements and thus make quality evaluations obsolete. This comment also applies to QoS, as broadcasting of content can achieve uncompromised transmission accuracy and efficiency. Multimedia field has reached a certain maturity, and previous ways of evaluation and monitoring are not sufficient anymore to provide information for future developments.

Appraisal of services is of capital importance. For instance, any development methodology, such as the AGILE process, requires evaluations before and after releasing a product to design the best service possible and making sure its implementation is satisfactory. A new measure had to be defined to cope with this need.

In the last decades, services were bounded by technology constraints. It is mostly not the case anymore, which justifies that other parameters can be taken into account from now. Service providers need to satisfy their users to gain market access, acceptance, and share. It was thus about time to add consumer-related considerations in evaluations.

To address this concern, the Qualinet consortium defined such a metric in 2013, the QoE. "Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/ or enjoyment of the application or service in the light of the user's personality and current state [Brunnström 2013].". This definition has been widely accepted by the research community and the ITU standardized this definition in 2016 [Amendment 2016].

Numerous works started defining and using QoE. However, either QoE was unwell understood and assessed as quality or QoS, or specific context-related influ-

ence factors were defined for the conducted analysis. QoE aims at evaluating any emerging or new technologies. There was then a need for recommendations about how to assess QoE, for instance how to derive influence factors from the context of the study. At this moment, to the best of our knowledge, no recommendations from standardization committees were provided.

We proposed in this thesis to address this need from the audiovisual field. From our outcomes, we derived guidelines for test designs of QoE assessment. Simultaneously, our research findings contributed to recommendations about the development of several content representations.

In the very near future, applications developers, compression engineers or light-field experts will need to be able to conduct their product/service analyses using QoE as a measure of consumers satisfaction. They need to be provided with standards, recommendations regarding the assessment design.

Proposing valid QoE evaluation procedures allows the community to conduct more precise and reliable tests. Eventually, QoE subjective tests will be performed to evaluate new services and product. Future applications depend on the accuracy and representativeness of these methodologies.

Also, objective models are usually based on subjective ratings or are validated by individual assessments. Ultimately, QoE models will be learnt or certified by studies implementing these appraisal procedures.

To develop QoE subjective assessment methodologies has a considerable impact on future improvements, both in QoE research field and industry.

1. First, we conducted two quality experiments to understand the lacks between current methodologies and QoE evaluation procedures to develop. We took advantage of conducted experiments to provide meaningful insights into the design of the delivery pipeline of HDR WCG imaging.

    - We investigated the difference between the legacy Y'CbCr and a new color model for HDR representations, the ICtCp. This new color space addresses limitations of Y'CbCr in terms of the achromatic channel I, isoluminance, hue linearity, perceptually more uniform colors, and bit depth limitation.

      A partial PC SbS, with repetition methodology, asking about representations preference was conducted.

      Our results do not show a clear preference of one color representation over the other. However, ICtCp provided slightly higher perceptual quality than Y'CbCr in low bandwidth constraints. Y'CbCr exhibited a little preference when representing contents at high bit rates.
    - A reshaper, an EETF optimizing code words to efficiently represent contents before compression, was evaluated for both Y'CbCr and ICtCp color space. The methodology selected was a partial SbS PC, with repetition.

This operation enables bit rate savings. We considered signals reshaped at 0, 10 and 20% of the initial bit rate.

Reshaper saving 20% of bit rate impacted the perceptual quality too much when compared to the anchor. Both other reshaper operations achieved similar or preferred perceptual quality. The reshaper configuration that reached higher preference scores is the reshaper saving 10% of the initial bit rate, especially if contents follow the ICtCp color model.

- We conducted a comparison of the three main approaches of compression in HDR. SLBC strategy is to send a compressed SDR stream and to predict HDR from this stream and transmitted metadata. SLNBC follows the same scheme but sends an HDR stream and retrieves the SDR. The DLBC builds on top of the SLBC and sends an enhancement stream containing encoded residuals of errors made by the prediction of HDR stream.

  Three proportions of the overall bit rate allocated to the enhancement layer of DLBC have been evaluated, namely 10, 15 and 20%. Overall bit rates of 4, 5, 6 and 8 Mbps were considered as bandwidth settings. The study evaluated both SDR and HDR streams levels of impairments. A SbS pairwise comparison test methodology was selected to present compressed contents and uncompressed reference contents.

  Regarding overall bandwidth, our results advocated tuning HDR delivery pipeline to 6 Mbps. Concerning compression schemes, DLBC performed as good as SLBC in the SDR evaluation and similarly to SLNBC in the HDR test. The most performing configuration was the DLBC using 6 Mbps of overall bit rate and allocating 15% of this bit rate to the enhancement layer.

With respect to QoE evaluations, we identified that assessments should include the review of QoE influence factors to understand viewers behavior better and relate scores to observable dimensions. These influence factors are context-based. Indeed, instead of providing a list of influence factors, we need to provide researchers and industrials with means of influence factors extractions.

2. The next presented work is an investigation of the interest of addressing SDR limitations in omnidirectional contents. Combining HDR imaging and 360° formats make experiences realistic, immersive and engaging.

   One one hand, we had to tackle the challenges of working on this format.

   - I acquired the first camera-consumer HDR 360° images dataset. The camera Ricoh Theta and its two wide-angle captors enabled the capture of five different exposure omnidirectional representations per scene.

     The 43 captured scenes, their HDR reconstruction and mapped versions are publicly available on the MMSPG website[1].

[1] http://mmspg.epfl.ch/360hdr-consumercamera

Among the captured contents we selected a set of eight images presenting a wide variety of spatial complexity, DR, and overall brightness. We designed a new characteristic for content selection, the variation in key values along viewports.

- No rendering system for omnidirectional contents are compliant with HDR. We thus had to map HDR contents to match the SDR display of the selected rendering system, a Merge HMD.

  The four widely used gamma, photographic, display adaptive and detail preserving TMOs were applied after the reconstruction of HDR contents to display contents on the HMD. An exposure fusion operator, bypassing the HDR reconstruction step, was also included in the test. TMOs and exposure fusion variants were the test contents.

- The experiment consisted of a comparison between the same content HDR mapped versions and original SDR. Mid exposure contents are representative of SDR imaging as neither highlights nor shadows are favored. They constituted our reference contents.

  To the best of our knowledge, no pair comparison methodology has been proposed for omnidirectional contents. We developed our own, namely the toggling. It allows subjects to go back and forth between test and reference content to exhaustively compare stimuli while reducing cognitive load and caused sickness effect.

  The Merge VR has two entries. Each of them controlled toggling or rating interactions. We built on top of the omnidirectional testbed [Upenik 2016] to implement the toggling methodology. Timestamps and locations of toggling were recorded.

On the other hand, we designed a post-questionnaire which investigates potential influence factors for this format.

- First, we defined influence factors extracted from a review of the HDR state of the art. We included most recurrent evaluations dimensions: details in the bright areas, details in the dark areas, unnatural colors, ghosting, noise, overall brightness, overall contrast and naturalness of the scene.
- Subjective ratings are content-dependent. Spatial and temporal complexity, textures colors impact the perception of a content. Besides, to our mind subjects appreciation of content as well. Hence, we investigated the aesthetic properties of the contents. Subjects were asked if they found contents boring, interesting, colorful, aesthetic, familiar, of quality, pleasurable and immersive.
- Experiences with HMD, and more generally VR, are known to generate sickness effect. This effect threatens the validity of our test. We thus decided to verify the extent of sickness generated by our test. We used the

simulator sickness questionnaire, asking about the following symptoms, general discomfort, fatigue, headache, eye strain, difficulty focusing, increased salivation, sweating, nausea, difficulty concentrating, "Fullness of the head", blurred vision, dizziness with eyes open, dizziness with eyes closed, vertigo, stomach awareness, and burping.

- Regarding omnidirectional contents, we investigated immersion, isolation, enjoyment, arousal/excitation and enjoyment due to the novelty of the visualization. Additionally, we added an open question about the subject's VR experience.

Our results exhibited no apparent increase in perceived quality. However, the exposure fusion, as well as the linear TMO to a lesser extent, showed promising results as being assessed as similar or slightly better than single exposure reference. Considering camera-consumer contents with DR under professional contents, this finding is sound and promising for the HDR omnidirectional imaging.

The processing of toggling information provided us with numerous interesting outcomes. Subjects paid attention primarily to furthest objects and loss of data. They did not focus much on bright or dark areas to verify the rendered level of details.

Regarding influence factors, all HDR dimensions were proved suitable for evaluation. Half of the subjects experienced up to slight sickness. Hence, the toggling methodology is not threatened by this effect. Evaluation of immersion, isolation, the subject's state of mind and novelty effect was not as informative as expected. It seems that, except for the navigation through the content, HDR aspects prevail over omnidirectional content factors.

Assessing multiple dimensions of QoE brings much richer analyses. To extract influence factors, we recommend an exhaustive review, the extraction of relevant aspects related to the context of the study, and a validation of selected dimensions (as done in this study). Our results pledged for the use of extracted HDR factors and showed the benefits to include aesthetic evaluation of contents.

For HDR 360°, it is recommended to assess stimuli regarding bright and dark areas. Several subjects complained during our experiment about feeling torn between preferring a stimulus in shadows, but the other in highlights.

3. In two studies, we investigated more user-centric aspects that influence QoE. Our motivation rooted in the possibility to evaluate influence factors of QoE which are common to all immersive technologies.

   Those aspects were evaluated on VR gaming experiences. Relatively new, VR gaming is all the rage for several years. It is thus a perfect use-case to study the novelty effect and expectations.

Each test implemented a SS evaluation methodology.

- Novelty effect is an effect that bias perceptions the first times a product/service is encountered. It potentially threatens the external validity of experiments. Hence, it has been the subject of an exploratory study to understand the source of such an effect.

  Novelty effect is rarely defined in immersive media literature. We proposed a definition for this effect: *The novelty effect is the transitory and unconscious tendency of subjects to evaluate their experience differently the first times they are using a product, not because of the product intrinsic features, but because the user perceives some of those features as unfamiliar.*

  We determined seven dimensions that impact or are impacted by the novelty effect. They were interest, experience, technical features, expectations, relative advantage, conscious newness, and current mood.

  We compared the scores of two populations having initially no experience at all with HMDs or VR. Before the test, one population took part in familiarization sessions targeting the reducing of the novelty effect.

  The only significant variation observed is that of expectations. We thus decided to focus on expectations analyses.

- We have conducted a cross-platform comparison between VR and typical gaming platforms, namely mobile and computer.

  We selected two games which have the same game design and level design on all three platforms. It must be noted that the Gunship Battle game on the computer was a mobile version played on an Android emulator.

  We reviewed most used questionnaires for gaming experiences and presence and selected the PENS as core questionnaire. Additionally, we added questions on overall quality, presence, immersion, and enjoyment. Indeed, we wanted to verify if questionnaires number of items could be reduced by asking one overall question instead of several precise ones.

  Expectations were evaluated in pre- and post-questionnaires according to core-questionnaire dimensions (i.e., quality, intuitiveness of controls, immersion, presence, enjoyment and competency).

  HMD provided higher presence than the two other platforms, especially physical presence. Although experiencing presence may have a positive effect on overall quality, the impact extent is not necessarily significant. However, we verified that competency increase leads to overall quality raise.

  Sickness effects, caused by the intense HMD Gunship Battle game, have been counterbalanced by the enjoyment of the experience. This remarkable finding must be verified for more extended gaming sessions.

Regarding expectations analysis, a first important outcome is that pre-expectations indicated which influence factor is more related to one of the platforms. Expectations are a good indicator for influence factors selection.

The gap model (expectations - MOSs) hinted that using a game on an emulator is a defect in the test design. Results were highly representative of subjects satisfaction and are thus valuable to understand scores further.

Expectations variations (revised expectations - expectations) presented insights that were not visible in usual analyses. We observed subjects behavior depending on preferred platforms. Such variations express user-specific conducts that are worth studying.

The combination of gap model and expectations revision offered another crucial observation. We could note that ideal expectations were adjusted based on the potential of technology and not necessarily on previous experiences.

Our results showed the tremendous interests to include expectations analyses in QoE evaluations.

4. Explicit self-assessments are biased by numerous effects such as experiment or learning effect. It is thus preferred to observe individuals reactions instead of what they say. In two works we analyzed the physiological response of subjects to stimuli of different levels of SoP.

   SoP is a high-level aspect of QoE which gains momentum every day, mainly because immersive technologies will form most of the future multimedia experiences. It is an influence factor mutual to all immersive multimedia.

   Brain, heart, and respiration activities were recorded to observe viewers implicit response to visualizations. Signals we collected thanks to the high-quality acquisition system EGI's GES 300 combined with a dense-array GSN of 256 electrodes to record EEG, two standard electrodes for ECG acquisition, and thoracic and abdomen respiratory inductive plethysmography belts which digitize respiration signals.

   We aimed to verify the possibility to predict SoP from physiological signals, to ultimately develop objective models of QoE and to better understand perceptual and cognitive processes.

   - Three levels of SoP were defined in a first study as parameters of contents immersiveness characteristics, namely quality (aka compression rate), resolution and sound system. Stimuli were presented in a SS fashion and rated accordingly to a 9-point ACR scales. Interest in audio and video content, perceived video quality, IL (i.e., presence), and surrounding awareness were the five criteria assessed per stimulus.

The scores analysis confirmed that three levels of SoP were experienced. Besides, high correlation demonstrated that video quality and surrounding awareness are influence factors of SoP. Accordingly, these factors must be considered in future tests about SoP, and eventually QoE.

Functional connectivity-based features were extracted from 20-electrode EEG signals while spectral analyses derived features from HRV and denoised respiration signals. A SVM with a RBF kernel was used as supervised learning tool. A leave-one-out cross-validation estimated its accuracy.

The classifier discriminating the three defined ILs reached a high accuracy level. 94 and 63 % of classification for high and low ILs were achieved, knowing that random classification is 33 %. It demonstrates the capability of using physiological signals for SoP measurements, towards objective metrics and in-depth knowledge in sensory and cognitive processes of QoE.

- In a second work, we examined scenarios more in line with today's consumption habits. ILs represented the best experiences possible on a phone, tablet and 4K screen. The significant variation of field-of-view coverage motivated this selection. A SS methodology was selected, with self-assessments graded on 9-point ACR scales. The six stimuli criteria of evaluation were interest in audio and video content, perceived audio and overall quality, IL, and surrounding awareness.

  Based on ratings, subjects only experienced two significantly different levels of SoP, middle and high. The predictor of SoP was thus binary. Overall quality and surrounding awareness were confirmed to be SoP influence factors. Besides, our results suggested, though did not establish, a relatedness between interest in video content and level of SoP. Future tests should investigate this connection.

  CSP analyses derived features from 20-electrode EEG signals while spectral analysis-based features were extracted from HRV and denoised respiration signals. A RBF-kerneled SVM was the used supervised learning tool. A leave-one-out cross-validation estimated its accuracy.

  We have explored several settings for EEG features extraction. Highest performance was reached with the 20-electrode system, when not performing EOG artifacts removal, and when considering the entire one-minute recordings for all frequency sub-bands. It must be noted that alpha and theta frequency bands showed to be particularly discriminative, thus related SoP to relaxation and information encoding.

  Two ground truth were used for learning. The first one is defined ILs, enabling the prediction of which device has been presenting a stimulus. This first classifier can discriminate 4K TV from iPhone experiences with an accuracy of 90 %. It showed that CSP efficiently extracts features for this purpose.

The second classifier was based on assessed levels of SoP. It performed poorly. Middle and high levels of presence were predicted with an accuracy of 61 %, knowing that random classification is 56 %. A difference in presence and immersion constructs are proved to be different. Being able to point at differences between these two concepts is a challenging though interesting future study. Also, with regards to the accuracy of the first classifier, this result also strongly advocates for the irrelevance of CSP features for the study of SoP. We thus recommend using functional connectivity or spectral analyses for SoP prediction and not spatial examination.

Overall we verified that immersion does not imply presence if experiences are not sufficiently different. We have confirmed that quality and presence must be SoP influence factors. We recommend future evaluations on the link between presence and QoE to interest in video content.

Regarding physiological signals, we advise not to analyze these signals spatially when prediction SoP. We, however, are thrilled to announce that physiological studies bring more knowledge about perceptual and cognitive functions while leading to accurate predictors for multimedia experiences SoP.

Our results are numerous and cover a wide range of applications for future immersive multimedia services. In the next section, we present our results sorted by technology-based outcomes and by QoE-related findings.

7.1 Sorted outcomes of our works

In this thesis, we investigated numerous methods for QoE evaluation and prediction. The performed studies generally had a two-fold impact. The first one is the context of study outcomes based on the analysis of the results. It brought new and valuable information on technologies evaluated to the research community. The second is the benefits towards guidelines for QoE assessment.

7.1.1 Study-based outcomes

We start by summarizing the valuable information for the research community, not related to QoE.

It should be noted that specific attention has been paid to evaluate realistic scenarios, which were selected to be as close as possible to typical real-life situations.

7.1.1.1 HDR and WCG

Along the studies towards the improvements of HDR and WCG representations, we observed that neither Y'CbCr nor ICtCp color spaces had been clearly preferred. However, results indicated that in case of dramatic bandwidth constraints, ICtCp is perceptually more performing.

Overall, the combination of ICtCp and the reshaper operator is an efficient representation as it guarantees a similar perceptual quality while making bit rate saving possible. Our results advocate for tuning the reshaper parameter at 10% of bit rate decrease. Higher visual preferences were observed when ICtCp use was coupled with the reshaper optimization.

Regarding compression solutions, the DLBC realized similar high-quality delivery as uncompromised SDR and HDR. It is thus the optimal compression solution for broadcasters due to its backward-compatibility characteristic. Our results strongly advise against the use of the SLBC scheme, HDR reconstructions were assessed perceptually as not high-quality.

We have shown that constant bandwidth constraints can be set at 6 Mbps for HDR delivery. It is recommended to tune the DLBC at 15 % of bitrate allocated to the enhancement layer.

7.1.1.2 HDR 360°

Through the work on HDR 360°, a images dataset containing the multi-exposures, the HDR reconstruction and the SDR variants of 43 omnidirectional contents was created. Gathered subjective scores, tracking, and toggle information are also included in the dataset.

We designed the first PC methodology for omnidirectional contents, along with propositions for data analysis. Its value has been proven by the richness of our results and the fact that at most slight sickness was felt.

Our results gave interesting insights in HDR 360° experiences: (1) the initial viewport has no influence on subjects' assessment, (2) the exploration of the content is mostly longitudinal, (3) subjects tend to focus on distant objects, and (4) the loss of information of light emitting areas is not considered in the quality judgment. Besides, the variation of key-value appeared to be a good indicator of the characteristics of omnidirectional contents.

7.1.1.3 VR gaming

The cross-platform gaming comparison showed that VR gaming induces a higher spatial presence in subjects when compared to mobile and computer platforms.

However, the impact of presence on overall quality was not necessarily significant. The dimension that had a significant relation with overall quality was competency. Game designers must pay attention to this factor when developing the level design of games.

Interestingly, sickness may be overcome if the SoP and enjoyment are high enough. Classification of games severity could include these two factors.

7.1.1.4 Typical scenarios of consumption

The SoP induced by an iPhone and iPad were not perceived as different. This finding is an observation of the fact that different ILs may generate the same levels

of SoP. Furthest investigations may provide compelling knowledge on what is the difference between the two concepts.

7.1.2 QoE-related results

With regards to QoE, several highly valuable findings result from the performed studies.

7.1.2.1 Unimodal studies

First of all, from quality evaluations we identified that unimodal assessment, may it be quality, preference or degree of impairments, is not sufficient to understand user behaviors.

We first lacked user-related information, such as their state of mind, preferences, expectations and prior experiences. Secondly, to further understand user behavior, one needs to gather details regarding the characteristics of the technology that influence the judgment of multimedia consumer.

We have emphasized the fact that evaluations of QoE must be multimodal and not unimodal. A trade-off must be found in questionnaire designs between asking too many questions, which are generating subjects' fatigue and lack of attention, or not enough, which prevents a thorough analysis of results. In a test, we have seen that one question may replace nine items questionnaire for presence evaluation. The balance is to find in the level of details required to evaluate a concept fully.

7.1.2.2 QoE influence factors extraction

Technology-based features are determined depending on the study context and are thus defined on the case-by-case basis. These pieces of information are common knowledge in the QoE field but are still not thoroughly evaluated.

In the HDR 360° study, we evaluated HDR-specific criteria, VR-related aspects, aesthetic properties of contents and asked subjects to self-describe their experience. Additionally, their level of familiarity with HMD was collected. Even though these data are not directly related to the preference of the TMO over the reference content, they give insight into the judgment process.

From VR technical features, we could validate the test method, as overall, slight or no sickness was experienced. The criterion unnatural color, despite being assessed as the 6^{th} most significant HDR factor, is of importance in the expression of subjects' opinion, based on aesthetic and technical factors. We thus strongly recommend to include the appraisal of precise technical features in QoE evaluations, as well as context-based features such as aesthetic perception of contents.

7.1.2.3 Novelty effect

During the exploratory analysis of the novelty effect, a new and broader definition has been created, along with the description of its influence factors.

Our study revealed that the novelty effect is highly related to users expectations. Consequently, we decided to focus on expectations analyses.

7.1.2.4 Expectations analyses

From the VR gaming evaluation, we have implemented the assessment of expectations. It should be noted that expectations are influence factors of QoE which are common to all immersive technologies. However, their assessment is context-based as they must be assessed on the technology-based features.

We have observed that pre-test expectations are a good indicator of which technical features are meaningful.

The E-P gap model gave additional insights into the subjects' assessment. Indeed, the gap model is the only result that clearly showed the low satisfaction of subjects regarding the computer version of Gunship Battle, which is due to the emulator use.

Likewise, the collection of these data enabled more in-depth expectation variations analysis, which showed us notable discrepancies between subjects, depending on users' preferred gaming platform. However, this result should be carefully interpreted, as the low number of subjects in each sub-category prevents from generalizing the findings. Nonetheless, the importance to observe such behavior is real.

It is interesting to conduct both gap model and expectations variation as it may give clues on the creation process of expectations. For instance, in this study, it seemed that ideal expectations were revised based on a prediction of future experiences rather than on the experience itself.

7.1.2.5 Physiological signals analyses

Last but not least, we have investigated two different approaches for the extraction of EEG features. In the first case, spectral analysis implementing functional connectivity-based features realized a reliable predictor for high SoP and a reasonably accurate one for middle IL.

When implementing a spatial-based feature extraction, the prediction of rendering systems reached high accuracy while the SoP classifier was almost random. We concluded that spatial-based features are not appropriate to physiologically evaluate the SoP. An identification of rendering device was nonetheless accurate base on these spatial-based features.

Those are promising results for the multimedia field, as many application may be derived from the rendering system or SoP prediction. For instance, from knowledge on perceptual and cognitive functions used when experiencing presence, accurate objective models can be created.

Two datasets containing subjective rates and physiological signals have been made available to the public.

7.2 Future perspective

Recommendations, datasets and experiments results presented above form a valuable framework for context-based QoE evaluations in immersive multimedia. This thesis sets a basis on which the research community can build upon to evaluate and develop richer multimedia experiences. The three databases are publicly available for this purpose.

Regarding the diversity of addressed issues, future perspectives are presented per technology and then for QoE research.

HDR representation and compression

Our tests on ICtCp aimed at indicating to standardization committees whether they should recommend this color model for HDR formats. ICtCp showed higher preference rate at low bit rates when compared to the conventional color space. In addition, several technical advantages, such as bit depth limitation advocate for the use of ICtCp.

The entire rendering pipeline used the PQ EOTF. Another TF, the HLG, is standardized in ITU recommendations and is equally as used as PQ in the field. ICtCp must be compared to Y'CpCr in less optimal conditions, meaning with the use of HLG OETF. After such study, one should be able to confirm or not the eligibility of ICtCp as HDR color space into standardization recommendations.

Similarly, the reshaper should be tested on a pipeline using the HLG. Its advantages must be confirmed for all stimuli, and this includes contents rendered through HLG application.

Let us assume in the following that the reshaper has been introduced in standardization recommendations. Regarding compression scheme selection, a comparison between reference, and reshaped and compressed HDR streams can be conducted. The use of the reshaper may dramatically impact the SLBC efficiency and put everything into perspective in the event that the DLBC codec is not able to compete anymore. I believe that by this time, the transition from SDR to HDR would have been implemented and that broadcasters will welcome a higher-efficiency HDR-based compression approach.

HDR 360°

Many opportunities and future developments are about to happen for this format.

We believe that our experiment can be extended to professional and stereoscopic contents. Professional contents will provide higher DR while stereoscopy will show more accurate depth information.

Based on our recommendations, TMOs dedicated to this immersive and realistic format can be developed. Their main difference with 2D TMOs is that in omnidirectional contents, subjects attention is not drawn on omnidirectional specific areas. Algorithms must be revised accordingly. Besides, several approaches can be consid-

ered. TMOs can be processed in real time per HMD viewport or being computed before visualization on the entire content.

We also have seen in the literature that the combination of two TMOs views (one TMO image for each eye) are perceived as richer experiences. Our experiment can also be extended with this visualization technique to assess which combination of TMOs provides the highest QoE.

More importantly, future tests must include influence factors evaluation at the end of every stimulus. We recall that these influence factors are details in the bright areas, details in the dark areas, unnatural colors, ghosting, noise, overall brightness, overall contrast and naturalness of the scene. We also support the inclusion of aesthetic post-evaluation.

VR gaming

We have shown that competency is a dimension significantly related to overall quality. Numerous influence factors of game quality, not included in our analysis, can be evaluated for VR gaming. For instance, various gaming experience factors, identified in [Möller 2013] (e.g., flow, tension, aesthetics, and learnability), may be appraised soon.

Our experiment may be extended regarding the equipment used. For instance, embedded-HMDs exhibit different experiences than consumer mobile-HMDs. Such a test would further explore the differences across platforms but may also show the difference between the two types of HMD. Such knowledge would provide game providers clues to better target parts of the market.

Other games genre can be evaluated following our test methodology. At the time of our study, only a highly limited number of games fulfilled our conditions of games having the same game design and level design. Considering the fast development of VR gaming, we can assume that shortly, new games will appear and could be evaluated in a cross-platform comparison. The more diverse games are evaluated, the more representative will be the results.

Physiological signals

Researchers can use our publicly available datasets to relate SoP to specific brain lobes activities and cognitive processes. This work has been performed by [Abromavičius 2017]. Authors conducted a spectral analysis of the EEG signals of our second dataset. As we did, they remarked significant activities in theta and beta waves in all lobes. Additionally, they noticed significant changes in high beta waves in parietal and temporal lobes. They linked SoP to specific frequency bands and brain lobes while defining its relatedness to mechanisms of memory and learning, emotional involvement as well as arousal or cognitive fatigue.

We showed that it is possible to evaluate high-level concepts of QoE through physiological signals. The next step is actually to predict QoE. Reliable classification of levels of QoE would be a powerful tool for the development of emerging

and new multimedia, especially immersive ones. Content creators would have more knowledge to create best experiences. Broadcasters could improve current services and develop user-centric new ones. The entire multimedia community could benefit from it. We, however, note the huge challenge to overcome that is to create the database on which predictors will be learnt.

Due to their promising results, physiological signals are becoming popular. Lately, a sensor-equipped HMD that records EEG, ECG and EOG has been built for studying QoE in immersive multimedia [Cassani 2018]. Many invested resources already help developing QoE models.

The high temporal resolution of physiological signals has not been exploited yet. First of all, the possibility to predict SoP and QoE in real-time would be beneficial for numerous applications such as streaming under variable bit rate, VoD, behavioral studies, game designs or VR monitoring.

A second application is going further into the study of what causes SoP and QoE. Biosignals could be split into presence phases or no presence phases. Then, physiological signals of presence phases would be analyzed to get more precise brain activities in specific lobes and frequency bands that are related to SoP and QoE. Based on such information, one could predict SoP or QoE of contents extracts. This would be a powerful tool for movie creators. The main challenge to tackle here is to be able to detect phases of presence. We have proposed a solution for this new direction towards a better prediction of SoP using physiological signals at the QEEMS in 2017 [Perrin 2017c]. We proposed to rely on relaxation states to identified phases of presences as our second study showed that alpha-band variations are an indicator of SoP.

Finally, the performed works cover a wide range of scientific fields. For instance, neuroscience, perceptual science, psychology, multimedia signal processing, and statistics are areas of expertise necessary to develop, design, prepare and analyze the conducted evaluations. Soon enough, projects involving researchers from those fields will be a necessity. The inclusion of physiological signals in subjective assessment in the multimedia field illustrates this statement. Indeed, the more multimedia researchers will need to understand perceptual and cognitive processes, the more exchange with researchers from other areas, such as neuroscience, will be required.

QoE evaluations

First of all, exposed results would have a higher impact on the field if they are replicated, reproduced and validated by peers. Naturally, we encourage researchers to use (or even overuse) our guidelines.

Many challenges are still to be tackled. We indicated that pre-experiment expectations are a good indicator of technical-based factors. A similar process must be defined to identify user-centric factors relevant for thorough QoE analysis.

At this moment, I suggest to start including new user-centric influence factors

based on an educated guess. One can then estimate its appropriateness. For instance, we have included considerations of content aesthetic which provided us with interesting insights and allowed us to deepen observations. Also, our results suggested a relation between SoP and interest in video content that must be inspected. To our mind, the more consumer-related aspects are studied, the more we will get knowledge of QoE.

Based on my research, I had the intuition that novelty effect happens when ideal expectations are unstable, the first times one encountered an experience. Ideal expectations are always evolving, but I would expect them to change more extensively during this "honeymoon phase". Such a study would bring much knowledge on expectations formation and could ultimately help to predict the extent of bias introduced by the novelty effect. This final outcome is valuable for all media providers and multimedia technology manufacturer.

General aspects

From a more general perspective, this thesis laid the foundations for expectations- or physiology-based objective metrics of QoE. It is the ultimate goal for QoE: to help the industry to measure QoE in a reliable and time-efficient and cost-effective (in terms of workforce and resources) manner.

Concerning immersive and emerging technologies, the forecast for the future is really exciting as mulsemedia, point-clouds and light fields representations, in addition of those included in this thesis, are opening the way for even more realistic, faithful, accurate, immersive, multisensory experiences. I believe that measuring QoE must be context-dependent as each of these technologies has specific features.

Emerging experiences during which users can interact with the content is nowadays possible. This use-case was out of the scope of this thesis, as we decided to investigate immersive multimedia. Interactions increase the engagement of users, not their immersion. Interactive contents have been around for several years now, but are coming forward. For instance, Netflix announcement of the release of an interactive episode of Black Mirror end of 2018 shows the momentum that this format got. QoE must also be evaluated on interactive contents. The evaluation of interactive contents presents some challenges such that the need to assess all contents possibilities. Besides, not only the experience is evaluated but also the enjoyment related to viewer's narrative decision-making.

A last more distant perspective is the following. The futurologist Ray Kurzweil, an Artificial Intelligence (AI) researcher, has made several predictions for the coming decades. He announced in [Kurzweil 2010] that in 2030, nanomachines could be introduced in ones' brain and control incoming and outcoming signals. It results in the possibility to create a truly full-immersion virtual reality without the need for any external equipment. This fact strengthens the importance of using physiological signals in evaluations. Besides assessing experiences based on their contexts is mandatory. Indeed, nowadays evaluating devices and TVs makes sense but it will possibly not be the case anymore in a decade.

Chapter 8

Appendix Example

8.1 Post questionnaires for HDR 360°

The following pages show the post-questionnaire form given to subjects.

In this evaluation, you are asked to compare two virtual reality scenes. Each scene is labelled either as 'T' for test or 'R' for reference at its center. You can toggle between the scenes using the right button on your HMD. You have to toggle at least twice before scoring. Once you have selected your choice press the left button to vote. You can vote only when you are viewing the Test image (T). These instructions will be repeated before you start the test on the Virtual Reality HMD.

At the end of the test, you will be asked to answer to the following questions. Please read them before you start the test.

Did you experience sickness? Yes ☐ No ☐

If yes, then on a scale from 1 to 4 rate the following symptoms: (1 – None; 2 – Slight; 3 – Moderate; 4 – Severe)

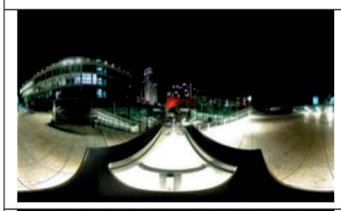
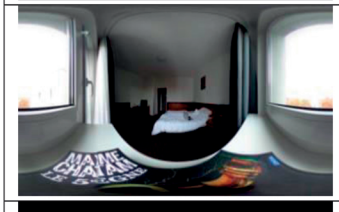
Symptom	Score
General Discomfort	
Fatigue	
Headache	
Eye strain	
Difficulty focusing	
Increased salivation	
Sweating	
Nausea	
Difficulty concentrating	
"Fullness of the Head"	
Blurred vision	
Dizziness with eyes open	
Dizziness with eyes closed	
Vertigo	
Stomach awareness	
Burping	
Others: _____	

Rank the following criteria considered when comparing the Test Scene 'T' and the Reference Scene 'R'? (1 being the most considered criterion while 8 is the least.)

Criteria*	Rank 1:Most considered 8: Least considered
Details in the bright areas	
Details in dark areas	
Unnatural colors	
Ghosting	
Noise	
Overall brightness	
Overall contrast	
Naturalness of the scene	

Others:

*Examples of the criteria are given on the last page.

Content	*Which of the following words best describes the scenes?* You can tick more than one.	*What did you like or dislike in this content?*
	☐ Boring ☐ Familiar ☐ Interesting ☐ Of quality ☐ Colorful ☐ Pleasurable ☐ Aesthetic ☐ Immersive ☐ Other: _____	
	☐ Boring ☐ Familiar ☐ Interesting ☐ Of quality ☐ Colorful ☐ Pleasurable ☐ Aesthetic ☐ Immersive ☐ Other: _____	
	☐ Boring ☐ Familiar ☐ Interesting ☐ Of quality ☐ Colorful ☐ Pleasurable ☐ Aesthetic ☐ Immersive ☐ Other: _____	
	☐ Boring ☐ Familiar ☐ Interesting ☐ Of quality ☐ Colorful ☐ Pleasurable ☐ Aesthetic ☐ Immersive ☐ Other: _____	
	☐ Boring ☐ Familiar ☐ Interesting ☐ Of quality ☐ Colorful ☐ Pleasurable ☐ Aesthetic ☐ Immersive ☐ Other: _____	
	☐ Boring ☐ Familiar ☐ Interesting ☐ Of quality ☐ Colorful ☐ Pleasurable ☐ Aesthetic ☐ Immersive ☐ Other: _____	
	☐ Boring ☐ Familiar ☐ Interesting ☐ Of quality ☐ Colorful ☐ Pleasurable ☐ Aesthetic ☐ Immersive ☐ Other: _____	
	☐ Boring ☐ Familiar ☐ Interesting ☐ Of quality ☐ Colorful ☐ Pleasurable ☐ Aesthetic ☐ Immersive ☐ Other: _____	

Select the right continuation of the following sentence "Prior to the test I …:

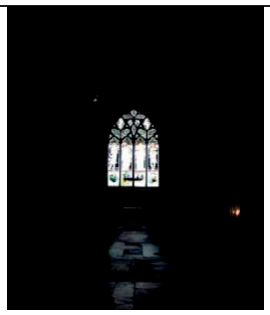
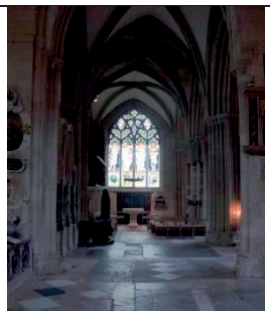
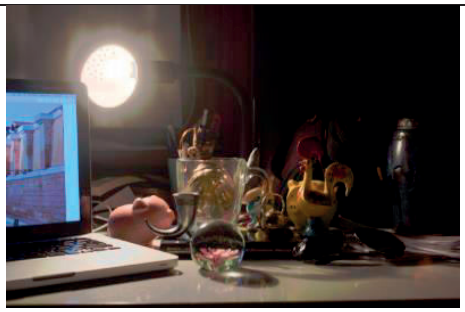
☐ had never used VR devices"
☐ had used a VR device a few times" (less than 5)
☐ had used a VR device several times" (more than 5)
☐ had used a VR device on a regular basis"
☐ Other: _____

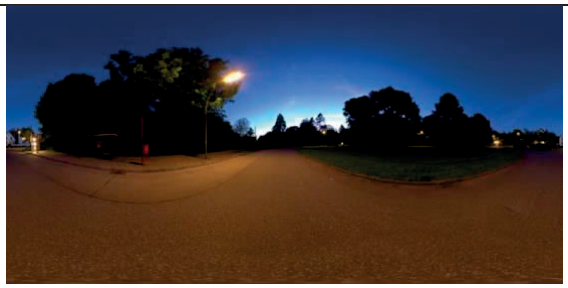
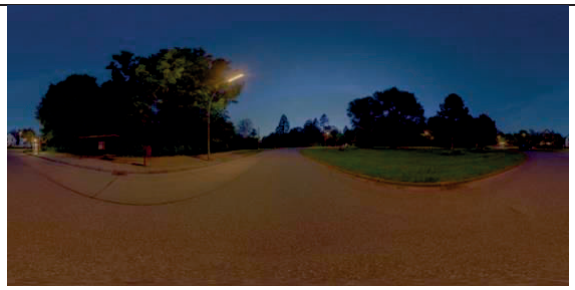
While visualizing the contents, did you experience?

Criterion	Strongly disagree	Disagree	Undecided	Agree	Strongly agree
Immersion					
Isolation					
Valence - Enjoyment					
Arousal – Excitation					
Enjoyment due to the novelty of the visualization					

Explain your choices above while describing your VR experience in a few words:

*Examples:

<table>
<tr><td colspan="2">Details in the bright areas</td></tr>
<tr><td> 
Original</td><td> 
Details in the bright areas</td></tr>
<tr><td colspan="2">Details in the dark areas</td></tr>
<tr><td> 
Original</td><td> 
Details in the dark areas</td></tr>
<tr><td colspan="2">Unnatural Colors</td></tr>
<tr><td> 
Original</td><td> 
Unnatural Colors</td></tr>
<tr><td colspan="2">Ghosting</td></tr>
<tr><td> 
Original</td><td> 
Ghosting</td></tr>
</table>

Noise	
	
Original	Noise (especially in the ceiling)
Overall Brightness	
	
Original	Low overall brightness
Overall Contrast	
	
Original	Poorly contrasted image
Naturalness of the scene	
	
Original	Less natural content

Definitions

General discomfort: Feeling of being uncomfortable physically and not feeling well (malaise).

Fatigue: Tiredness.

Headache: A pain you feel inside your head.

Eyestrain: Tired or painful eyes as a result of too much reading, looking at a computer screen, etc.

Difficulty focusing: Difficulties to adapt your clear vision when moving your focus from far objects to close objects and vice versa.

Increased salivation: Increase in the act of producing saliva in the mouth.

Sweating: Cold sweating is sweating in the absence of thermal stimulus.

Nausea: The feeling that you are going to vomit. It begins with stomach awareness, progresses to queasiness and then vomiting.

Difficulty concentrating: Difficulty to direct your attention or your efforts towards a particular activity, subject, or problem.

"Fullness of the Head": Having the impression that your head is full as what happens when all your blood is in your head (when your head is upside down).

Blurred vision: When looking around everything is unclear.

Dizziness with eyes open: The feeling as if everything is turning around, and that you are not able to balance and may fall down, when you have your eyes open.

Dizziness with eyes closed: The feeling as if everything is turning around, and that you are not able to balance and may fall down, when you have your eyes closed.

Vertigo: The sensation of moving around in space, or objects moving around a person. It is a disturbance of equilibrium.

Stomach awareness: Stomach awareness is usually used to indicate a feeling of discomfort which is just short of nausea.

Burping: Voluntary or involuntary release of gas from stomach through the mouth.

Bibliography

[Abbas 2017] Adeel Abbas and David Newman. *A novel projection for omni-directional video.* In Applications of Digital Image Processing XL, volume 10396, page 103960V. International Society for Optics and Photonics, 2017. (Cited on page 42.)

[Abromavičius 2017] Vytautas Abromavičius, Aurimas Gedminas and Artūras Serackis. *Detecting sense of presence changes in EEG spectrum during perception of immersive audiovisual content.* In Electrical, Electronic and Information Sciences (eStream), 2017 Open Conference of, pages 1–4. IEEE, 2017. (Cited on pages 245, 252 and 268.)

[Acqualagna 2015] Laura Acqualagna, Sebastian Bosse, Anne K Porbadnigk, Gabriel Curio, Klaus-Robert Müller, Thomas Wiegand and Benjamin Blankertz. *EEG-based classification of video quality perception using steady state visual evoked potentials (SSVEPs).* Journal of neural engineering, vol. 12, no. 2, page 026012, 2015. (Cited on page 241.)

[Akyüz 2013] Ahmet Oğuz Akyüz and Aslı Gençtav. *A reality check for radiometric camera response recovery algorithms.* Computers & Graphics, vol. 37, no. 7, pages 935–943, 2013. (Cited on page 108.)

[Alreshoodi 2013] Mohammed Alreshoodi and John Woods. *Survey on QoE\QoS correlation models for multimedia services.* arXiv preprint arXiv:1306.0221, 2013. (Cited on page 46.)

[Amara 1960] Roy Amara. *We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run.* `https://web.archive.org/web/20180410135130/https://spotlessdata.com/blog/amaras-law`, 1960. [Online; accessed 30-Sep-2018]. (Cited on page 159.)

[Amendment 2016] ITU Amendment. *5: New definitions for inclusion in recommendation: ITU-T P. 10/G. 100*, July 2016. (Cited on pages 47 and 255.)

[Aminghafari 2006] Mina Aminghafari, Nathalie Cheze and Jean-Michel Poggi. *Multivariate denoising using wavelets and principal component analysis.* Computational Statistics & Data Analysis, vol. 50, no. 9, pages 2381 – 2398, 2006. (Cited on pages 207 and 228.)

[Antons 2012] Jan-Niklas Antons, Robert Schleicher, Sebastian Arndt, Sebastian Moller, Anne K Porbadnigk and Gabriel Curio. *Analyzing speech quality perception using electroencephalography.* IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 6, pages 721–731, 2012. (Cited on page 210.)

[Antons 2013] Jan-Niklas Antons, Sebastian Arndt, Robert Schleicher, Sebastian Möller, Douglas O'Shaughnessy, Tiago H Falk*et al.* *Cognitive, affective, and experience correlates of speech quality perception in complex listening conditions.* In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 3672–3676. IEEE, 2013. (Cited on page 210.)

[Armstrong 2014] Richard A Armstrong. *When to use the Bonferroni correction.* Ophthalmic and Physiological Optics, vol. 34, no. 5, pages 502–508, 2014. (Cited on pages 64 and 94.)

[Arndt 2013] Sebastian Arndt, Jan-Niklas Antons, Rishabh Gupta, Robert Schleicher, Sebastian Möller, Tiago H Falk*et al.* *Subjective quality ratings and physiological correlates of synthesized speech.* In Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on, pages 152–157. IEEE, 2013. (Cited on page 210.)

[Arndt 2014] Sebastian Arndt, Jan-Niklas Antons, Robert Schleicher, Sebastian Moller and Gabriel Curio. *Using electroencephalography to measure perceived video quality.* IEEE Journal of Selected Topics in Signal Processing, vol. 8, no. 3, pages 366–376, 2014. (Cited on page 210.)

[Arndt 2016] Sebastian Arndt, Kjell Brunnström, Eva Cheng, Ulrich Engelke, Sebastian Möller and Jan-Niklas Antons. *Review on using physiology in quality of experience.* Electronic Imaging, vol. 2016, no. 16, pages 1–9, 2016. (Cited on pages 192, 194, 203, 209, 210 and 211.)

[Artusi 2011] Alessandro Artusi, Francesco Banterle, Kurt Debattista and Alan Chalmers. Advanced high dynamic range imaging: theory and practice. AK Peters/CRC Press, 2011. (Cited on page 41.)

[Artusi 2016] Alessandro Artusi, Rafal K Mantiuk, Thomas Richter, Pavel Korshunov, Philippe Hanhart, Touradj Ebrahimi and Massimiliano Agostinelli. *JPEG XT: A Compression Standard for HDR and WCG Images [Standards in a Nutshell].* IEEE Signal Processing Magazine, vol. 33, no. 2, pages 118–124, 2016. (Cited on page 38.)

[Ary 2018] Donald Ary, Lucy Cheser Jacobs, Christine K Sorensen Irvine and David Walker. Introduction to research in education. Cengage Learning, 2018. (Cited on page 149.)

[ATS 2015] ATSC TG3/S34-1. Dolby proposal in response to the atsc 3.0 hdr call for contributions using adi technology, September 2015. (Cited on page 75.)

[Azimi 2015] Maryam Azimi, Ronan Boitard, Basak Oztas, Stelios Ploumis, Hamid Reza Tohidypour, Mahsa T Pourazad and Panos Nasiopoulos. *Compression efficiency of HDR/LDR content.* In Quality of Multimedia Experi-

ence (QoMEX), 2015 Seventh International Workshop on, pages 1–6. IEEE, 2015. (Cited on page 40.)

[Aziz 2015] Nur Sukinah Aziz and Adzhar Kamaludin. *Using pre-test to validate the Questionnaire for Website Usability (QWU).* In Software Engineering and Computer Systems (ICSECS), 2015 4th International Conference on, pages 107–111. IEEE, 2015. (Cited on page 144.)

[Bang 2014] Jae Won Bang, Hwan Heo, Jong-Suk Choi and Kang Ryoung Park. *Assessment of eye fatigue caused by 3D displays based on multimodal measurements.* Sensors, vol. 14, no. 9, pages 16467–16485, 2014. (Cited on pages 193 and 201.)

[Baños 2004] Rosa María Baños, Cristina Botella, Mariano Alcañiz, Víctor Liaño, Belén Guerrero and Beatriz Rey. *Immersion and emotion: their impact on the sense of presence.* CyberPsychology & Behavior, vol. 7, no. 6, pages 734–741, 2004. (Cited on page 49.)

[Barten 1999] Peter GJ Barten. Contrast sensitivity of the human eye and its effects on image quality, volume 19. Spie optical engineering press Bellingham, WA, 1999. (Cited on page 35.)

[Bartko 1966] John J Bartko. *The intraclass correlation coefficient as a measure of reliability.* Psychological reports, vol. 19, no. 1, pages 3–11, 1966. (Cited on page 144.)

[Başar 2001] Erol Başar, Canan Başar-Eroglu, Sirel Karakaş and Martin Schürmann. *Gamma, alpha, delta, and theta oscillations govern cognitive processes.* International journal of psychophysiology, vol. 39, no. 2-3, pages 241–248, 2001. (Cited on pages 203 and 204.)

[Bassett 2006] Danielle Smith Bassett and ED Bullmore. *Small-world brain networks.* The neuroscientist, vol. 12, no. 6, pages 512–523, 2006. (Cited on page 205.)

[Baumann 2015] Olie Baumann. *The Interaction between Transfer Function and Compression in High Dynamic Range Video.* In Annual Technical Conference and Exhibition, SMPTE 2015, pages 1–13. SMPTE, 2015. (Cited on page 40.)

[Beatty 1982] Jackson Beatty. *Task-evoked pupillary responses, processing load, and the structure of processing resources.* Psychological bulletin, vol. 91, no. 2, page 276, 1982. (Cited on page 193.)

[Belardinelli 2004] Marta Olivetti Belardinelli, Carlo Sestieri, Rosalia Di Matteo, Franco Delogu, Cosimo Del Gratta, Antonio Ferretti, Massimo Caulo, Armando Tartaro and Gian Luca Romani. *Audio-visual crossmodal interactions*

in environmental perception: an fMRI investigation. Cognitive Processing, vol. 5, no. 3, pages 167–174, 2004. (Cited on page 210.)

[Berntson 1997] Gary G Berntson, J Thomas Bigger, Dwain L Eckberg, Paul Grossman, Peter G Kaufmann, Marek Malik, Haikady N Nagaraja, Stephen W Porges, J Philip Saul, Peter H Stone*et al.* *Heart rate variability: origins, methods, and interpretive caveats.* Psychophysiology, vol. 34, no. 6, pages 623–648, 1997. (Cited on page 207.)

[Beyer 2014] Justus Beyer, Viktor Miruchna and Sebastian Möller. *Assessing the impact of display size, game type, and usage context on mobile gaming QoE.* In Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on, pages 69–70. IEEE, 2014. (Cited on page 167.)

[Bilchick 2006] Kenneth C Bilchick and Ronald D Berger. *Heart rate variability.* Journal of cardiovascular electrophysiology, vol. 17, no. 6, page 691, 2006. (Cited on page 207.)

[Blythe 1999] Jim Blythe. *Innovativeness and newness in high-tech consumer durables.* Journal of Product & Brand Management, vol. 8, no. 5, pages 415–429, 1999. (Cited on pages 149, 150 and 151.)

[Blythe 2018] Mark Blythe and Andrew Monk, editors. Funology 2 - from usability to enjoyment, second edition. Human-Computer Interaction Series. Springer, 2018. (Cited on page 18.)

[Boitard 2012] Ronan Boitard, Kadi Bouatouch, Remi Cozot, Dominique Thoreau and Adrien Gruson. *Temporal Coherency for Video Tone Mapping.* In SPIE Optical Engineering+ Applications, pages 84990D–84990D. International Society for Optics and Photonics, 2012. (Cited on page 114.)

[Bonmatí Coll 2016] Ester Bonmatí Coll*et al.* *Study of brain complexity using information theory tools.* 2016. (Cited on page 197.)

[Boone 1995] Marinus M Boone, Edwin NG Verheijen and Peter F Van Tol. *Spatial sound-field reproduction by wave-field synthesis.* Journal of the Audio Engineering Society, vol. 43, no. 12, pages 1003–1012, 1995. (Cited on page 27.)

[Borer 2015a] T Borer and A Cotton. *ARIB STD-B67-Essential Parameter Values for the Extended Image Dynamic Range Television System for Programme Production.* ARIB Standard. Association of Radio Industries and Businesses, 2015. (Cited on page 36.)

[Borer 2015b] Tim Borer and Andrew Cotton. *A "display independent" high dynamic range television system.* 2015. (Cited on pages 37 and 148.)

[Borer 2016] Tim Borer and Andrew Cotton. *A Display-Independent High Dynamic Range Television System.* SMPTE Motion Imaging Journal, vol. 125, no. 4, pages 50–56, 2016. (Cited on page 36.)

[Bosse 2016] Sebastian Bosse, Klaus-Robert Müller, Thomas Wiegand and Wojciech Samek. *Brain-computer interfacing for multimedia quality assessment.* In Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on, pages 002834–002839. IEEE, 2016. (Cited on pages 192, 195 and 211.)

[Bouchard 2004] Stéphane Bouchard, Geneviéve Robillard, Julie St-Jacques, Stéphanie Dumoulin, Marie-Josée Patry and Patrice Renaud. *Reliability and validity of a single-item measure of presence in VR.* In Haptic, Audio and Visual Environments and Their Applications, 2004. HAVE 2004. Proceedings. The 3rd IEEE International Workshop on, pages 59–61. IEEE, 2004. (Cited on page 49.)

[Bracht 1968] Glenn H Bracht and Gene V Glass. *The external validity of experiments.* American educational research journal, vol. 5, no. 4, pages 437–474, 1968. (Cited on page 149.)

[Bradley 1994] Margaret M Bradley and Peter J Lang. *Measuring emotion: the self-assessment manikin and the semantic differential.* Journal of behavior therapy and experimental psychiatry, vol. 25, no. 1, pages 49–59, 1994. (Cited on page 126.)

[Brunnström 2013] Kjell Brunnström, Sergio Ariel Beker, Katrien De Moor, Ann Dooms, Sebastian Egger, Marie-Neige Garcia, Tobias Hossfeld, Satu Jumisko-Pyykkö, Christian Keimel, Mohamed-Chaker Larabi*et al. Qualinet white paper on definitions of quality of experience.* 2013. (Cited on pages 17, 18, 19, 47, 173, 176 and 255.)

[Capraş 2017] Roxana-Denisa Capraş, Tudor C Drugan and Sorana D Bolboacă. *Development and validation of a questionnaire to assess evidence based-practice.* In E-Health and Bioengineering Conference (EHB), 2017, pages 129–132. IEEE, 2017. (Cited on page 144.)

[Cassani 2018] Raymundo Cassani, Marc-Antoine Moinnereau and Tiago H Falk. *A Neurophysiological Sensor-Equipped Head-Mounted Display for Instrumental QoE Assessment of Immersive Multimedia.* In 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX), pages 1–6. IEEE, 2018. (Cited on page 269.)

[Chadha 2010] Monima Chadha. *Perceptual experience and concepts in classical indian philosophy.* 2010. (Cited on page 17.)

[Chanel 2006] Guillaume Chanel, Julien Kronegg, Didier Grandjean and Thierry Pun. Emotion assessment: Arousal evaluation using eeg's and peripheral physiological signals, pages 530–537. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. (Cited on page 65.)

[Chanel 2009] Guillaume Chanel. *Emotion assessment for affective computing based on brain and peripheral signals*. PhD thesis, University of Geneva, 2009. (Cited on page 207.)

[Chen 2015] Yanjiao Chen, Kaishun Wu and Qian Zhang. *From QoS to QoE: A tutorial on video quality assessment*. IEEE Communications Surveys & Tutorials, vol. 17, no. 2, pages 1126–1165, 2015. (Cited on pages 45, 46 and 48.)

[Chinnock 2016] Chris Chinnock. *Dolby Vision and HDR10*. White Paper of Insight Media, 2016. (Cited on page 39.)

[Cobb 1999] Sue VG Cobb, Sarah Nichols, Amanda Ramsey and John R Wilson. *Virtual reality-induced symptoms and effects (VRISE)*. Presence: teleoperators and virtual environments, vol. 8, no. 2, pages 169–186, 1999. (Cited on page 133.)

[Craster 1929] JEE Craster. *Some equal-area projections of the sphere*. The Geographical Journal, vol. 74, no. 5, pages 471–474, 1929. (Cited on page 117.)

[Croft 2000] Rodney J Croft and Robert J Barry. *Removal of ocular artifact from the EEG: a review*. Neurophysiologie Clinique/Clinical Neurophysiology, vol. 30, no. 1, pages 5–19, 2000. (Cited on page 201.)

[Cronbach 1951] Lee J. Cronbach. *Coefficient alpha and the internal structure of tests*. Psychometrika, vol. 16, no. 3, pages 297–334, Sep 1951. (Cited on page 144.)

[Csikszentmihalyi 1999] Mihaly Csikszentmihalyi. *If we are so rich, why aren't we happy?* American psychologist, vol. 54, no. 10, page 821, 1999. (Cited on page 47.)

[De Abreu 2017] Ana De Abreu, Cagri Ozcinar and Aljosa Smolic. *Look around you: Saliency maps for omnidirectional images in vr applications*. In Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on, pages 1–6. IEEE, 2017. (Cited on page 136.)

[de Borda 1781] Jean C de Borda. *Mémoire sur les élections au scrutin*. 1781. (Cited on page 130.)

[Debevec 1998] Paul Debevec and Jitendra Malik. *Recovering High Dynamic Range Radiance Maps from Photographs*. In Proceedings of the 24th annual conference on Computer graphics and interactive techniques (SIGGRAPH '97), pages 369–378. ACM Press/Addison-Wesley Publishing Co., 1998. (Cited on page 108.)

[Debevec 2008] Paul Debevec. *Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography*. In ACM SIGGRAPH 2008 classes, page 32. ACM, 2008. (Cited on page 106.)

[Devlin 2002] Kate Devlin. *A review of tone reproduction techniques*. Computer Science, University of Bristol, Tech. Rep. CSTR-02-005, 2002. (Cited on page 37.)

[Diaz 2016] Raul Diaz, Sam Blinstein and Sheng Qu. *Integrating HEVC Video Compression with a High Dynamic Range Video Pipeline*. SMPTE Motion Imaging Journal, vol. 125, no. 1, pages 14–21, 2016. (Cited on pages 38 and 89.)

[Dolby 2016] Dolby. *Dolby Vision ^{TM} for the home*. White paper, 2016. (Cited on page 36.)

[Dufaux 2009] Frédéric Dufaux, Gary J Sullivan and Touradj Ebrahimi. *The JPEG XR image coding standard [Standards in a Nutshell]*. IEEE Signal Processing Magazine, vol. 26, no. 6, 2009. (Cited on page 38.)

[Ebner 1998a] Fritz Ebner. *Derivation and modelling hue uniformity and development of the ipt color space*. 1998. (Cited on page 31.)

[Ebner 1998b] Fritz Ebner and Mark D Fairchild. *Development and testing of a color space (IPT) with improved hue uniformity*. In Color and Imaging Conference, volume 1998, pages 8–13. Society for Imaging Science and Technology, 1998. (Cited on page 31.)

[Edge 2013] C.J. Edge. *Gamut mapping using hue-preserving color space*, may 2013. US Patent 8,437,053 ; accessed 31 October 2016. (Cited on page 73.)

[Egan 2016] Darragh Egan, Sean Brennan, John Barrett, Yuansong Qiao, Christian Timmerer and Niall Murray. *An evaluation of Heart Rate and ElectroDermal Activity as an objective QoE evaluation method for immersive virtual reality environments*. In Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on, pages 1–6. IEEE, 2016. (Cited on pages 166 and 210.)

[Eilertsen 2013] Gabriel Eilertsen, Robert Wanat, Rafał K Mantiuk and Jonas Unger. *Evaluation of Tone Mapping Operators for HDR-Video*. In Computer Graphics Forum, volume 32, pages 275–284. Wiley Online Library, 2013. (Cited on page 118.)

[Eilertsen 2017] Gabriel Eilertsen, Rafal Konrad Mantiuk and Jonas Unger. *A comparative review of tone-mapping algorithms for high dynamic range video*. In Computer Graphics Forum, volume 36, pages 565–592. Wiley Online Library, 2017. (Cited on page 37.)

[embedded vision alliance 2016] embedded vision alliance. *Consumer Virtual Reality Head-Mounted Display Shipments to Reach 130 Million Worldwide by 2021*. https://www.embedded-vision.com/industry-analysis/market-analysis/

consumer-virtual-reality-head-mounted-display-shipments-reach-130-, 2016. [Online; accessed 11-May-2018]. (Cited on page 43.)

[Engelke 2016] Ulrich Engelke, Daniel P Darcy, Grant H Mulliken, Sebastian Bosse, Maria G Martini, Sebastian Arndt, Jan-Niklas Antons, Kit Yan Chan, Naeem Ramzan and Kjell Brunnstrom. *Psychophysiology-Based QoE Assessment: A Survey.* IEEE Journal of Selected Topics in Signal Processing, 2016. (Cited on page 64.)

[Engelke 2017] Ulrich Engelke, Daniel P Darcy, Grant H Mulliken, Sebastian Bosse, Maria G Martini, Sebastian Arndt, Jan-Niklas Antons, Kit Yan Chan, Naeem Ramzan and Kjell Brunnström. *Psychophysiology-based QoE assessment: a survey.* IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 1, pages 6–21, 2017. (Cited on pages 192, 193, 194, 195, 200, 203, 209 and 253.)

[ETS 2017] ETSI GS CCM 001 V1.1.1. Compound content management specification, February 2017. (Cited on pages 39 and 88.)

[Fei 2006] Jin Fei and Ioannis Pavlidis. *Analysis of breathing air flow patterns in thermal imaging.* In Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE, pages 946–952. IEEE, 2006. (Cited on page 200.)

[Field 2013] Andy Field. Discovering statistics using ibm spss statistics. Sage, 2013. (Cited on page 177.)

[Fingelkurts 2005] Andrew A Fingelkurts, Alexander A Fingelkurts and Seppo Kähkönen. *Functional connectivity in the brain - is it an elusive concept?* Neuroscience & Biobehavioral Reviews, vol. 28, no. 8, pages 827–836, 2005. (Cited on page 228.)

[Flynn 2016] David Flynn, Detlev Marpe, Matteo Naccari, Tung Nguyen, Chris Rosewarne, Karl Sharman, Joel Sole and Jizheng Xu. *Overview of the range extensions for the HEVC standard: tools, profiles, and performance.* IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, no. 1, pages 4–19, 2016. (Cited on page 38.)

[Forgas 1999] Joseph P Forgas. *On feeling good and being rude: Affective influences on language use and request formulations.* Journal of Personality and Social Psychology, vol. 76, no. 6, page 928, 1999. (Cited on page 192.)

[Freeman 1999] Jonathan Freeman, Steve E Avons, Don E Pearson and Wijnand A IJsselsteijn. *Effects of sensory information and prior experience on direct subjective ratings of presence.* Presence: Teleoperators & Virtual Environments, vol. 8, no. 1, pages 1–13, 1999. (Cited on page 49.)

[Friston 2011] Karl J Friston. *Functional and effective connectivity: a review.* Brain connectivity, vol. 1, no. 1, pages 13–36, 2011. (Cited on pages 204 and 205.)

[Gardner 1999] William G Gardner. *3D audio and acoustic environment modeling.* Wave Arts, Inc.–1999.–109 p, 1999. (Cited on page 27.)

[Geng 2010] Xiao Geng. *Cultural differences influence on language.* Review of European Studies, vol. 2, no. 2, page 219, 2010. (Cited on page 192.)

[Ghinea 2008] Gheorghita Ghinea and Sherry Y Chen. *Measuring quality of perception in distributed multimedia: Verbalizers vs. imagers.* Computers in Human Behavior, vol. 24, no. 4, pages 1317–1329, 2008. (Cited on page 20.)

[Ghinea 2014] Gheorghita Ghinea, Christian Timmerer, Weisi Lin and Stephen R Gulliver. *Mulsemedia: State of the art, perspectives, and challenges.* ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 11, no. 1s, page 17, 2014. (Cited on page 42.)

[Gravetter 2018] Frederick J Gravetter and Lori-Ann B Forzano. Research methods for the behavioral sciences. Cengage Learning, 2018. (Cited on pages 149, 150 and 176.)

[Gupta 2015] Rishabh Gupta, Hubert J Banville and Tiago H Falk. *PhySyQX: A database for physiological evaluation of synthesised speech quality-of-experience.* In Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on, pages 1–5. IEEE, 2015. (Cited on page 211.)

[Hanhart 2016] Philippe Hanhart, Martin Řeřábek and Touradj Ebrahimi. *Subjective and objective evaluation of HDR video coding technologies.* In Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on, pages 1–6. Ieee, 2016. (Cited on pages 40 and 45.)

[Hastie 2001] Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The elements of statistical learning.* 2001. (Cited on pages 208 and 209.)

[Hatchett 2018] Jonathan Hatchett, Kurt Debattista, Ratnajit Mukherjee, Thomas Bashford-Rogers and Alan Chalmers. *An evaluation of power transfer functions for HDR video compression.* The Visual Computer, vol. 34, no. 2, pages 167–176, 2018. (Cited on page 37.)

[Hausner 2004] Alexander Hausner and Marc Stamminger. *Tone Mapping in VR Environments.* In M. Alexa and E. Galin, editors, Eurographics 2004 - Short Presentations. Eurographics Association, 2004. (Cited on page 109.)

[Haynes 2013] Winston Haynes. Student's t-test, pages 2023–2025. Springer New York, 2013. (Cited on page 63.)

[He 2004] Ping He, G Wilson and C Russell. *Removal of ocular artifacts from electro-encephalogram by adaptive filtering.* Medical and biological engineering and computing, vol. 42, no. 3, pages 407–412, 2004. (Cited on page 201.)

[Herre 2015] Jürgen Herre, Johannes Hilpert, Achim Kuntz and Jan Plogsties. *MPEG-H audio the new standard for universal spatial/3D audio coding.* Journal of the Audio Engineering Society, vol. 62, no. 12, pages 821–830, 2015. (Cited on page 27.)

[Hewage 2013] Chaminda TER Hewage and Maria G Martini. *Quality of experience for 3D video streaming.* IEEE Communications Magazine, vol. 51, no. 5, pages 101–107, 2013. (Cited on page 48.)

[Higgs 2005] Bronwyn Higgs, Michael Jay Polonsky and Mary Hollick. *Measuring expectations: forecast vs. ideal expectations. Does it really matter?* Journal of Retailing and Consumer Services, vol. 12, no. 1, pages 49–64, 2005. (Cited on pages 161 and 163.)

[Houghton 2005] Tony Houghton. *Ubiquitous services and applications: Focus on what customers think not what they say.* Proceedings of EURESCOM 2005, pages 11–17, 2005. (Cited on page 162.)

[Hua 2004] Jianping Hua, Zixiang Xiong, James Lowey, Edward Suh and Edward R Dougherty. *Optimal number of features as a function of sample size for various classification rules.* Bioinformatics, vol. 21, no. 8, pages 1509–1515, 2004. (Cited on page 208.)

[Hulsebos 2002] Edo Hulsebos, Diemer de Vries and Emmanuelle Bourdillat. *Improved microphone array configurations for auralization of sound fields by wave-field synthesis.* Journal of the Audio Engineering Society, vol. 50, no. 10, pages 779–790, 2002. (Cited on page 27.)

[Hupont 2015] Isabelle Hupont, Joaquin Gracia, Luis Sanagustin and Miguel Angel Gracia. *How do new visual immersive systems influence gaming QoE? A use case of serious gaming with Oculus Rift.* In Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on, pages 1–6. IEEE, 2015. (Cited on page 166.)

[Hwang 2006] Jane Hwang, Jaehoon Jung and Gerard Jounghyun Kim. *Hand-held virtual reality: a feasibility study.* In Proceedings of the ACM symposium on Virtual reality software and technology, pages 356–363. ACM, 2006. (Cited on page 153.)

[Ibraheem 2012] Noor A Ibraheem, Mokhtar M Hasan, Rafiqul Z Khan and Pramod K Mishra. *Understanding color models: a review.* ARPN Journal of science and technology, vol. 2, no. 3, pages 265–275, 2012. (Cited on page 31.)

[IJsselsteijn 2008] WA IJsselsteijn, YAW De Kort and Karolien Poels. *The game experience questionnaire*. Manuscript in preparation, 2008. (Cited on page 153.)

[IJsselsteijn 2013] WA IJsselsteijn, YAW De Kort and K Poels. *The Game Experience Questionnaire: Development of a self-report measure to assess the psychological impact of digital games. Manuscript in Preparation*. 2013. (Cited on pages 164 and 165.)

[Ikeda 2003] Sei Ikeda, Tomokazu Sato and Naokazu Yokoya. *High-resolution panoramic movie generation from video streams acquired by an omnidirectional multi-camera system*. In Multisensor Fusion and Integration for Intelligent Systems, MFI2003. Proceedings of IEEE International Conference on, pages 155–160. IEEE, 2003. (Cited on page 42.)

[Insko 2003] Brent E Insko. *Measuring presence: Subjective, behavioral and physiological methods*. 2003. (Cited on pages 192, 193, 194 and 195.)

[Ishihara 1960] Shinobu Ishihara. Tests for colour-blindness. Kanehara Shuppan Company, 1960. (Cited on page 58.)

[ISO 2016a] ISO/IEC JTC1/SC29/WG11. Description of the reshaper parameters derivation process in etm reference software, San Diego, February, February 2016. Doc. W0031. (Cited on page 75.)

[ISO 2016b] ISO/IEC JTC1/SC29/WG11. Overview of ictcp, San Diego, February, February 2016. Doc. W0050. (Cited on page 74.)

[ISO 2016c] ISO/IEC JTC1/SC29/WG11. Verification test plan for hdr/wcg coding using hevc main 10 profile, San Diego, February 2016. Doc. W0018. (Cited on page 76.)

[ITU-T Recommendation P.911 1998] ITU-T Recommendation P.911. *Subjective Audiovisual Quality Assessment Methods for Multimedia Applications. International Telecommunication*, 1998. (Cited on page 168.)

[ITU 1994] ITU-R BT.814-1. Specifications and alignment procedures for setting of brightness and contrast of displays., 1992-1994. (Cited on page 53.)

[ITU 1996] ITU-T P.800. Methods for subjective determination of transmission quality, August 1996. (Cited on page 56.)

[ITU 2005] ITU-R BT.1729. Common 16:9 or 4:3 aspect ratio digital television reference test pattern., January 2005. (Cited on page 53.)

[ITU 2007a] ITU-R BT.1886. Reference electro-optical transfer function for flat panel displays used in HDTV studio production., March 2007. (Cited on pages 53 and 109.)

[ITU 2007b] ITU-T P.910. P. 10/g. 100 amendment 1, new appendix i–definition of quality of experience (qoe), January 2007. (Cited on page 47.)

[ITU 2008a] ITU-R BT.1201-1. Extremely high resolution imagery., June 2008. (Cited on page 28.)

[ITU 2008b] ITU-R BT.1769. Parameter values for an expanded hierarchy of lsdi image formats for production and international programme exchange., June 2008. (Cited on page 222.)

[ITU 2008c] ITU-R E.800. Overall network operation, telephone service, service operation and human factors., September 2008. (Cited on pages 44 and 46.)

[ITU 2008d] ITU-T P.910. Subjective video quality assessment methods for multimedia applications, April 2008. (Cited on pages 51, 53, 54, 56, 78 and 168.)

[ITU 2010] ITU-R BT.1845. Guidelines on metrics to be used when tailoring television programmes to broadcasting applications at various image quality levels, display sizes and aspect ratios., March 2010. (Cited on page 52.)

[ITU 2011a] ITU-R BT.1788. Methodology for the subjective assessment of video quality in multimedia applications., 2011. (Cited on pages 57, 58 and 76.)

[ITU 2011b] ITU-R BT.601-7. Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratio., March 2011. (Cited on pages 28, 32 and 53.)

[ITU 2012a] ITU-R BS.775-3. Multi-channel stereophonic sound system with or without accompanying picture, 2012. (Cited on pages 27 and 220.)

[ITU 2012b] ITU-R BT.2022-0. General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays., August 2012. (Cited on pages 51, 52, 53, 72 and 233.)

[ITU 2012c] ITU-R BT.500-13. Methodology for the subjective assessment of the quality of television pictures., January 2012. (Cited on pages 51, 52, 54, 56, 58, 60, 61, 79, 92, 120, 129 and 235.)

[ITU 2012d] ITU-T P.1401. Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models., August 2012. (Cited on pages 61 and 62.)

[ITU 2013a] ITU-R BT.2035-0. A reference viewing environment for evaluation of hdtv program material or completed programmes., July 2013. (Cited on pages 51, 52, 53, 109 and 118.)

[ITU 2013b] ITU-T COM.12-039. Investigating the subjective judgment process using physiological data., 2013. (Cited on page 211.)

[ITU 2013c] ITU-T COM.12-103. Using physiological data for assessing subjective video quality ratings., 2013. (Cited on page 211.)

[ITU 2013d] ITU-T COM.12-112. Using physiological data for assessing variations of the cognitive state evoked by quality profiles., 2013. (Cited on page 211.)

[ITU 2014a] ITU-R F.734. Definitions, requirements and use cases for telepresence systems., October 2014. (Cited on page 49.)

[ITU 2014b] ITU-T COM.12-202. Using physiological data for assessing the audiovisual quality of longer stimuli., 2014. (Cited on page 211.)

[ITU 2014c] ITU-T G.1031. Qoe factors in web-browsing., February 2014. (Cited on page 48.)

[ITU 2014d] ITU-T P.913. Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment, Januarry 2014. (Cited on pages 51, 52, 53, 54, 58, 59, 61 and 62.)

[ITU 2015a] ITU-R BT.2020-2. Parameter values for ultra-high definition television systems for production and international programme exchange., October 2015. (Cited on pages 32 and 87.)

[ITU 2015b] ITU-R BT.2021-1. Subjective methods for the assessment of stereoscopic 3DTV systems., February 2015. (Cited on pages 29, 52, 58 and 60.)

[ITU 2015c] ITU-R BT.2246-5. The present state of ultra-high definition television., July 2015. (Cited on pages 28, 35, 49 and 52.)

[ITU 2015d] ITU-R BT.2390-0. High dynamic range television for production and international programme exchange., October 2015. (Cited on pages 34, 35, 53, 73 and 74.)

[ITU 2015e] ITU-R BT.709-6. Parameter values for the HDTV standards for production and international programme exchange., June 2015. (Cited on pages 32 and 51.)

[ITU 2015f] ITU-R H.265. High efficiency video coding, April 2015. (Cited on page 38.)

[ITU 2016] ITU-R BT.2100-0. Image parameter values for high dynamic range television for use in production and international programme exchange., July 2016. (Cited on pages 35, 36, 37, 51, 72, 74 and 88.)

[Jang 2015] Eun-Hye Jang, Byoung-Jun Park, Mi-Sook Park, Sang-Hyeob Kim and Jin-Hun Sohn. *Analysis of physiological signals for recognition of boredom, pain, and surprise emotions.* Journal of physiological anthropology, vol. 34, no. 1, page 25, 2015. (Cited on page 207.)

[Jennett 2008] Charlene Jennett, Anna L Cox, Paul Cairns, Samira Dhoparee, Andrew Epps, Tim Tijs and Alison Walton. *Measuring and defining the experience of immersion in games*. International journal of human-computer studies, vol. 66, no. 9, pages 641–661, 2008. (Cited on pages 166, 173 and 180.)

[Jennett 2010] Charlene Ianthe Jennett. *Is game immersion just another form of selective attention? An empirical investigation of real world dissociation in computer game immersion*. PhD thesis, UCL (University College London), 2010. (Cited on pages 50, 164 and 165.)

[Jerritta 2011] S Jerritta, M Murugappan, R Nagarajan and Khairunizam Wan. *Physiological signals based human emotion recognition: a review*. In Signal Processing and its Applications (CSPA), 2011 IEEE 7th International Colloquium on, pages 410–415. IEEE, 2011. (Cited on page 211.)

[Johnson 2006] Chris Johnson. The practical zone system: for film and digital photography. Taylor & Francis, 2006. (Cited on page 109.)

[Johnson 2014] Daniel Johnson, Christopher Watling, John Gardner and Lennart E Nacke. *The edge of glory: the relationship between metacritic scores and player experience*. In Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play, pages 141–150. ACM, 2014. (Cited on pages 164 and 180.)

[Joyce 2004] Carrie A Joyce, Irina F Gorodnitsky and Marta Kutas. *Automatic removal of eye movement and blink artifacts from EEG data using blind component separation*. Psychophysiology, vol. 41, no. 2, pages 313–325, 2004. (Cited on page 201.)

[Kalawsky 2000] Roy S Kalawsky. *The validity of presence as a reliable human performance metric in immersive environments*. In proceedings of the Presence Workshop'00, 2000. (Cited on page 49.)

[Keimel 2011] Christian Keimel, Martin Rothbucher, Hao Shen and Klaus Diepold. *Video is a cube*. IEEE Signal Processing Magazine, vol. 28, no. 6, pages 41–49, 2011. (Cited on pages 45 and 47.)

[Kennedy 1993] Robert S Kennedy, Norman E Lane, Kevin S Berbaum and Michael G Lilienthal. *Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness*. The international journal of aviation psychology, vol. 3, no. 3, pages 203–220, 1993. (Cited on pages 125 and 128.)

[Ko 2011] Li-Wei Ko, Chun-Shu Wei, Tzyy-Ping Jung and Chin-Teng Lin. *Estimating the level of motion sickness based on EEG spectra*. In International Conference on Foundations of Augmented Cognition, pages 169–176. Springer, 2011. (Cited on page 173.)

[Kober 2012] Silvia Erika Kober and Christa Neuper. *Using auditory event-related EEG potentials to assess presence in virtual reality.* International Journal of Human-Computer Studies, vol. 70, no. 9, pages 577–587, 2012. (Cited on page 211.)

[Koelstra 2012] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt and Ioannis Patras. *Deap: A database for emotion analysis; using physiological signals.* IEEE Transactions on Affective Computing, vol. 3, no. 1, pages 18–31, 2012. (Cited on pages 207 and 211.)

[Kolasinski 1995] Eugenia M Kolasinski. *Simulator Sickness in Virtual Environments.* Technical report, DTIC Document, 1995. (Cited on pages 125, 126, 173 and 176.)

[Kroupi 2014a] Eleni Kroupi. *Emotion Detection and Recognition based on Brain and Peripheral Physiological Signals.* 2014. (Cited on pages 204, 207 and 228.)

[Kroupi 2014b] Eleni Kroupi, Philippe Hanhart, Jong-Seok Lee, Martin Rerabek and Touradj Ebrahimi. *EEG correlates during video quality perception.* In Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European, pages 2135–2139, 2014. (Cited on page 210.)

[Kroupi 2014c] Eleni Kroupi, Philippe Hanhart, Jong-Seok Lee, Martin Rerabek and Touradj Ebrahimi. *Predicting subjective sensation of reality during multimedia consumption based on EEG and peripheral physiological signals.* In Multimedia and Expo (ICME), 2014 IEEE International Conference on, pages 1–6. IEEE, 2014. (Cited on pages 199 and 210.)

[Kuang 2004] Jiangtao Kuang, Hiroshi Yamaguchi, Garrett M Johnson and Mark D Fairchild. *Testing HDR image rendering algorithms.* In Color and Imaging Conference, volume 2004, pages 315–320. Society for Imaging Science and Technology, 2004. (Cited on page 118.)

[Kurzweil 2010] Ray Kurzweil. The singularity is near. Gerald Duckworth & Co, 2010. (Cited on page 270.)

[Labs 2016a] Dolby Labs. *Subsampling in ICtCp vs Y'C'bC'r.* Technical report, Dolby, 2016. (Cited on page 32.)

[Labs 2016b] Dolby Labs. *White Paper on ICtCp.* Technical report, Dolby, 2016. version 7.2. (Cited on page 74.)

[Lassalle 2011] Julie Lassalle, Laetitia Gros and Gilles Coppin. *Combination of physiological and subjective measures to assess quality of experience for audiovisual technologies.* In Quality of Multimedia Experience (QoMEX), 2011

Third International Workshop on, pages 13–18. IEEE, 2011. (Cited on page 210.)

[Lasserre 2016] S Lasserre, E François, F Le Léannec and D Touzé. *Single-layer HDR video coding with SDR backward compatibility*. In SPIE Optical Engineering+ Applications, pages 997108–997108. International Society for Optics and Photonics, 2016. (Cited on page 38.)

[Latora 2001] Vito Latora and Massimo Marchiori. *Efficient behavior of small-world networks*. Physical review letters, vol. 87, no. 19, page 198701, 2001. (Cited on pages 205 and 228.)

[Lau 2015] Christie Pei-hang Lau. *The Relationship Between Subjective Questionnaire Scores and Metacritic Scores*. 2015. (Cited on pages 164 and 173.)

[Laugwitz 2008] Bettina Laugwitz, Theo Held and Martin Schrepp. *Construction and evaluation of a user experience questionnaire*. In Symposium of the Austrian HCI and Usability Engineering Group, pages 63–76. Springer, 2008. (Cited on pages 164 and 165.)

[LaViola Jr 2000] Joseph J LaViola Jr. *A discussion of cybersickness in virtual environments*. ACM SIGCHI Bulletin, vol. 32, no. 1, pages 47–56, 2000. (Cited on pages 125, 173 and 176.)

[Le Feuvre 2014] Jean Le Feuvre, JM Thiesse, Matthieu Parmentier, Mickaël Raulet and Christophe Daguet. *Ultra high definition HEVC DASH data set*. In Proceedings of the 5th ACM Multimedia Systems Conference, pages 7–12. ACM, 2014. (Cited on page 89.)

[Ledda 2004] Patrick Ledda, Luis Paulo Santos and Alan Chalmers. *A local model of eye adaptation for high dynamic range images*. In Proceedings of the 3rd international conference on Computer graphics, virtual reality, visualisation and interaction in Africa, pages 151–160. ACM, 2004. (Cited on pages 30 and 33.)

[Ledda 2005] Patrick Ledda, Alan Chalmers, Tom Troscianko and Helge Seetzen. *Evaluation of tone mapping operators using a high dynamic range display*. In ACM Transactions on Graphics (TOG), volume 24, pages 640–648. ACM, 2005. (Cited on page 118.)

[Lessiter 2001] Jane Lessiter, Jonathan Freeman, Edmund Keogh and Jules Davidoff. *A cross-media presence questionnaire: The ITC-Sense of Presence Inventory*. Presence: Teleoperators & Virtual Environments, vol. 10, no. 3, pages 282–297, 2001. (Cited on page 49.)

[Li 2009a] Congcong Li and Tsuhan Chen. *Aesthetic visual quality assessment of paintings*. IEEE Journal of selected topics in Signal Processing, vol. 3, no. 2, pages 236–252, 2009. (Cited on page 124.)

[Li 2009b] Mu Li and Bao-Liang Lu. *Emotion classification based on gamma-band EEG*. In Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE, pages 1223–1226. IEEE, 2009. (Cited on page 241.)

[Li 2014] Xiaoou Li, Xun Chen, Yuning Yan, Wenshi Wei and Z Jane Wang. *Classification of EEG signals using a multiple kernel learning support vector machine*. Sensors, vol. 14, no. 7, pages 12784–12802, 2014. (Cited on page 229.)

[Liao 2011] Wei Liao, Jurong Ding, Daniele Marinazzo, Qiang Xu, Zhengge Wang, Cuiping Yuan, Zhiqiang Zhang, Guangming Lu and Huafu Chen. *Small-world directed networks in the human brain: multivariate Granger causality analysis of resting-state fMRI*. Neuroimage, vol. 54, no. 4, pages 2683–2694, 2011. (Cited on page 205.)

[Lindemann 2011] Lea Lindemann and Marcus Magnor. *Assessing the quality of compressed images using EEG*. In Image Processing (ICIP), 2011 18th IEEE International Conference on, pages 3109–3112. IEEE, 2011. (Cited on page 210.)

[Linden 2003] Alexander Linden and Jackie Fenn. *Understanding Gartner's hype cycles*. Strategic Analysis Report N° R-20-1971. Gartner, Inc, 2003. (Cited on page 159.)

[Litwic 2016] Lukasz Litwic, Olie Baumann, Philip White and Matthew S Goldman. *Bit Rate Requirements for High Dynamic Range Video*. SMPTE Motion Imaging Journal, vol. 125, no. 5, pages 52–60, 2016. (Cited on pages 40, 72 and 92.)

[Liu 2016] Mingyang Liu, Di Fan, Xiaohan Zhang and Xiaopeng Gong. *Human emotion recognition based on galvanic skin response signal feature selection and svm*. In Smart City and Systems Engineering (ICSCSE), International Conference on, pages 157–160. IEEE, 2016. (Cited on page 229.)

[Love 2013] Adam Love, Andreas Kavazis, Alan Morse and KC Mayer. *Soccer-specific stadiums and attendance in major league soccer: Investigating the novelty effect*. Journal of Applied Sports Management, 2013. (Cited on page 149.)

[Lu 2016a] Taoran Lu, Fangjun Pu, Peng Yin, Tao Chen, Walt Husak, Jaclyn Pytlarz, Robin Atkins, Jan Fr-hlich and Guan-Ming Su. *ITP Colour Space and Its Compression Performance for High Dynamic Range and Wide Colour Gamut Video Distribution*. ZTE COMMUNICATIONS, vol. 14, pages 32–38, February 2016. (Cited on pages 73, 74 and 75.)

[Lu 2016b] Taoran Lu, Fangjun Pu, Peng Yin, Jaclyn Pytlarz, Tao Chen and Walter Husak. *Adaptive reshaper for high dynamic range and wide color gamut video*

compression. In SPIE Optical Engineering+ Applications, pages 99710B–99710B. International Society for Optics and Photonics, 2016. (Cited on pages 75 and 76.)

[Luthra 2015] Ajay Luthra, Edouard Francois and Walt Husak. *Call for evidence (CfE) for HDR and WCG video coding*. ISO/IEC JTC1/SC29/WG11 MPEG2014 N, vol. 15083, 2015. (Cited on page 40.)

[Maaoui 2008] C Maaoui and A Pruski. *A comparative study of SVM kernel applied to emotion recognition from physiological signals*. In Systems, Signals and Devices, 2008. IEEE SSD 2008. 5th International Multi-Conference on, pages 1–6. IEEE, 2008. (Cited on page 229.)

[MacAdam 1942] David L MacAdam. *Visual sensitivities to color differences in daylight*. JOSA, vol. 32, no. 5, pages 247–274, 1942. (Cited on page 74.)

[Mahmalat 2016] Samir Mahmalat, Nikolce Stefanoski, Daniel Luginbühl, Tunç Ozan Aydın and Aljosa Smolic. *Luminance independent chromaticity preprocessing for HDR video coding*. In Image Processing (ICIP), 2016 IEEE International Conference on, pages 1389–1393. IEEE, 2016. (Cited on page 32.)

[Mansilla 2013] Wendy Ann C Mansilla. *Quality of Aesthetic Experience and Implicit Modulating Factors*. PhD thesis, Norwegian University of Science and Technology (NTNU), 2013. (Cited on page 125.)

[Manski 2004] Charles F Manski. *Measuring expectations*. Econometrica, vol. 72, no. 5, pages 1329–1376, 2004. (Cited on pages 162, 191 and 248.)

[Mantiuk 2006a] Rafał Mantiuk, Alexander Efremov, Karol Myszkowski and Hans-Peter Seidel. *Backward compatible high dynamic range MPEG video compression*. In ACM Transactions on Graphics (TOG), volume 25, pages 713–723. ACM, 2006. (Cited on page 38.)

[Mantiuk 2006b] Rafal Mantiuk, Karol Myszkowski and Hans-Peter Seidel. *A perceptual framework for contrast processing of high dynamic range images*. ACM Transactions on Applied Perception (TAP), vol. 3, no. 3, pages 286–308, 2006. (Cited on page 110.)

[Mantiuk 2007] Rafał Mantiuk, Grzegorz Krawczyk, Radosław Mantiuk and Hans-Peter Seidel. *High Dynamic Range Imaging Pipeline: Perception-motivated Representation of Visual Content*. In Bernice E. Rogowitz, Thrasyvoulos N. Pappas and Scott J. Daly, editors, Human Vision and Electronic Imaging XII, volume 6492 of *Proceedings of SPIE*, San Jose, USA, February 2007. SPIE. (Cited on page 109.)

[Mantiuk 2008] Rafał Mantiuk, Scott Daly and Louis Kerofsky. *Display adaptive tone mapping*. In ACM Transactions on Graphics (TOG), volume 27, page 68. ACM, 2008. (Cited on page 109.)

[Mantiuk 2015] Rafał K Mantiuk, Karol Myszkowski and Hans-Peter Seidel. High dynamic range imaging. Wiley Online Library, 2015. (Cited on pages 34 and 41.)

[Marchesotti 2011] Luca Marchesotti, Florent Perronnin, Diane Larlus and Gabriela Csurka. *Assessing the aesthetic quality of photographs using generic image descriptors*. In Computer Vision (ICCV), 2011 IEEE International Conference on, pages 1784–1791. IEEE, 2011. (Cited on page 124.)

[Mauchly 1940] John W Mauchly. *Significance test for sphericity of a normal n-variate distribution*. The Annals of Mathematical Statistics, vol. 11, no. 2, pages 204–209, 1940. (Cited on page 177.)

[McQuiston 1989] Daniel H McQuiston. *Novelty, complexity, and importance as causal determinants of industrial buyer behavior*. The Journal of Marketing, pages 66–79, 1989. (Cited on pages 149 and 150.)

[Meehan 2002] Michael Meehan, Brent Insko, Mary Whitton and Frederick P Brooks Jr. *Physiological measures of presence in stressful virtual environments*. ACM Transactions on Graphics (TOG), vol. 21, no. 3, pages 645–652, 2002. (Cited on page 211.)

[Meline 2009] Timothy Meline. A research primer for communication sciences and disorders. Allyn & Bacon, 2009. (Cited on pages 149 and 150.)

[Melo 2015] Miguel Melo, Maximino Bessa, Kurt Debattista and Alan Chalmers. *Evaluation of Tone-Mapping Operators for HDR Video Under Different Ambient Luminance Levels*. In Computer Graphics Forum, volume 34, pages 38–49, 2015. (Cited on page 118.)

[Mertens 2009] Tom Mertens, Jan Kautz and Frank Van Reeth. *Exposure fusion: A simple and practical alternative to high dynamic range photography*. In Computer Graphics Forum, volume 28, pages 161–171. Wiley Online Library, 2009. (Cited on pages 34, 109 and 110.)

[Mikamo 2016] Michihiro Mikamo, Kotaro Mori, Bisser Raytchev, Toru Tamaki and Kazufumi Kaneda. *Binocular tone reproduction display for an HDR panorama image*. In Proceedings of the ACM Symposium on Applied Perception, pages 131–131. ACM, 2016. (Cited on pages 109 and 145.)

[Miller 2009] Liz Miller. Mood mapping: plot your way to emotional health and happiness. Pan Macmillan, 2009. (Cited on pages 153 and 154.)

[Möller 2013] Sebastian Möller, Steven Schmidt and Justus Beyer. *Gaming taxonomy: An overview of concepts and evaluation methods for computer gaming qoe*. In Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on, pages 236–241. IEEE, 2013. (Cited on pages 164 and 268.)

[Möller 2014] Sebastian Möller and Alexander Raake. Quality of experience: advanced concepts, applications and methods. Springer, 2014. (Cited on pages 18, 19, 45 and 48.)

[Möller 2017] Sebastian Möller. Quality engineering: Qualität kommunikationstechnischer systeme. Springer-Verlag, 2017. (Cited on page 47.)

[Moon 2015] Seong-Eun Moon and Jong-Seok Lee. *Perceptual experience analysis for tone-mapped HDR videos based on EEG and peripheral physiological signals*. IEEE Transactions on Autonomous Mental Development, vol. 7, no. 3, pages 236–247, 2015. (Cited on pages 210 and 211.)

[Moon 2017] Seong-Eun Moon and Jong-Seok Lee. *Implicit Analysis of Perceptual Multimedia Experience Based on Physiological Response: A Review*. IEEE Transactions on Multimedia, vol. 19, no. 2, pages 340–353, 2017. (Cited on pages 192, 193, 194, 195, 209, 210, 211 and 253.)

[Munafo 2017] Justin Munafo, Meg Diedrick and Thomas A. Stoffregen. *The virtual reality head-mounted display Oculus Rift induces motion sickness and is sexist in its effects*. Experimental Brain Research, vol. 235, no. 3, pages 889–901, 2017. (Cited on pages 173 and 180.)

[Mustafa 2012] Maryam Mustafa, Stefan Guthe and Marcus Magnor. *Single-trial EEG classification of artifacts in videos*. ACM Transactions on Applied Perception (TAP), vol. 9, no. 3, page 12, 2012. (Cited on page 210.)

[Myszkowski 2008] Karol Myszkowski, Rafal Mantiuk and Grzegorz Krawczyk. *High dynamic range video*. Synthesis Lectures on Computer Graphics and Animation, vol. 1, no. 1, pages 1–158, 2008. (Cited on page 38.)

[Narwaria 2012] Manish Narwaria, Matthieu Perreira Da Silva, Patrick Le Callet and Romuald Pépion. *Effect of tone mapping operators on visual attention deployment*. In SPIE Optical Engineering+ Applications, pages 84990G–84990G. International Society for Optics and Photonics, 2012. (Cited on pages 123 and 139.)

[Narwaria 2014] Manish Narwaria, Matthieu Perreira Da Silva, Patrick Le Callet and Romuald Pepion. *Single exposure vs tone mapped high dynamic range images: a study based on quality of experience*. In Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European, pages 2140–2144. IEEE, 2014. (Cited on page 118.)

[Nayar 1997] Shree K Nayar. *Catadioptric omnidirectional camera*. In Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, pages 482–488. IEEE, 1997. (Cited on page 42.)

[Nidal 2014] Kamel Nidal and Aamir Saeed Malik. Eeg/erp analysis: Methods and applications. Crc Press, 2014. (Cited on page 204.)

[Pan 1985] Jiapu Pan and Willis J Tompkins. *A real-time QRS detection algorithm*. IEEE transactions on biomedical engineering, no. 3, pages 230–236, 1985. (Cited on page 206.)

[Parasuraman 1985] Anantharanthan Parasuraman, Valarie A Zeithaml and Leonard L Berry. *A conceptual model of service quality and its implications for future research*. the Journal of Marketing, pages 41–50, 1985. (Cited on pages 148, 161 and 174.)

[Parasuraman 1994] Ananthanarayanan Parasuraman, Valarie A Zeithaml and Leonard L Berry. *Reassessment of expectations as a comparison standard in measuring service quality: implications for further research*. the Journal of Marketing, pages 111–124, 1994. (Cited on page 161.)

[Parasuraman 2002] A Parasuraman, V Zeithaml and L Berry. *SERVQUAL: a multiple-item scale for measuring consumer perceptions of service quality*. Retailing: critical concepts, vol. 64, no. 1, page 140, 2002. (Cited on pages 161, 163 and 173.)

[Perkis 2006] Andrew Perkis, Solveig Munkeby and Odd Inge Hillestad. *A model for measuring quality of experience*. In Signal Processing Symposium, 2006. NORSIG 2006. Proceedings of the 7th Nordic, pages 198–201. IEEE, 2006. (Cited on page 162.)

[Perrin 2017a] Anne-Flore Perrin, Cambodge Bist, Rémi Cozot and Touradj Ebrahimi. *Measuring quality of omnidirectional high dynamic range content*. In Applications of Digital Image Processing XL, volume 10396, page 1039613. International Society for Optics and Photonics, 2017. (Cited on page 106.)

[Perrin 2017b] Anne-Flore Perrin, Touradj Ebrahimi, Saman Zadtootaghaj, Steven Schmidt and Sebastian Möller. *Towards the Need Satisfaction in Gaming: A comparison of different gaming platforms*. In Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on, pages 1–3. IEEE, 2017. (Cited on page 163.)

[Perrin 2017c] Anne-Flore Nicole Marie Perrin and Touradj Ebrahimi. *Towards an implicit time-continuous assessment of Quality of Experience for immersive multimedia*. Technical report, 2017. (Cited on pages 253 and 269.)

[Perrin 2018a] Anne-Flore Perrin, Martin Rerabek, Walt Husak and Touradj Ebrahimi. *Evaluation of ICtCp color space and an Adaptive Reshaper for HDR and WCG*. IEEE CONSUMER ELECTRONICS MAGAZINE, 2018. (Cited on page 86.)

[Perrin 2018b] Anne-Flore Perrin, Martin Rerabek, Walt Husak and Touradj Ebrahimi. *Quality assessment of an HDR dual-layer backward-compatible codec compared to uncompromised SDR and HDR solutions*. IEEE transaction on broadcastion, special issue on Quality of Experience for advanced broadcast services, 2018. (Cited on page 98.)

[Pitas 2000] Ioannis Pitas. Digital image processing algorithms and applications. John Wiley & Sons, 2000. (Cited on page 31.)

[Plataniotis 2013] Konstantinos N Plataniotis and Anastasios N Venetsanopoulos. Color image processing and applications. Springer Science & Business Media, 2013. (Cited on pages 30 and 31.)

[Plutchik 2001] Robert Plutchik. *The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice*. American scientist, vol. 89, no. 4, pages 344–350, 2001. (Cited on page 126.)

[Poynton 1995] Charles A Poynton. *A Guided Tour of Colour Space*. In Advanced Television and Electronic Imaging Conference, New Foundation for Video Technology: The SMPTE, pages 167–180. SMPTE, 1995. (Cited on page 31.)

[Poynton 2002] Charles Poynton. *Chroma subsampling notation*. Retrieved June, vol. 19, page 2004, 2002. (Cited on page 32.)

[Pratt 2007] William K Pratt. Digital image processing: Piks scientific inside, volume 4. Wiley Online Library, 2007. (Cited on pages 30 and 31.)

[Rabenstein 2006] Rudolf Rabenstein, Sascha Spors and Peter Steffen. *Wave field synthesis techniques for spatial sound reproduction*. Topics in Acoustic Echo and Noise Control, vol. 5, pages 517–545, 2006. (Cited on page 27.)

[Rai 2017] Yashas Rai, Patrick Le Callet and Philippe Guillotel. *Which saliency weighting for omnidirectional image quality assessment?* In 9th International Conference on Quality of Multimedia Experience (QoMEX 2017). IEEE, 2017. (Cited on pages 135 and 137.)

[Reason 1978] J Reason. *Motion sickness: Some theoretical and practical considerations*. Applied ergonomics, vol. 9, no. 3, pages 163–167, 1978. (Cited on page 125.)

[Reinhard 2002] Erik Reinhard, Michael Stark, Peter Shirley and James Ferwerda. *Photographic tone reproduction for digital images*. ACM transactions on graphics (TOG), vol. 21, no. 3, pages 267–276, 2002. (Cited on page 109.)

[Reinhard 2007] Erik Reinhard, Timo Kunkel, Yoann Marion, Jonathan Brouillat, Rémi Cozot and Kadi Bouatouch. *Image display algorithms for high-and low-dynamic-range display devices.* Journal of the Society for Information Display, vol. 15, no. 12, pages 997–1014, 2007. (Cited on page 130.)

[Reinhard 2010] Erik Reinhard, Wolfgang Heidrich, Paul Debevec, Sumanta Pattanaik, Greg Ward and Karol Myszkowski. High dynamic range imaging: acquisition, display, and image-based lighting. Morgan Kaufmann, 2010. (Cited on pages 30, 31, 33, 34, 41 and 57.)

[Řeřábek 2016] Martin Řeřábek, Evgeniy Upenik and Touradj Ebrahimi. *JPEG backward compatible coding of omnidirectional images.* In Applications of Digital Image Processing XXXIX, volume 9971, page 99711O. International Society for Optics and Photonics, 2016. (Cited on page 43.)

[Richardson 2004] Iain E Richardson. H. 264 and mpeg-4 video compression: video coding for next-generation multimedia. John Wiley & Sons, 2004. (Cited on page 38.)

[Rigby 2007] Scott Rigby and Richard Ryan. *The Player Experience of Need Satisfaction (PENS): An applied model and methodology for understanding key components of the player experience.* Retrieved from immersyve. com/PENS_Sept07. pdf, 2007. (Cited on pages 50, 164, 165 and 176.)

[Robitza 2016] Werner Robitza and Alexander Raake. *(Re-) actions speak louder than words? A novel test method for tracking user behavior in web video services.* In Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on, pages 1–6. IEEE, 2016. (Cited on page 162.)

[Romero 2015] Ana Carla Leite Romero, Simone Fiuza Regacone, Daiane Damaris Baptista de Lima, Pedro de Lemos Menezes and Ana Cláudia Figueiredo Frizzo. *Event-related potentials in clinical research: guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400.* Audiology-Communication Research, vol. 20, no. 2, pages VII–VIII, 2015. (Cited on page 203.)

[Rosa 2015] Ana Filipa Rosa, Ana Isabel Martins, Victor Costa, Alexandra Queirós, Anabela Silva and Nelson Pacheco Rocha. *European Portuguese validation of the post-study system usability questionnaire (PSSUQ).* In Information Systems and Technologies (CISTI), 2015 10th Iberian Conference on, pages 1–5. IEEE, 2015. (Cited on page 144.)

[Rubinov 2010] Mikail Rubinov and Olaf Sporns. *Complex network measures of brain connectivity: uses and interpretations.* Neuroimage, vol. 52, no. 3, pages 1059–1069, 2010. (Cited on page 205.)

[Ryan 2006] Richard M Ryan, C Scott Rigby and Andrew Przybylski. *The motivational pull of video games: A self-determination theory approach.* Motivation and emotion, vol. 30, no. 4, pages 344–360, 2006. (Cited on page 173.)

[Sackl 2012] Andreas Sackl, Kathrin Masuch, Sebastian Egger and Raimund Schatz. *Wireless vs. wireline shootout: How user expectations influence quality of experience.* In Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on, pages 148–149. IEEE, 2012. (Cited on page 162.)

[Sackl 2014] Andreas Sackl and Raimund Schatz. *Got what you want? modeling expectations to enhance web qoe prediction.* In Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on, pages 57–58. IEEE, 2014. (Cited on page 162.)

[Sawyer 2011] Keith Sawyer. *The cognitive neuroscience of creativity: a critical review.* Creativity research journal, vol. 23, no. 2, pages 137–154, 2011. (Cited on pages 203 and 204.)

[Schild 2012] Jonas Schild, Joseph LaViola and Maic Masuch. *Understanding user experience in stereoscopic 3D games.* In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 89–98. ACM, 2012. (Cited on page 167.)

[Scholler 2012] Simon Scholler, Sebastian Bosse, Matthias Sebastian Treder, Benjamin Blankertz, Gabriel Curio, Klaus-Robert Muller and Thomas Wiegand. *Toward a direct measure of video quality perception using EEG.* IEEE transactions on Image Processing, vol. 21, no. 5, pages 2619–2629, 2012. (Cited on page 210.)

[Schubert 2001] Thomas Schubert, Frank Friedmann and Holger Regenbrecht. *The experience of presence: Factor analytic insights.* Presence, vol. 10, no. 3, pages 266–281, 2001. (Cited on pages 50, 164 and 165.)

[Seufert 2015] Michael Seufert, Sebastian Egger, Martin Slanina, Thomas Zinner, Tobias Hossfeld and Phuoc Tran-Gia. *A survey on quality of experience of HTTP adaptive streaming.* IEEE Communications Surveys & Tutorials, vol. 17, no. 1, pages 469–492, 2015. (Cited on page 48.)

[Sharples 2008] Sarah Sharples, Sue Cobb, Amanda Moody and John R Wilson. *Virtual reality induced symptoms and effects (VRISE): Comparison of head mounted display (HMD), desktop and projection display systems.* Displays, vol. 29, no. 2, pages 58–69, 2008. (Cited on page 125.)

[Sherman 2002] William R Sherman and Alan B Craig. Understanding virtual reality: Interface, application, and design. Elsevier, 2002. (Cited on page 43.)

[Skodras 2001] Athanassios Skodras, Charilaos Christopoulos and Touradj Ebrahimi. *The JPEG 2000 still image compression standard*. IEEE Signal processing magazine, vol. 18, no. 5, pages 36–58, 2001. (Cited on page 38.)

[Slater 1997] Mel Slater and Sylvia Wilbur. *A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments*. Presence: Teleoperators & Virtual Environments, vol. 6, no. 6, pages 603–616, 1997. (Cited on pages 49, 164 and 176.)

[Slater 2009] Mel Slater, Beau Lotto, Maria Marta Arnold and Maria V Sanchez-Vives. *How we experience immersive virtual environments: the concept of presence and its measurement*. Anuario de psicología, vol. 40, no. 2, 2009. (Cited on pages 211 and 217.)

[SMP 2014a] SMPTE ST 2084:2014. High dynamic range electro-optical transfer function of mastering reference displays, August 2014. (Cited on pages 35 and 88.)

[SMP 2014b] SMPTE ST 2086:2014. Mastering display color volume metadata supporting high luminance and wide color gamut images, 2014. (Cited on page 39.)

[SMP 2017] SMPTE ST 2094:2017. Dynamic metadata for color volume transform - klv encoding and mxf mapping, February 2017. (Cited on pages 38, 39, 72, 88 and 89.)

[Snellen 1868] Herman Snellen. Test-types for the determination of the acuteness of vision. Williams and Norgate, 1868. (Cited on page 58.)

[Statista2017 2017] C Statista2017. *Projected virtual reality headsets unit sales worldwide in 2016 (in million), by device*. https://www.statista.com/statistics/458037/virtual-reality-headsets-unit-sales-worldwide/, 2017. [Online; accessed 13-April-2017]. (Cited on page 167.)

[Strohmeier 2010] Dominik Strohmeier, Satu Jumisko-Pyykkö and Kristina Kunze. *Open profiling of quality: a mixed method approach to understanding multimodal quality perception*. Advances in multimedia, vol. 2010, 2010. (Cited on page 20.)

[Sullivan 2012] Gary J Sullivan, Jens Ohm, Woo-Jin Han and Thomas Wiegand. *Overview of the high efficiency video coding (HEVC) standard*. IEEE Transactions on circuits and systems for video technology, vol. 22, no. 12, pages 1649–1668, 2012. (Cited on page 40.)

[Takahashi 2008] Akira Takahashi, David Hands and Vincent Barriac. *Standardization activities in the ITU for a QoE assessment of IPTV*. IEEE Communications Magazine, vol. 46, no. 2, 2008. (Cited on page 46.)

[Tan 2015] Chek Tien Tan, Tuck Wah Leong, Songjia Shen, Christopher Dubravs and Chen Si. *Exploring gameplay experiences on the Oculus Rift.* In Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play, pages 253–263. ACM, 2015. (Cited on page 166.)

[Teas 1993] R Kenneth Teas. *Expectations, performance evaluation, and consumers' perceptions of quality.* The journal of marketing, pages 18–34, 1993. (Cited on page 161.)

[Teplan 2002] Michal Teplan*et al.* *Fundamentals of EEG measurement.* Measurement science review, vol. 2, no. 2, pages 1–11, 2002. (Cited on page 200.)

[Tian 2010] Yonghong Tian, Jaideep Srivastava, Tiejun Huang and Noshir Contractor. *Social multimedia computing.* Computer, vol. 43, no. 8, pages 27–36, 2010. (Cited on page 195.)

[Tosic 2009] Ivana Tosic and Pascal Frossard. *Low bit-rate compression of omnidirectional images.* In Picture Coding Symposium, 2009. PCS 2009, pages 1–4. IEEE, 2009. (Cited on page 43.)

[Tyagi 2018] Vipin Tyagi. Content-based image retrieval: Ideas, influences, and current trends. Springer, 2018. (Cited on pages 30 and 31.)

[Unger 2016] Jonas Unger, Francesco Banterle, Rafal Mantiuk and Gabriel Eilertsen. *The HDR-video pipeline: From capture and image reconstruction to compression and tone mapping.* In Eurographics 2016, 2016. (Cited on page 34.)

[Upenik 2016] Evgeniy Upenik, Martin Rerabek and Touradj Ebrahimi. *A Testbed for Subjective Evaluation of Omnidirectional Visual Content.* In 32nd Picture Coding Symposium, number EPFL-CONF-221560, 2016. (Cited on pages 117 and 258.)

[Upenik 2017] Evgeniy Upenik and Touradj Ebrahimi. *A simple method to obtain visual attention data in head mounted virtual reality.* In IEEE International Conference on Multimedia and Expo 2017, number EPFL-CONF-227457, 2017. (Cited on page 136.)

[Wang 2015] Danhong Wang, Randy L Buckner, Michael D Fox, Daphne J Holt, Avram J Holmes, Sophia Stoecklein, Georg Langs, Ruiqi Pan, Tianyi Qian, Kuncheng Li*et al.* *Parcellating cortical functional networks in individuals.* Nature neuroscience, vol. 18, no. 12, page 1853, 2015. (Cited on page 193.)

[Watson 1988] David Watson, Lee Anna Clark and Auke Tellegen. *Development and validation of brief measures of positive and negative affect: the PANAS scales.* Journal of personality and social psychology, vol. 54, no. 6, page 1063, 1988. (Cited on page 126.)

[Watts 1998] Duncan J Watts and Steven H Strogatz. *Collective dynamics of "small-world" networks.* nature, vol. 393, no. 6684, pages 440–442, 1998. (Cited on pages 205 and 228.)

[Welch 1967] Peter Welch. *The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms.* IEEE Transactions on audio and electroacoustics, vol. 15, no. 2, pages 70–73, 1967. (Cited on page 204.)

[Wells 2010] John D Wells, Damon E Campbell, Joseph S Valacich and Mauricio Featherman. *The effect of perceived novelty on the adoption of information technology innovations: a risk/reward perspective.* Decision Sciences, vol. 41, no. 4, pages 813–843, 2010. (Cited on pages 149, 150 and 151.)

[Witmer 1998] Bob G Witmer and Michael J Singer. *Measuring presence in virtual environments: A presence questionnaire.* Presence, vol. 7, no. 3, pages 225–240, 1998. (Cited on pages 49, 50, 164 and 165.)

[Wojciechowski 2004] Rafal Wojciechowski, Krzysztof Walczak, Martin White and Wojciech Cellary. *Building virtual and augmented reality museum exhibitions.* In Proceedings of the ninth international conference on 3D Web technology, pages 135–144. ACM, 2004. (Cited on page 43.)

[Wu 2009] Dan Wu, Lei Wang, Yuan-Ting Zhang, Bang-Yu Huang, Bo Wang, Shao-Jie Lin and Xiao-Wen Xu. *A wearable respiration monitoring system based on digital respiratory inductive plethysmography.* In Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE, pages 4844–4847. IEEE, 2009. (Cited on page 200.)

[Wu 2017] Hong Ren Wu and Kamisetty Ramamohan Rao. Digital video image quality and perceptual coding. CRC press, 2017. (Cited on page 45.)

[Yang 2012] Xuan S Yang, Linling Zhang, Tien-Tsin Wong and Pheng-Ann Heng. *Binocular tone mapping.* ACM Trans. Graph., vol. 31, no. 4, pages 93–1, 2012. (Cited on pages 109 and 145.)

[Yang 2013] Yuan Yang. *EEG signal analysis for brain-computer interfaces for large public applications.* PhD thesis, Télécom ParisTech, 2013. (Cited on page 241.)

[Yoshida 2005] Akiko Yoshida, Volker Blanz, Karol Myszkowski and Hans-Peter Seidel. *Perceptual evaluation of tone mapping operators with real-world scenes.* In Electronic Imaging 2005, pages 192–203. International Society for Optics and Photonics, 2005. (Cited on page 118.)

[Youngblut 2003] Christine Youngblut and Odette Huie. *The relationship between presence and performance in virtual environments: Results of a VERTS*

study. In Virtual Reality, 2003. Proceedings. IEEE, pages 277–278. IEEE, 2003. (Cited on page 49.)

[Yu 2015a] Matt Yu. *Dynamic Tone Mapping with Head-Mounted Displays*. 2015. (Cited on page 109.)

[Yu 2015b] Matt Yu, Haricharan Lakshman and Bernd Girod. *A framework to evaluate omnidirectional video coding schemes*. In Mixed and Augmented Reality (ISMAR), 2015 IEEE International Symposium on, pages 31–36. IEEE, 2015. (Cited on pages 42 and 117.)

[Zakharchenko 2016] Vladyslav Zakharchenko, Kwang Pyo Choi and Jeong Hoon Park. *Quality metric for spherical panoramic video*. In SPIE Optical Engineering+ Applications, pages 99700C–99700C. International Society for Optics and Photonics, 2016. (Cited on page 117.)

[Zakzanis 2001] Konstantine K Zakzanis. *Statistics to tell the truth, the whole truth, and nothing but the truth: formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers*. Archives of clinical neuropsychology, vol. 16, no. 7, pages 653–667, 2001. (Cited on page 62.)

[Zare 2016] Alireza Zare, Alireza Aminlou, Miska M Hannuksela and Moncef Gabbouj. *HEVC-compliant tile-based streaming of panoramic video for virtual reality applications*. In Proceedings of the 2016 ACM on Multimedia Conference, pages 601–605. ACM, 2016. (Cited on page 43.)

[Zeithaml 1993] Valarie A Zeithaml, Leonard L Berry and Arantharanthan Parasuraman. *The nature and determinants of customer expectations of service*. Journal of the academy of Marketing Science, vol. 21, no. 1, pages 1–12, 1993. (Cited on pages 148, 161, 163 and 174.)

[Zerman 2017] Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, Rafal Mantiuk and Frédéric Dufaux. *Effect of color space on high dynamic range video compression performance*. In Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on, pages 1–6. IEEE, 2017. (Cited on page 86.)

[Zhang 2010] Liwei Zhang, Guozhong Liu and Ying Wu. *Wavelet and common spatial pattern for EEG signal feature extraction and classification*. In Computer, Mechatronics, Control and Electronic Engineering (CMCE), 2010 International Conference on, volume 5, pages 243–246. IEEE, 2010. (Cited on page 241.)

[Zhang 2016] Yang Zhang, Matteo Naccari, Dimitris Agrafiotis, Marta Mrak and David R Bull. *High dynamic range video compression exploiting luminance masking*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, no. 5, pages 950–964, 2016. (Cited on page 38.)

[Zyda 2005] Michael Zyda. *From visual simulation to virtual reality to games*. Computer, vol. 38, no. 9, pages 25–32, 2005. (Cited on page 43.)

Last Updated on 25th May 2018

Anne-Flore Perrin

Digital signal processing and multimedia quality assessment engineer
7 rue des carrieres,
25370 JOUGNE, FRANCE
anne-flore.perrin25@gmail.com
Linked In | +33 674 84 1122

RESEARCH INTERESTS

Ultra High Definition (UHD), High Dynamic Range (HDR), High Frame Rate (HFR), Light Field, Virtual Reality (VR) and 360° Imaging, Multimedia Processing and Quality Assessment, Image and Video Compression, Quality of Experience and Story telling.

EDUCATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE (EPFL)
PhD in electronic engineering
Jul 2015 - Present | Multimedia Signal Processing Group (MMSPG) | Lausanne, Switzerland
Swiss Federal Institute of Technology
Supervisor: Touradj Ebrahimi

ÉCOLE SUPERIEURE D'INGÉNIEURS DE RENNES (ESIR)
Master Degree in computer science
Sep 2011 - Nov 2014 | Universitée Rennes 1, Rennes, France
Engineer school, specializing in image processing and computer graphics areas

CLASSE PRÉPARATOIRES AUX GRANDES ÉCOLES (CPGE)
Bachelor Degree in mathematics and physics
Sep 2009 - Jul 2011 | Lycé carnot, Dijon, France
Two-year highly selective classes to prepare for national competitive entrance exams to French engineer schools.

SKILLS

PROGRAMMING

Matlab • C/C++ • Java
ffmpeg • LaTeX• Caml
Python • C • Gstreamer 0.10
iOS • Linux • Windows
git • svn • Shell

LANGUAGES

French: Native speaker
English: Fluent, TOEIC 850/990 (2015)
Spanish: Basic user

RESEARCH

EPFL | PhD student, Research Assistant
Jul 2015 – Present | Multimedia Signal Processing Group (MMSPG) | Lausanne, Switzerland
Context-based Quality of Experience in immersive multimedia. | Supervisor: Prof. Touradj Ebrahimi

- **Quality evaluation for High Dynamic Range (HDR) and Wide Color Gamut (WCG).**
  - In partnership with Dolby Laboratory -
  Evaluations of compression operations, color space representation and compression-optimized transfer function.
- **360° HDR imaging exploration.**
  - In partnership with IRT B<>com -
  Camera-consumer omnidirectional HDR images dataset creation. Pair comparison methodology design tackling 360° constraints.
- **Expectations and novelty effect investigations in Virtual Reality (VR) gaming.**
  - In partnership with Deutch Telecom -
  Comparison between typical gaming platforms and VR gaming Head Mounted Display (HMD). Novelty effect analysis within VR gaming.
- **Physiological signals use to predict Sense of Presence (SoP).**
  SoP prediction under resolution, quality/compression, sound system variations and within typical multimedia consumption scenario.

EPFL | Video compression and quality researcher - Intern
Dec 2014 – May 2015 | Multimedia Signal Processing Group (MMSPG) | Lausanne, Switzerland
Supervisor: Prof. Touradj Ebrahimi

- SoP prediction of multimedia contents through physiological signals (EEG, ECG and respiration) analyses.
- Deep learning prospection.
- Transcoding from JPEG 2000 to HEVC, ToFuTV project (Transcoders of the future television).

ORANGE LABS | Video compression research engineer - Intern
Mar 2014 – Aug 2014 | TV Laboratory | Cesson Sevigné, France
Supervisor: Dr. Pierrick Philippe
Increased HEVC intra-coding by 1% by implementing adaptive incomplete transforms to compete with DST and DCT.

EXPERIENCE

AVIWEST | IT developer - Intern
Jun 2013 – Sep 2013 | Saint-Gregoire, France
Developed a video streaming application on a linux distribution for laptop. Prototype presented at the International Broadcasting Convention (IBC) 2013.

PUBLICATION LIST

JOURNAL PAPERS

Anne-Flore Perrin, Martin Rerabek, Walt Husak, and Touradj Ebrahimi. Quality assessment of an HDR dual-layer backward-compatible codec compared to uncompromised SDR and HDR solutions. IEEE transaction on broadcastion, special issue on Quality of Experience for advanced broadcast services, 2018.

Anne-Flore Perrin, Martin Rerabek, Walt Husak, and Touradj Ebrahimi. Evaluation of ICtCp color space and an adaptive reshaper for HDR and WCG. IEEE CONSUMER ELECTRONICS MAGAZINE, 2018.

CONFERENCE PAPERS

Anne-Flore Perrin, Cambodge Bist, Rémi Cozot, and Touradj Ebrahimi. Measuring quality of omnidirectional high dynamic range content. In Applications of Digital Image Processing XL, volume 10396, page 1039613. International Society for Optics and Photonics, 2017.

Anne-Flore Perrin, Touradj Ebrahimi, Saman Zadtootaghaj, Steven Schmidt, and Sebastian Möller. Towards the need satisfaction in gaming: a comparison of different gaming platforms. In Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on, pages 1–3. IEEE, 2017.

Anne-Flore Perrin, Martin Rerabek, and Touradj Ebrahimi. Towards prediction of sense of presence in immersive audiovisual communications. Electronic Imaging, 2016(16):1–8, 2016.

Anne-Flore Perrin, He Xu, Eleni Kroupi, Martin Rerabek, and Tourajd Ebrahimi. Multimodal dataset for assessment of quality of experience in immersive multimedia. In Proceedings of the 23rd ACM international conference on Multimedia, pages 1007–1010. ACM, 2015.

Adrià Arrufat, Anne-Flore Perrin, and Pierrick Philippe. Image coding with incomplete transform competition for hevc. In Image Processing (ICIP), 2015 IEEE International Conference on, pages 2349–2353. IEEE, 2015.

PROJECT

QOE-NET Early stage researcher (ESR)
Jul 2015 – Jun 2018
The innovative QoE maNagement in Emerging mulTimedia services (QoE-Net) is a Marie Skłodowska-Curie Initial Training Network funded by European Union's Horizon 2020 research and innovation programme under grant agreement No 643072A.

- Publicity responsible in organizational committee of for International Young Researcher Summit on QoE in Emerging Multimedia Services (QEEMS 2017), May 29–30, 2017 in Erfurt.
- Organizer of QoE-Net summer school on QoE management and implementation, September 11–15, 2017 in Lausanne.

REVIEWER

- Quality and User Experience (Springer)
- European Association for Signal Processing (EURASIP) Journal on Image and Video Processing
- Institute of Electrical and Electronics Engineers (IEEE) transaction on multimedia
- Association for Computing Machinery MultiMedia (ACM MM)

MISCELLANEOUS

COMPANIEROS FORMATION | Label handi-management
2014 | Online, France
Training course about management for disabled people.

VOLUNTEERING

- Club responsible in the ESIR students association (2012-2013)
- Animation and coaching of seven 10-year-old children, in the ACE association (2007-2009)

TRAVELING

United States and Europe (e.g. UK, Germany, Norway, Portugal).

HOBBIES

- Sports: Dance, Running, Ski, Rugby and Badminton.
- Arts: Drawing, Image editing, Play the piano.

Ce document a été imprimé au Centre d'impression EPFL,
imprimerie climatiquement neutre, certifiée myClimate.