The Quality of Experience of Emerging Display Technologies

Peter Andras Kara

Faculty of Science, Engineering and Computing

Kingston University

A thesis submitted in partial fulfillment of the requirements of the University for the award of

Doctor of Philosophy

March 2019

I dedicate this thesis to my parents.

Declaration

The thesis does not contain any material that has been previously submitted for an award at an institute of Higher Education either in the UK or overseas. I declare that all content provided in the thesis is my own original work, and that any references to or use of other sources have been clearly acknowledged within the text.

Acknowledgments

There are numerous great people I need to thank before beginning my doctoral thesis, as they all contributed to the successful execution and dissemination of my work. I am grateful to the following individuals, categorized by their institutions:

Kingston University: Maria G. Martini, for being an absolutely outstanding supervisor, Fatima M. Felisberti, for providing scientific and spiritual guidance, Vesna Brujic-Okretic, for her unwavering vision of long-term goals, and Nabajeet Barman, for continuously encouraging me to disseminate my scientific results and for repeatedly acknowledging my progress. I also wish to thank the institution itself for granting countless hours of access to the state-of-the-art Center of Augmented and Virtual Environments (CAVE), where I carried out the majority of my local research. Furthermore, I wish to thank the IT support of the institution for all the technical assistance I received on Penrhyn Road Campus.

Holografika: Tibor Balogh, for sharing his vast experience in light field with me, Attila Barsi, for his exemplary dedication to research and development, Aron Cserkaszky, for his scientific excellence in our work together, Peter T. Kovacs, for setting the initial direction of research on perceived quality, Zsolt Nagy, for his prompt support in research configurations and dissemination, and every other employee of the company who assisted my activities either in the office or in the laboratory.

Budapest University of Technology and Economics: Sandor Imre, for his supervision and support of my work during B.Sc., M.Sc. and Ph.D., and Laszlo Bokor, for his contributions to research. I wish to thank the entire Department of Networked Systems and Services (BME-HIT) for supporting and encouraging me, and for providing the knowledge and skills that were required for my activities in the European project that led to the results presented in this thesis.

I wish to thank Luigi Atzori and all the seniors of EU H2020 QoE-Net for successfully carrying out the project, and thus, providing an excellent scientific opportunity for twelve Early Stage Researchers (ESRs), including me. From these ESRs, I wish to particularly thank Roopak R. Tamboli and Werner Robitza for the fruitful collaborations we had together. I wish to thank every single individual who participated in the subjective tests. Finally, I wish to thank Aniko Simon for her great efforts to support me over the past decade, and I wish to thank my parents, simply for everything.

Abstract

As new display technologies emerge and become part of everyday life, the understanding of the visual experience they provide becomes more relevant. The cognition of perception is the most vital component of visual experience; however, it is not the only cognition that contributes to the complex overall experience of the end-user. Expectations can create significant cognitive bias that may even override what the user genuinely perceives. Even if a visualization technology is somewhat novel, expectations can be fuelled by prior experiences gained from using similar displays and, more importantly, even a single word or an acronym may induce serious preconceptions, especially if such word suggests excellence in quality.

In this interdisciplinary Ph.D. thesis, the effect of minimal, one-word labels on the Quality of Experience (QoE) is investigated in a series of subjective tests. In the studies carried out on an ultra-high-definition (UHD) display, UHD video contents were directly compared to their HD counterparts, with and without labels explicitly informing the test participants about the resolution of each stimulus.

The experiments on High Dynamic Range (HDR) visualization addressed the effect of the word "premium" on the quality aspects of HDR video, and also how this may affect the perceived duration of stalling events. In order to support the findings, additional tests were carried out comparing the stalling detection thresholds of HDR video with conventional Low Dynamic Range (LDR) video.

The third emerging technology addressed by this thesis is light field visualization. Due to its novel nature and the lack of comprehensive, exhaustive research on the QoE of light field displays and content parameters at the time of this thesis, instead of investigating the labeling effect, four phases of subjective studies were performed on light field QoE. The first phases started with fundamental research, and the experiments progressed towards the concept and evaluation of the dynamic adaptive streaming of light field video, introduced in the final phase.

List of Publications

Publications relevant to the Thesis

2019

- P. A. Kara, W. Robitza, N. Pinter, A. Raake, M. G. Martini, A. Simon, "Comparison of HD and UHD video quality with and without the influence of the labeling effect" in Springer Quality and User Experience, vol. 4, no. 4, pp. 1–29.
- P. A. Kara, A. Cserkaszky, M. G. Martini, L. Bokor, A. Simon, "The effect of labeling on the perceived quality of HDR video transmission" in Springer Cognition, Technology & Work, pp. 1–17.
- P. A. Kara, R. R. Tamboli, O. Doronin, A. Cserkaszky, A. Barsi, Zs. Nagy, M. G. Martini, A. Simon, "The key performance indicators of projection-based light field visualization" in Taylor & Francis Journal of Information Display, vol. 20, no. 2, pp. 81–93.

2018

- P. A. Kara, A. Cserkaszky, M. G. Martini, A. Barsi, L. Bokor, T. Balogh, "Evaluation of the Concept of Dynamic Adaptive Streaming of Light Field Video" in IEEE Transactions on Broadcasting, vol. 64, no. 2, pp. 407–421.
- P. A. Kara, R. R. Tamboli, A. Cserkaszky, M. G. Martini, A. Barsi, L. Bokor, "The Viewing Conditions of Light-Field Video for Subjective Quality Assessment", International Conference on 3D Immersion (IC3D), Brussels, Belgium, December.
- P. A. Kara, R. R. Tamboli, A. Cserkaszky, M. G. Martini, A. Barsi, L. Bokor, "The perceived quality of light-field video services", SPIE Applications of Digital Image Processing, San Diego, USA, August.
- P. A. Kara, A. Cserkaszky, M. G. Martini, "Premium HDR: The Impact of a Single Word on the Quality of Experience of HDR Video", International Conference on Multimedia and Expo (ICME) Emerging Multimedia Systems and Applications (EMSA), San Diego, USA, July.

2017

- P. A. Kara, A. Cserkaszky, A. Barsi, M. G. Martini, T. Balogh, "Towards Adaptive Light Field Video Streaming" in IEEE COMSOC MMTC Communications – Frontiers, vol. 12, no. 4, pp. 50–55.

- P. A. Kara, P. T. Kovacs, S. Vagharshakyan, M. G. Martini, S. Imre, A. Barsi, K. Lackner, T. Balogh, "Perceptual Quality of Reconstructed Medical Images on Projection-based Light Field Displays", in Springer eHealth 360, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (LNICST), vol. 181, pp. 476–483.

- P. A. Kara, A. Cserkaszky, A. Barsi, T. Papp, M. G. Martini, L. Bokor, "The Interdependence of Spatial and Angular Resolution in the Quality of Experience of Light Field Visualization", International Conference on 3D Immersion (IC3D), Brussels, Belgium, December.

- P. A. Kara, A. Cserkaszky, S. Darukumalli, A. Barsi, M. G. Martini, "On the Edge of the Seat: Reduced Angular Resolution of a Light Field Cinema with Fixed Observer Positions", International Conference on Quality of Multimedia Experience (QoMEX), Erfurt, Germany, May.

- P. A. Kara, W. Robitza, A. Raake, M. G. Martini, "The Label Knows Better: The Impact of Labeling Effects on Perceived Quality of HD and UHD Video Streaming", International Conference on Quality of Multimedia Experience (QoMEX), Erfurt, Germany, May.

2016

- P. A. Kara, M. G. Martini, P. T. Kovacs, S. Imre, A. Barsi, K. Lackner, T. Balogh, "Perceived Quality of Angular Resolution for Light Field Displays and the Validity of Subjective Assessment", International Conference on 3D Imaging (IC3D), Liege, Belgium, December.

- P. A. Kara, M. G. Martini, C. T. Hewage, F. M. Felisberti, "Times change, stalling stays: Subjective Quality Assessment over Time of Stalling in Autostereoscopic 3D Video Services", International Conference on Signal Image Technology & Internet Based Systems (SITIS) International Workshop on Quality of Multimedia Services (QUAMUS), Naples, Italy, November.

- P. A. Kara, P. T. Kovacs, S. Vagharshakyan, M. G. Martini, A. Barsi, T. Balogh, A. Chuchvara, A. Chehaibi, "The Effect of Light Field Reconstruction and Angular Resolution Reduction on the Quality of Experience", International Conference on Signal Image Technology & Internet Based Systems (SITIS) International Workshop on Quality of Multimedia Services (QUAMUS), Naples, Italy, November.

- P. A. Kara, P. T. Kovacs, M. G. Martini, A. Barsi, K. Lackner, T. Balogh, "Viva la Resolution: The Perceivable Differences between Image Resolutions for Light Field Displays", ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS), Berlin, Germany, August.

- P. A. Kara, M. G. Martini, S. Rossi, "One Spoonful or Multiple Drops: Investigation of Stalling Distribution and Temporal Information for Quality of Experience over Time", International Conference on Telecommunications and Multimedia (TEMU), Heraklion, Greece, July.

- P. A. Kara, W. Robitza, M. G. Martini, C. T. Hewage, F. M. Felisberti, "Getting Used to or Growing Annoyed: How Perception Thresholds and Acceptance of Frame Freezing Vary Over Time in 3D Video Streaming", IEEE International Conference on Multimedia and Expo (ICME) International Packet Video Workshop (PV), Seattle, USA, July.

- P. A. Kara, P. T. Kovacs, M. G. Martini, A. Barsi, K. Lackner, T. Balogh, "From a Different Point of View: How the Field of View of Light Field Displays affects the Willingness to Pay and to Use", International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, Portual, June.

Presentations

2018

- "Let there be Light(-Field): on the present & future of consumer-grade & professional applications of glasses-free 3D technology", Stereopsia Thematic Conference: Trends in stereo 3D, Brussels, Belgium, December.

2017

- "Quality of Experience of 3D Visualization on Light Field Displays", ETSI Workshop on Multimedia Quality in Virtual, Augmented or other Realities, Sophia Antipolis, France, May.

Further Publications

2019

- A. Cserkaszky, P. A. Kara, R. R. Tamboli, A. Barsi, M. G. Martini, L. Bokor, T. Balogh, "Angularly-Continuous Light-field Format: Concept, Implementation and Evaluation" in Wiley Journal of the Society for Information Display, vol. 27, no. 7, pp. 442–461.

2018

- R. R. Tamboli, A. Cserkaszky, P. A. Kara, A. Barsi, M. G. Martini, "Objective quality evaluation of an angularly-continuous light-field format", International Conference on 3D Immersion (IC3D), Brussels, Belgium, December.
- K. K. Vupparaboina, R. R. Tamboli, S. Manne, P. A. Kara, M. G. Martini, A. Barsi, A. Richhariya, S. Jana, "Towards true-to-scale 3D reconstruction of the human face using structured light projection and off-the-shelf cameras", International Conference on 3D Immersion (IC3D), Brussels, Belgium, December.
- R. R. Tamboli, P. A. Kara, N. Bisht, A. Barsi, M. G. Martini, S. Jana, "Objective Quality Assessment of 2D Synthesized Views for Light-field Visualization", International Conference on 3D Immersion (IC3D), Brussels, Belgium, December.
- A. Cserkaszky, A. Barsi, Zs. Nagy, G. Puhr, T. Balogh, P. A. Kara, "Real-time light-field 3D telepresence", 7th European Workshop on Visual Information Processing (EUVIP), Tampere, Finland, November.

- A. Cserkaszky, P. A. Kara, A. Barsi, M. G. Martini, T. Balogh, "Light-fields of Circular Camera Arrays", European Signal Processing Conference (EUSIPCO), Rome, Italy, September.

- R. R. Tamboli, P. A. Kara, A. Cserkaszky, A. Barsi, M. G. Martini, S. Jana, "Canonical 3D object orientation for interactive light-field visualization", SPIE Applications of Digital Image Processing, San Diego, USA, August.

- R. R. Tamboli, K. K. Vupparaboina, S. Manne, P. A. Kara, A. Cserkaszky, M. G. Martini, A. Richhariya, S. Jana, "Towards Euclidean auto-calibration of stereo camera arrays", SPIE Optical System Alignment, Tolerancing, and Verification, San Diego, USA, August.

- A. Cserkaszky, P. A. Kara, R. R. Tamboli, A. Barsi, M. G. Martini, T. Balogh, "Light-field capture and display systems: limitations, challenges, and potentials", SPIE Novel Optical Systems Design and Optimization, San Diego, USA, August.

- A. Cserkaszky, P. A. Kara, A. Barsi, M. G. Martini, "The potential synergies of visual scene reconstruction and medical image reconstruction", SPIE Novel Optical Systems Design and Optimization, San Diego, USA, August.

- R. R. Tamboli, P. A. Kara, A. Cserkaszky, A. Barsi, M. G. Martini, B. Appina, S. S. Channappayya, S. Jana, "3D Objective Quality Assessment of Light Field Video Frames", 3DTV Conference, Stockholm, Sweden, June.

- A. Cserkaszky, P. A. Kara, A. Barsi, M. G. Martini, "Expert Evaluation of a Novel Light-field Visualization Format", 3DTV Conference, Stockholm, Sweden, June.

- R. R. Tamboli, M. S. Reddy, P. A. Kara, M. G. Martini, S. S. Channappayya, S. Jana, "A High-angular-resolution Turntable Data-set for Experiments on Light Field Visualization Quality", International Conference on Quality of Multimedia Experience (QoMEX), Pula, Italy, May.

- R. R. Tamboli, P. A. Kara, B. Appina, M. G. Martini, S. S. Channappayya, S. Jana, "Effect of Primitive Features of Content on Perceived Quality of Light Field Visualization", International Conference on Quality of Multimedia Experience (QoMEX), Pula, Italy, May.

2017

- W. Robitza, A. Ahmad, P. A. Kara, L. Atzori, M. G. Martini, L. Sun, A. Raake, "Challenges of Future Multimedia QoE Monitoring for Internet Service Providers" in Springer Multimedia Tools and Applications, vol. 76, no. 21, pp. 22243–22266.

- A. Cserkaszky, P. A. Kara, A. Barsi, M. G. Martini, "Towards Display-Independent Light-Field Formats", International Conference on 3D Immersion (IC3D), Brussels, Belgium, December.

- O. Doronin, P. A. Kara, A. Barsi, M. G. Martini, "Ray Tracing for HoloVizio Light Field Displays", International Conference on 3D Immersion (IC3D), Brussels, Belgium, December.

- P. A. Kara, Zs. Nagy, M. G. Martini, A. Barsi, "Cinema as Large as Life: Large-scale Light Field Cinema System", International Conference on 3D Immersion (IC3D), Brussels, Belgium, December.

- A. Cserkaszky, P. A. Kara, A. Barsi, M. G. Martini, "To Interpolate or not to Interpolate: Subjective Assessment of Interpolation Performance on a Light Field Display", IEEE International Conference on Multimedia and Expo (ICME) Workshop on Hot Topics in 3D Multimedia (Hot3D), Hong Kong, China, July.

- T. Balogh, P. A. Kara, "VR versus LF: Towards the limitation-free 3D", SPIE Digital Optical Technologies International Symposium, Munich, Germany, June.

- O. Doronin, P. A. Kara, A. Barsi, M. G. Martini, "Screen-Space Ambient Occlusion for Light Field Displays", International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG), Pilsen, Czech Republic, May.

- S. Darukumalli, P. A. Kara, A. Barsi, M. G. Martini, "How to Rate the New Glasses-free Experience: Subjective Quality Assessment Methodologies for Experiments on Light Field Displays", International Young Researcher Summit on Quality of Experience in Emerging Multimedia Services (QEEMS), Erfurt, Germany, May.

- P. A. Kara, A. Cserkaszky, A. Barsi, M. G. Martini, "The Couch, the Sofa, and Everything in between: Discussion on the Use Case Scenarios for Light Field Video Streaming Services", International Young Researcher Summit on Quality of Experience in Emerging Multimedia Services (QEEMS), Erfurt, Germany, May.

2016

- S. Darukumalli, P. A. Kara, A. Barsi, M. G. Martini, T. Balogh, "Subjective Quality Assessment of Zooming Levels and Image Reconstructions based on Region of Interest for Light Field Displays", International Conference on 3D Imaging (IC3D), Liege, Belgium, December.
- S. Darukumalli, P. A. Kara, A. Barsi, M. G. Martini, T. Balogh, A. Chehaibi, "Performance Comparison of Subjective Assessment Methodologies for Light Field Displays", IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Limassol, Cyprus, December.
- W. Robitza, P. A. Kara, M. G. Martini, A. Raake, "On the Experimental Biases in User Behavior and QoE Assessment in the Lab", IEEE Global Communications Conference, Exhibition & Industry Forum (GLOBECOM) IEEE International Workshop on Quality of Experience for Multimedia Communications (QoEMC), Washington DC, USA, December.

Glossary of Terms

ACR	Absolute Category Rating
ANOVA	Analysis of Variance
AR	Augmented Reality
CI	Confidence Interval
DASH	Dynamic Adaptive Streaming over HTTP
DCR	Degradation Category Rating
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
EEG	Electroencephalogram
fMRI	functional Magnetic Resonance Imaging
FOV	Field of View
FP	Full Parallax
FR	Full Reference
HD	High Definition
HDR	High Dynamic Range
HEVC	High Efficiency Video Coding
HFR	High Frame Rate

HLG	Hybrid Log-Gamma
HPO	Horizontal Parallax Only
HSD	Honestly Significant Difference
HTTP	HyperText Transfer Protocol
HUD	Head-up Display
HVS	Human Visual System
IDMS	International Display Measurement Standard
ITU	International Telecommunication Union
JND	Just Noticeable Difference
KPI	Key Performance Indicator
LDR	Low Dynamic Range
LF	Light Field
MLLFD	Multi-Layer Light Field Display
MOS	Mean Opinion Score
MS-SSIM	Multi-Scale Structural Similarity Index
OLED	Organic Light-Emitting Diode
PAL	Phase Alternating Line
PC	Paired Comparison
PoE	Preference of Experience
ppcm	Pixel Per Centimeter

ppi	Pixel Per Inch
PQ	Perceptual Quantizer
PSNR	Peak Signal-to-Noise Ratio
QoE	Quality of Experience
SAMVIQ	Subjective Assessment Methodology for Video Quality
SID	Society for Information Display
SSIM	Structural Similarity Index
TI	Temporal Information
TMO	Tone Mapping Operator
UEQ	User Experience Questionnaire
UHD	Ultra High Definition
VIF	Visual Information Fidelity
VoD	Video on Demand
VQEG	Video Quality Experts Group
VQM	Video Quality Metric
VSNR	Visual Signal-to-Noise Ratio
VR	Virtual Reality
VVA	Valid Viewing Area
WVGA	Wide Video Graphics Array
WTP	Willingness to Pay

Contents

List of Figures

Chapter 1

Introduction

We live in a visual world. The vast majority of the information our brains gather comes from visual sources, and in this era of rapidly developing technology, it is not surprising that we spend more and more of our time looking at screens. Some of these screens are small and fit into the palms of our hands, while some others are large and need vast spaces to accommodate them. We watch some of them from afar, while others are situated mere centimeters from our eyes. We look at screens as a part of our daily occupations, in pursuit of entertainment, and for countless other reasons. The importance of display technologies has been continuously escalating in the past century, and it is expected to accelerate exponentially in the years to come.

My father always used to tell me — still tells me, and hopefully, he will continue to tell me for many more years — things to think about, things to consider. One of these things was,

> "Get yourself a fine bed with a high-quality mattress. After all, you spend nearly a third of your life in it, so make sure that the quality is right."

I have to admit, he was right, and I thank him for his words of wisdom. But how are beds and mattresses connected to what was written in the previous paragraph? Well, if a person spends more and more time looking at displays, then there is an evident need for excellence in visual quality. Honestly speaking, I am quite certain there are many of us who now spend more time looking at displays than resting in bed.

Quality can be measured in many different ways. No matter which one or ones we personally prefer in the assessment of display technologies, we have to face the fact that at the end of the day it is the Quality of Experience (QoE) that determines the real value of such systems. We can have dazzling numbers objectively describing how amazing a given display is, it is utterly irrelevant if the users are not satisfied during practical usage.

At the time of this thesis, we have multiple simultaneously emerging display technologies. While some of them enable the content to be watched in either 2D or 3D without the need for near-eye viewing devices, others, such as stereoscopic 3D, Virtual Reality (VR) and Augmented Reality (AR), rely on such gears. Personally, I have never really been a big fan of display technologies that force me to have something on my head while viewing the content, and this preference is reflected in my research directions as well.

In this thesis, the QoE of three emerging display technologies — that do not rely on near-eye viewing devices — is addressed. First, ultra-high-definition (UHD) displays are investigated, particularly 4K televisions, which are already present on the consumer market with rapidly growing content types, especially video. Second, High Dynamic Range (HDR) displays are studied, which are currently surfacing in everyday life, yet true HDR contents are still scarce, but HDR is expected to become a common visualization format in the very near future. Lastly, the perceived quality of light field (LF) displays is assessed, which are only expected to emerge in home multimedia consumption scenarios somewhere in the following decade.

UHD and HDR are three-letter acronyms that are now basically everywhere in the commercial world. With 8K resolution reaching the shops, 4K is becoming a de facto standard of visualization: new movies come out already in 4K and old ones are being remastered, live programs converge towards 4K broadcasting, online real-time video sharing platforms embrace UHD resolution and users are provided with affordable UHD-capable recording devices. However, this aforementioned transition towards UHD is still heavily in progress, even though it happens at a fast pace. HDR televisions are now also available to regular customers, but due to the lack of appropriate content, it is more favored and sought after in the use case scenario of gaming.

Both UHD and HDR are rapidly becoming default in gaming visualization, especially for PC gaming and consoles. Although the strength of the global gaming industry is unquestionable and it has become one of the main driving forces of display technology, gaming is not directly included in the scope of the research presented in this thesis. My work mainly focused on the perceived quality of videos.

As potential customers meet these acronyms regularly, they develop a sense of quality attached to them, sometimes even without personal visual experience. Yet such expectations may fundamentally distort experience and satisfaction. This phenomenon is commonly known as the labeling effect, when the known attributes and properties of entities create a bias in how such entities are perceived.

On the other hand, light field displays are so novel that most users are not even aware that such technology exists. Due to this novelty, the labeling effect does not apply to them, or at least, not in the manner as it does for UHD and HDR. Furthermore, light field visualization is essentially different from conventional 2D televisions, while UHD and HDR are primarily the same, but come with the added values of a higher resolution, and a greater bit depth and dynamic range, respectively.

Therefore, I chose a more fundamental assessment of light field QoE, which does not involve the labeling effect. As the research on perceived quality using real light field displays was very limited at the time I started working on them in 2015 — during my first industrial secondment at Holografika Ltd. in Hungary — I began addressing basic display and content parameters, and continued towards the sophisticated exploitation of perceptual thresholds, visual sensitivity, and user preference.

This thesis is comprised of three main chapters, that separately address these three display technologies. In Chapter 2, I present my research on 4K UHD. Two subjective studies were carried out to compare UHD and upscaled HD videos visually. One of them used blind testing methodology, while the other informed the observers about the resolution. In the latter, observers were not only informed, but also sometimes misinformed by false resolution labels, in order to measure the power of the induced bias. The primary aim was to compare the genuine perceptual differences with the results of the labeling effect.

Chapter 3 also deals with the labeling effect, but it is applied to HDR visualization. As multiple standards are now competing with each other — such as HDR10 and Dolby Vision — the user might come across different presentations of the three-letter acronym. Therefore, instead of comparing HDR to conventional Low Dynamic Range (LDR) visualization, HDR is basically compared to itself, with the addition of a label. Inspired by the naming convention of modern systems and services, videos in a visualization format called "Premium HDR" were shown to observers, who compared them to identical HDR videos without this label in mock-up experiments. A total of four experiments were carried out, investigating the effect of this given label. The first one compared the primary attributes of visualization, such as luminance, colors, frame rate and image quality. The second and the third investigated the perceptual thresholds for stalling detection in HDR and LDR videos, respectively, as these different formats may come with different cognitive demands, and the so-called "wow factor" or "wow effect" may affect visual attention as well. The fourth and final experiment examined the labeling effect with regards to the perceived length of stalling events, which is a key indicator of the quality of real-time video transmission.

Table 1.1: Demographic statistics of the test participants. The category of LF+ includes subjective tests that were carried on different platforms to support experiments on light field.

	UHD	**HDR**	**LF**	**LF+**
Number of subjective tests	2	4	10	2
Number of test participants	60	116	120	36
Number of male participants	46	85	86	25
Number of female participants	14	31	34	11
Lowest test participants age	18	20	19	18
Highest test participants age	40	60	59	37
Average test participants age	25.4	28.5	29	26.7

Chapter 4 reports the results of the series of experiments that were performed to address light field QoE, carried out in four phases. Phase 1 was the initial research on perceived quality, exploring display Field of View (FOV), content spatial and angular resolution, and view synthesis. Phase 2 involved viewing conditions, interpolation and the interdependence between spatial and angular resolution. Both Phases 1 and 2 used static models with plain-colored backgrounds, and Phases 3 and 4 targeted video QoE. In Phase 3, video resolutions and viewing conditions were investigated. Finally, in Phase 4, the concept of the dynamic adaptive streaming of light field video content was proposed and evaluated.

The findings presented in these three chapters are summarized in Chapter 5, which concludes the thesis. A total of 18 subjective studies were carried out and are presented in this work. The 10 studies on light field were performed at Holografika Ltd., the research on LDR was at the Budapest University of Technology and Economics, and the rest took place at Kingston University. The experiments involved the participation of 332 individuals. The demographic statistics of the test participants for each experiment type is given in Table 1.1.

Prior to each test, participants were subject to a screening based on Snellen charts and Ishihara plates, filtering for correct visual acuity and color vision, respectively. Each test participant was sufficiently informed about the experiment before participating at free will, and a form of consent was read and signed, acknowledging the conditions of participation and the confidential handling of data.

This thesis is an interdisciplinary work, as the presented work combines computer science and applied psychology — particularly in the experiments of Chapter 2 and Chapter 3. The presentation of research aims to make the content of this thesis approachable by the experts of both areas.

1.1 Experimental Methodology

As detailed earlier in the chapter, the results presented in this thesis are the outputs of numerous subjective tests. These tests did not always use the exact same measurement techniques and experimental setups, as each chosen methodology was specially tailored to the experiment, based on the research aim.

If there is a specific research question that can be addressed by subjective tests, then creating the experimental setup is the way to determine *how* that question should be answered. It typically includes all the necessary information that is required to carry out the tests, and therefore, it is vital to the repeatability — also known as the test-retest reliability — of the experiment.

In the scope of the experiments covered by the thesis, the first component of the experimental setup is the physical environment in which the tests took place. It applied to each and every subjective test that they were located in laboratory environments, with fixed parameters (e.g., room illumination), and every test environment was isolated from audiovisual distractions. Although having tests "out in the wild" imitate more realistic use case scenarios, such tests are affected by countless external factors — which sometimes can be challenging to control — thus, degrading the clear focus of the experiment (unless the research focus itself aims to measure the effect of the environment). Furthermore, tests carried out in laboratories provide de facto better test-retest reliability, especially if the other tests in the more realistic scenarios involve environmental randomness (e.g., an audio distraction that only occurs during the evaluation performed by a single individual).

As the topic of the thesis is built around visualization quality, the next important item on the list is the device itself, on which the test stimuli were shown to the test participants. Even though all of these devices (see Appendix B) were capable of audio playback (either by built-in speakers or by external audio systems), the tests did not involve any audio whatsoever; the stimuli were purely visual and the audio components were completely unused during the experiments. The experimental setup also defined how these displays were to be viewed by the test participants. By "how", the general viewing conditions are meant, including viewing distance (based on the height of the screen) and viewing angle (commonly center view but it may vary). One could immediately state that viewing height may also be subject to variation, for instance during a subject test using a light field display with vertical parallax support. Yet as no such display was involved in the research, the center view was applied along the vertical axis as well. The distance and the angle of observation are constant in

most of the experimental designs in the field; however, observer movement may be included in subjective tests. While the tests of Chapters 2 and 3 were all performed with static viewing positions, those of Chapter 4 involved observer movement due to the angle-dependent visualization of light field technology (unless if the research aim directly targeted static observation). Moreover, the chapter also presents a research where the investigated topic was particularly the effect of viewing conditions.

The test conditions — also known as test cases — define the varying and unvarying parameters of the visual stimuli. The emphasis is evidently on the varying parameters, as they serve as the very core of the experiment. Depending on the research, it is possible to select either a single parameter or multiple parameters that differentiate the test cases. However, there is a third option as well: to use only unvarying parameters, thus, presenting identical stimuli to the test participants. There may be several reasons to carry out a test like that. The most relevant to this thesis is to investigate the effect of cognitive bias. In practice, it means that the exact same stimulus is shown to a test participant, but different pieces of information are used to describe them, and therefore, the effect of the given label can be measured by quantifying the perceived difference. Different examples can be to investigate QoE over time or the recency effect by presenting a given stimulus at different times. In this thesis, there are examples for all three variations. In Chapter 3, the video sequences in the stimulus pairs were identical, in order to address the labeling effect. In Chapter 4, there are experiments where only a single parameter of light field visualization differed (e.g., spatial resolution, angular resolution, etc.), and in others, multiple parameters were changed between test conditions (e.g., spatial resolution *and* angular resolution).

When the test conditions are defined, they are applied to source contents, and thus, creating the test stimuli. These contents can be any audiovisual material, which can even include an entire virtual scenario [1]. In the experiments covered by the thesis, the contents were videos, and also still models and scenes in the first two phases of light field research. Furthermore, all the experiments used the so-called full test matrix methodology, meaning that every test condition was applied to every source content, and every test participant evaluated every visual stimulus. In the selection of the source videos, temporal information (TI) analysis was performed, which is a metric that describes how different adjacent frames are, providing an estimation regarding content dynamism diversity.

It is also the form of evaluation itself where many experiments of the thesis differed. Quality assessment is typically performed by using rating scales. Certain scales evaluate stimuli on their own, and they are commonly known as Absolute Category

Rating (ACR) scales, as they rate along well-defined numerical or qualitative values. Other scales may not address the quality of a given stimulus as it is, but they compare it to a different stimulus. To name the method, the terms "pair comparison" and "paired comparison" both frequently appear in the scientific literature. During paired comparisons, the quality of the two stimuli may differ, but they are evidently always the same content; technically, one could design an experimental setup where the source contents in a stimulus pair are different, but that would immensely reduce the clarity of the focus on the quality parameters. It is also possible to compare the quality to a flawless reference stimulus in order to measure the degradation caused by varying parameters. These are Degradation Category Rating (DCR) scales. If one stimulus is provided at a time and it is rated during or after its visualization, then that is a single stimulus method. If two stimuli are shown before quality assessment, then that is a double stimulus method. They do not necessarily need to follow each other, as comparisons may also happen via a simultaneous stimulus approach, where the two stimuli are presented at the same time, commonly side by side. Rating may be performed by using given rating options, or it can be continuous. Furthermore, quasi-continuous scales attempt to be the best of both worlds, as they appear continuous from the perspective of the test participant, but such a scale provides a simple numerical value in a limited assessment space. The subjective tests of this thesis used many different rating scales, all tailored for the experiment at hand. The tests in Chapter 2 involved 3-point and 7-point comparison scales; half of the test participants used one of these scales, and the other half compared via the other scale. As explained in the chapter, the decision of using two scales of the same type but different levels of grain was required as the impact of the scale itself was investigated as well. Beyond quality assessment, a post-experiment questionnaire collected information about the test procedure and the state of the test participant using the 20-point quasi-continuous scale. The experiments of Chapter 3 also used the more fine-grained 7-point comparison scale when comparing the identical stimuli, and the 5-point DCR scale was used to measure degradation detection and toleration. Due to the diversity in the research aims of the ten subjective tests of Chapter 4, the rating scales varied more. To distinguish perceptually similar yet genuinely different visual stimuli, the 10-point ACR scale and the 25-point quasi-continuous scale were used. The latter was also supported by the binary scale on quality acceptance, in order to enhance the understanding of the gathered data. Quality degradation was similarly recorded via the 5-point DCR scale, and the 7-point comparison scale was frequently utilized as well. Although the 5-point ACR scale is probably the most common in

the field of QoE, it did not appear in these experiments, as most of the evaluation tasks were comparisons, and absolute ratings used more fine-grained scale choices.

The test protocol describes the procedure in which the subjective test is carried out. It defines the temporal structure, the sequence order (i.e., randomization), and the possible clustering of the test conditions and/or contents, the separative pauses — also known as stimulus separations or separation screens — and the evaluation periods. For example, Figure 2.2 demonstrates this in Chapter 2, and Figure 3.5 serves the same purpose in Chapter 3, both for paired comparisons. The most common duration for separation screens in QoE studies is 5 seconds, which applied to almost every research covered by this thesis. The only deviation from this is in Chapter 2, where the reduction to 2 seconds was necessary due to the repeatedly presented labels (also 2 seconds) and the large test matrix. The assessment period was typically 10 seconds long. The guideline for protocol timing in general is specified by ITU-R Rec. BT.500-13[1] and ITU-T P.910[2]. As examples for randomization and clustering, both the test condition and the source stimulus order were uniquely randomized for every single test participant, extended with the requirement that adjacent stimulus pairs always had different contents in order to avoid repetition. One of the experiments in Chapter 4 involved test conditions with different rating tasks — and different rating scale types as well — and therefore, the conditions were clustered by the given task, in order to enable a better focus on the task at hand by preventing the cumbersome procedure of switching back and forth between the two subjective assessment tasks.

Before running the test itself, a training phase is required to help the test participants familiarize themselves with the assessment task, the device, the rating scale, and the test protocol. This is also where the aforementioned Snellen charts and Ishihara plates are involved. Based on the experimental setup, this phase prior to the subjective tests enables the communication of special instructions and information, such as the presence of labels, as it is detailed in Chapters 2 and 3. This is particularly important when there is a constant allocation of labels (e.g., label A always applies to the first stimulus and label B to the second one). During the task of a paired comparison, it is crucial that test participants clearly understand which stimulus is being compared to which. In the presented experiments, it was always the second stimulus in a pair that was compared to the first one. As it shall be seen later in Chapter 4, certain unusual visual phenomena generated by novel display technologies require extensive training, in order to achieve the desired level of experimental validity.

[1]Rec. BT.500: Methodologies for the subjective assessment of the quality of television images
[2]Rec. P.910: Subjective video quality assessment methods for multimedia applications

When all the tests are completed, the gathered data is then analyzed. The most common metric in the field for presenting quality assessment is the Mean Opinion Score (MOS). It is the arithmetic average of the collected data. On its own, it tends to serve its purpose very well, but it is not always sufficient [2]. Let us just think of the evaluation of a novel display technology, the subjective quality assessment of which may provide polarized results; one group of the test participants finds it excellent, while others completely reject it, resulting in MOS values around the middle of the scale. MOS values are typically accompanied by confidence intervals (CIs), based on deviation and sample size. In this thesis, the confidence level of 95% is used, which is the standard value in statistics. It is used to determine statistically significant differences, along with analysis of variance (ANOVA), Tukey's HSD (honestly significant difference) test, the Holm and the Bonferroni multiple comparison tests. All the subjective ratings collected from the 332 test participants are taken into account in the thesis and none was discarded. Discarding test results is a common practice when they deviate too much from ratings of the majority; however, even though many test conditions throughout the entire thesis received completely different subjective scores (e.g., in Chapter 2, one test participant rated all the stimuli in the paired comparisons to be the same, while another test participant did not use this rating option even once), these all contain scientifically valuable information (e.g., different manifestations of cognitive bias), and therefore, they should be kept, analyzed and presented. The presentation of the results is either performed via figures or tables. Figures may primarily visualize mean scores (e.g., Figures 3.8, 3.12, 4.6) and overall scoring distribution (e.g., Figures 2.3b, 3.20, 4.9), but individual ratings, personal scoring distributions and other types of subjective assessment information may be presented as well. For instance, in Chapter 2, Figures 2.6 and 2.7 show how each test participant used the rating options during the test. Figures 2.8 and 2.9 portray values aggregated from the ratings of separate test participants and sort them in ascending order. The same applies to Figures 3.10, 3.11, 3.22, 3.23, and 4.11. In these figures, the different individuals are along the horizontal x-axis, and the investigated values are along the vertical y-axis. However, while Figure 4.31 is essentially similar in this manner, it is important to note that the data in the figure was sorted by three independent variables (i.e., test conditions), and therefore, the values aligned vertically may not necessarily belong to the same test participant. Figure 4.11 also involves multiple vertically aligned values, yet these belong to the same test participants, as only a single test condition was used for data sorting.

1.2 Demographic Distribution in Research

In the real world, multimedia consumption is somewhat balanced when it comes to male and female viewers. This suggests that in order to have representative subjective experiments, the distribution of the test participants should be balanced as well regarding this matter; the male-to-female ratio is expected to be about 1:1, so roughly half of the test participants ought to be male, and thus, the other half should be female.

Although the subjective ratings generated by both demographic groups are unquestionably equally important, sometimes there are certain differences between how males and females perceive quality, and these differences can be recorded and measured in the assessment scores they provide during experiments. When Casal *et al.* [3] presented their results on the device characterization for conditional encoding at the meeting of the Video Quality Experts Group (VQEG), it was pointed out that the MOS of female test participants was higher in every test scenario, meaning that they tolerated the degraded stimuli more. Regarding audio quality, Sax [4] found that a female individual may hear the same sound with a higher level of sensitivity, in comparison to a male individual. There are also works in the field of QoE that address the effect of gender-based differences regarding the experimental scenario itself. For example, the work of Hyder *et al.* [5] targets virtual acoustic environments and how the size of the virtual conferencing rooms affects the localization performance of male and female test participants. Based on the collected data, it is stated that males may localize more concurrent talkers in smaller-sized rooms, while females perform their best in this regard in middle-sized environments. On the other hand, the publication of Zündorf *et al.* [6] concludes that in a multi-source sound environment (such as a cocktail party), males perform significantly better in collecting spatial information. The two genders may also differ in tasks such as searching on the web [7], which was the reason why Lamm *et al.* decided to only include female test participants in their work [8].

The ITU-T P.800 recommendation[3] on the methods for subjective determination of transmission quality (particularly on telephony) states the following in its annex:

> "No steps are taken to balance the numbers of male and female subjects unless the design of the experiment requires it."

[3]Rec. P.800: Methods for subjective determination of transmission quality

However, a more recent recommendation, P.913[4] highlights gender balance:

> "Likewise, participants will be approximately 50% female and 50% male, unless otherwise dictated by the experimental design (i.e., surveying females' perception of audio quality)."

As shown in the data of Table 1.1, the subjective tests presented in this thesis had a distribution of male and female participants that was far from being balanced; globally, out of the 332 individuals, 242 were male and only 90 were female. Naturally, the aim was to have a gender-balanced pool of test participants. Unfortunately, this was only possible to the extent of best effort; it was a very specific goal to involve as many female individuals in the experiments as possible, but it was constrained by what was available within the institution and within the social reach of the organizers. There was indeed the option to halt the research until a target percentage in gender balance (e.g., at least 40% female test participants) was achieved; however, that could have resulted in either of the following two very undesirable consequences: (a) the target is reached by reducing the total number of test participants per experiment, and thus, reducing the statistical strength of the collected assessment data, or (b) the target is reached by delaying the completion of the tests, unpredictably affecting both the publication of research results and the completion of the thesis.

Although gender balance in QoE research is unquestionably essential, it is a goal, a requirement that is not always met when it comes to subjective tests in practice. Due to rather simple circumstances, it can be quite a challenging issue to reach such numbers in research. If we take, for instance, experiments taking place at the information science and electrical engineering faculty of a technical university, at the time of this thesis, in many of such institutions, there is an apparent male dominance in population. Of course, it should be added that sometimes it is actually the other way around; in certain studies, females have superiority in numbers. Table 1.2 lists the relevant numbers on achieved gender distribution in the related scientific literature. Finally, it needs to be noted that, unfortunately, there are *many* publications in the field of QoE which involve test participants for subjective tests but they do not report gender distribution (or any demographic distribution) at all.

[4]Rec. 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

Table 1.2: Examples of gender distribution in related works.

Publication	**Participants**	**Males (%)**	**Females (%)**
Lamm *et al.* [8]	89	0 (0%)	89 (100%)
Szybillo *et al.* [9]	90	0 (0%)	90 (100%)
Bouchard *et al.* [1]	31	5 (16.1%)	26 (83.9%)
Burton *et al.* [10]	500	160 (32%)	340 (68%)
Berger *et al.* [11]	24	9 (37.5%)	15 (62.5%)
Cserkaszky *et al.* [12]	24	10 (41.6%)	14 (58.4%)
Verbeke *et al.* [13]	303	129 (42.6%)	174 (57.4%)
Li *et al.* [14]	42	18 (42.9%)	24 (57.1%)
Szanja *et al.* [15]	159	69 (43.4%)	90 (56.6%)
Sackl *et al.* [16]	49	24 (49%)	25 (51%)
Korshunov *et al.* [17]	24	12 (50%)	12 (50%)
Tanaka *et al.* [18]	36	18 (50%)	18 (50%)
Hamzaoui *et al.* [19]	292	155 (53%)	137 (47%)
Lick *et al.* [20]	161	86 (53.4%)	75 (46.6%)
Al-juboori *et al.* [21]	28	15 (53.6%)	13 (46.4%)
Narwaria *et al.* [22]	48	28 (58.3%)	20 (41.7%)
Moon *et al.* [23]	5	3 (60%)	2 (40%)
Shi *et al.* [24]	23	14 (60.9%)	9 (39.1%)
Bist *et al.* [25]	16	10 (62.5%)	6 (37.5%)
Rieh *et al.* [26]	15	10 (66.6%)	5 (33.3%)
Gachter *et al.* [27]	120	82 (68%)	38 (32%)
Viola *et al.* [28]	35	24 (68.6%)	11 (31.4%)
Darukumalli *et al.* [29]	20	14 (70%)	6 (30%)
Garber *et al.* [30]	120	84 (70%)	36 (30%)
Marton *et al.* [31]	33	24 (72.7%)	9 (27.3%)
Hulusic *et al.* [32]	20	15 (75%)	5 (25%)
Paudyal *et al.* [33]	20	15 (75%)	5 (25%)
Tamboli *et al.* [34]	20	16 (80%)	4 (20%)
Kovács *et al.* [35]	53	43 (81.1%)	10 (18.9%)
Van Wallendael *et al.* [36]	63	59 (93.7%)	4 (6.3%)

1.3 Contributions to Knowledge

- UHD and HD video contents were systematically compared, taking into account the labeling effect. The obtained subjective results show the statistically significant impact of cognitive bias and the difference in rating behavior between assessment scale option numbers, i.e., 3-point and 7-point comparison scales. The subjective scores conclude the lack of significant difference between the perceived quality of UHD and HD videos while avoiding labels and using the standardized test methodology, particularly the standard viewing distance for UHD content.

- The subjective assessment of identical HDR video stimuli was exposed to the labeling effect via the label "Premium HDR". The ratings indicate that the perception of quality aspects such as luminance, color, and image quality were positively biased by the label, while the frame rate was negatively assessed multiple times, due to the cognition of a trade-off between visual quality and frame rate. The results of the subjective test on the perceived duration of stalling events in HDR video were consistent with this, rating the identical stalling events in "Premium HDR" videos to be significantly longer. Without the inclusion of the labeling effect, the comparison of HDR and LDR stalling event detection thresholds showed that it can be easier for frame repetition to go unnoticed or to be more tolerated in HDR videos due to the higher level of cognitive load.

- The fundamental research on light field QoE determined levels of perceived quality regarding content spatial and angular resolution, visualization FOV, and addressed the perceptual effects of light field reconstruction and interpolation. The findings of the experiments on viewing conditions indicate tolerance towards angular disturbance in case of static observation methodology, in contrast to sideways observer movement. Multiple subjective studies revealed that the blur caused by low spatial resolution could improve the smoothness of the horizontal motion parallax if the angular resolution is insufficient. Based on these results, the concept of dynamic adaptive streaming of light field video is proposed in the thesis, using quality switches in content spatial and angular resolution, and exploiting the interdependence between them. The work is supported by the results of subjective evaluation.

1.4 Acknowledgment of Joint Research Efforts

- The work on UHD visualization was performed in collaboration with Werner Robitza. He contributed to the preparation of the visual stimuli; he generated the sequences from the available individual video frames, aided by his prior experience on the usage of the Lanczos filter, which also evidently supported the decision making process of the experimental setup.

- As the experiment on LDR stalling event detection (from the HDR test series) was carried out at the Budapest University of Technology and Economics, Laszlo Bokor provided invaluable support in setting up the local test environment and in organizing the participation of individuals from the institution. The latter also benefited multiple subjective tests on light field visualization.

- The research on light field view synthesis was enabled by Suren Vagharshakyan, as he applied his own implementation of the Shearlet transform to the rendered content, thus, creating half of the visual stimuli used in the experiment.

- Aron Cserkaszky utilized two interpolation techniques to create the contents of the research on light field interpolation. Our collaboration is detailed in the chapter, and it is clearly declared that only a given portion of the research is presented in the thesis.

- Attila Barsi generated all light field video sequences, which was a highly time- and resource-consuming process, and therefore, his expertise on light field video was imperative to achieve efficient rendering.

- The test stimuli for the evaluation of the concept of the dynamic adaptive streaming of light field video was objectively assessed via the metric of Roopak R. Tamboli, who personally applied the code of his implementation to the videos of the experiment, and thus, provided the corresponding output.

Chapter 2

Ultra High Definition Visualization

2.1 Introduction

The technical term "UHD" — referring to ultra-high-definition displays and contents — has entered our everyday lives in the past decade, and at the time of writing this thesis, it is slowly becoming a common format of multimedia consumption, on TVs, tablets, computer screens, cinema, and other technologies. The rise of UHD content is enabled by the fact that more and more UHD-capable displays emerge on the consumer market, thus, creating a vigorous competition which continuously reduces prices — especially entry-level prices — making such displays available to a wider range of consumers. Also, content creation and provision on amateur and professional levels shift towards UHD resolution as well, including real-time streaming services.

By now, it can be stated that most multimedia consumers have come across the term UHD in one way or another, even if they have not experienced the visuals of a true UHD content on a UHD display yet. These three letters are found highlighted on stickers and labels on displays in shops, they are emphasized in commercials on TV and the Internet, and content providers promote this attribute whenever they can, particularly when selling UHD on top of existing HD programming. At the same time, the first demos of "8K" entertainment systems are emerging.

UHD in the context of home entertainment can also be labeled "4K". To be more accurate, UHD TV can either refer to "UHD-1" (3840×2160 pixels) or "UHD-2" (7680×4320 pixels), standardized by ITU-R Rec. BT.2020[1]. The formats named "4K" (4096×2160) and "8K" (8192×4320) are standardized for cinema, defined by the Digital Cinema System Specification. In practice, UHD/4K commonly refers to

[1]Rec. BT.2020: Parameter values for ultra-high definition television systems for production and international programme exchange

the resolution of 3840×2160 pixels. In the scope of this thesis, this specific resolution is addressed and it is compared with Full HD (1920×1080).

The research in this chapter addresses the perceived differences between HD and UHD video while taking into consideration the labeling effect, that is, the effect a certain label (like "UHD") may have on users' opinions or decisions. The core of the work investigates the visual quality achieved by these resolutions, which plays an important role in the overall QoE of a user. As the study addresses perceived differences, one could assume the intention to question the added value of UHD video compared to HD, or the doubt regarding the relevance of the presence of UHD-capable displays on the market. This, however, is not the purpose of this research.

The work is mainly motivated by the appearance and the usage of the term "UHD" on commercial levels. As most of such cases strongly suggest superior visual quality through the higher amount of pixels on the screen, user expectations evidently rise. Expectations can not only influence the overall experience, but they can also affect the actual perception of visual quality.

Another motivation is that studies found in the literature are not entirely conclusive on whether UHD content can, in general, provide a statistically significant perceived quality difference compared to HD. At the very least, a high level of content dependency was found in many independent tests. Indeed, while electronics shops typically show so-called "eye candy" video contents on their displays — which are meant to push the limits of the displays' capabilities to show the potential buyers what visuals such displays can achieve — the average user does not spend the majority of his or her time watching short demo videos.

The inflated expectations, combined with a potential lack of major visual differences can lead to persistent forms of cognitive bias, resulting in severe distortions of QoE. Such distortions are present in everyday life, via given cognitive processes. Therefore, understanding these effects is just as important as the efforts to avoid or eliminate them from subjective studies. As the cognition of perception can be affected by other cognitions evoked by expectations, what the users observe can be viewed as a sort of illusion. Although they can make general QoE research more difficult, as it has been stated, they are indeed a part of our everyday lives, and should not be looked at as something inherently bad. After all, as French Enlightenment philosopher Voltaire wrote,

> "Illusion is the first of all pleasures."

The chapter primarily investigates research questions on the quality of HD and UHD video in the presence (or absence) of the labeling effect. This phenomenon is addressed by describing a series of subjective tests that were conducted — a total of four studies with a different set of participants. In each of these tests, participants had to compare the visual quality of HD and UHD videos on a UHD display and choose the relative quality difference on a rating scale.

In one set of tests, subjects were made aware of the content resolution: before each video sequence, a label was shown ("HD" or "UHD"). However, some of these labels were intentionally presented in a misleading fashion, providing false information on video resolution, that is, certain paired comparisons involved two identical video sequences, but the labels suggested that they differed. The test paradigm was also repeated without labels, while keeping all other parameters of the experiment unchanged, in order to obtain quality ratings unbiased by labels, and to check how big the impact of labels would be.

As the overall perceived differences between two sequences — also considering the labeling effect — may not be great enough to be registered on a coarse 3-point comparison scale ("*Worse*", "*Same*", "*Better*"), both tests were additionally run using another rating scale with 7 option numbers.

With the set of four tests (labels/no labels, 3/7-point rating scale) the following two questions can be answered: 1) Can users see a difference between HD and UHD videos? 2) When rating this difference, are users more impacted by the labels than the actual visual quality of the content?

2.2 Related Research on UHD Video QoE

2.2.1 Standards and Recommendations

As with most technologies finding widespread use among the consumers, there are standards that govern how a technology is to be developed, evaluated, and integrated with other technologies. Standards or international recommendations provide for that interoperability. Among the most relevant international standards on the topic of UHD are documents from the International Telecommunication Union's telecommunication and broadcasting sectors (ITU-T and ITU-R).

While ITU-R Rec. BT.709 addresses HDTV (i.e., TV up to 1080p resolution), Rec. BT.2020 covers UHD and the corresponding specifications of dynamic range, color gamut and primaries, bit depths, frame rates, and pixel resolutions. Additional recommendations include Rec. BT.1769 which specifies parameter values for large

screen digital imagery and how to design a system that gives viewers visual experiences of a high-sensation of reality — which UHD was also developed for.

Subjective quality assessment tests are typically carried out in a rigorous fashion: users are placed in a dedicated testing room with specific lighting conditions and a certain viewing distance to the screen. Guidelines in ITU-T Recommendations P.910, P.911, and P.913, as well as ITU-R Rec. BT.500-13 may be applied in those tests. When subjectively evaluating the quality of UHD systems, the aspect of viewing distance plays a crucial role. Typically, for HDTV applications, human testers are seated at a distance of about 3 H to the TV, where H is the height of the display under study. This is specified in ITU-R Rec. BT.500-13 and ITU-T P.910, where the latter says that "the viewing distance should be defined taking into account not only the screen size, but also the type of screen, the type of application and the goal of the experiment." This preferred viewing distance, therefore, mainly depends on user preferences and may be determined empirically, but 3 H has emerged as the standard for HD testing. For UHD screens, however, another recommendation, Rec. BT.2022[2] was developed, which furthermore distinguishes between the preferred viewing distance and the *design* viewing distance. The latter is the most optimal distance at which "two adjacent pixels subtend an angle of 1 arc-min at the viewer's eye". This distance is 1.6 H for UHD-1 resolution. This distance is also employed by the method specified in ITU-R Rec. BT.2095, which is a test protocol for expert viewing.

Let us now consider a practical example. If we take a content with HD resolution visualized on a display also with HD resolution and a screen height of 100 cm, then the aforementioned 3 H viewing distance for a test participant in a subjective study is 300 cm. However, it is important to note that this example assumes that the content is visualized using the entire screen. In case the visual stimulus only covers 50% of the height of this specific screen, then the 3 H distance corresponds to 150 cm instead. Concerning the standardized 1.6 H distance, we have to consider that for a 55-inch screen, this distance is 1 meter, which may be much too close for many environments — in fact, it is hard to imagine that in a traditional living room scenario, viewers would sit that close to the screen. It should, therefore, be virtually impossible for users to distinguish HD from UHD at distances any further, which is very likely to happen in real life.

[2]Rec. BT.2022: General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays

2.2.2 Appearance of UHD in QoE Studies

We find the first major wave of literature covering UHD with regard to QoE in the years around 2013–2015, as tools and technical equipment to display UHD content become more readily available. Before that, research on this topic is scarce. Surveying the literature, particularly with a focus on subjectively comparing HD against UHD, it can be observed that this topic has not yet been studied conclusively.

Bae *et al.* [37] conducted a subjective study in 2013, in which HEVC-encoded UHD video sequences were presented at different target bitrates, color formats (YUV 4:4:4 and YUV 4:2:0), and viewing distances (0.75 H and 1.5 H). The authors used two source sequences from the year 2010 for their test, so it can be considered one of the first to investigate QoE for UHD video. Choosing a double-stimulus method (DSIS, see ITU-T Rec. P.910), the original source videos were compared against the encoded ones. The main results of the experiment were subjective ratings for the clips, however the authors did not specifically compare HD and UHD.

In 2014, Tanaka *et al.* [18] investigated the use of a double-stimulus (DSCQS, see ITU-R Rec. BT.500-13) method for subjectively evaluating 4K video quality. They found it to be viable but did not compare HD and UHD either.

Li *et al.* [14] compared different upscaling algorithms for use in content preparation for UHD transmissions via a subjective test. The authors took source material at 2160p, 1080p, and 720p resolutions and upscaled the latter two to UHD using several different algorithms. Their choice of viewing distance was motivated as a compromise between 0.75 H suggested in the literature, and 1.6 H based on the preferred viewing distance (as mentioned in the previous section) — they chose 1 H. In a paired-comparison test, which they deemed to be the most reliable for such tasks, they asked subjects to pick the preferred sequence, resulting in a Preference of Experience (PoE). Notably, users saw HD and UHD video at the same time, rendered in vertical stripes on the display. As expected, original UHD sequences achieved higher preference values, with Lanczos upscaling performing better than other methods. The authors also noted that "visual acuity on high motion content on 4K screen is significantly lower than in the normal HD condition", and that there is a center bias when viewing content at such low distances. However, the test methodology itself may be questioned: were users really rating the quality of the upscaling, or was the ability of the users to perceive striped patterns tested?

In the work of Weerakkody *et al.* [38], the authors conducted subjective verification tests for the HEVC standard, in which the potential for bitrate savings against H.264 was the main research target. Using five source sequences and H.264 and HEVC

encoding, sequences were presented to subjects at two viewing distances (1.5 H and 2 H). An analysis concluded that for subjects sitting at a greater distance, MOS values were higher, but only for HEVC content. This suggests that impairments of the codec could only be visible at a closer viewing range.

In 2015, Berger *et al.* [11] conducted a subjective study in which 15 contents and 4 sets of bitrates were chosen per content to investigate the impact of lossy video encoding with the HEVC codec. Subjects were asked to rate the visual quality of the processed stimuli. The authors specifically investigated the practical case of transmission chains, considering that in real-life, bottlenecks in networks may require bitrate and/or resolution reduction for video in order to be still viewable. The authors found that downsampling and later upsampling video (i.e., reducing the resolution during transmission) did not yield perceivable visual quality degradations. Quite the contrary, for some contents, perceived quality *improved* due to the contained camera noise or fast motion in the original source videos. They also found that a UHD transmission chain required only a slightly higher bitrate for the same visual quality when compared to HD.

In a 2016 study, Xie *et al.* [39] determined the required HEVC encoding bitrates for UHD transmissions through subjective tests. They found that naïve viewers could not distinguish quality above 5.6 MBit/s. They also found a strong content dependency in their results. Further, they mentioned that in a scenario where viewers have no access to the original source material, the quality differences at high bitrates may generally not be perceptible.

Van Wallendael *et al.* [36] conducted a subjective study in which UHD was compared against HD, choosing a "striped" test method similar to Li (see above). Based on a set of 31 source sequences, they found that, in general, UHD content was determined to be sharper than HD, and that the likelihood of adequately detecting real UHD sequences ranged from about 40% to 80%, depending on the content. They also note that there are learning effects and biases inherent in the method that may lead to distorted preference ratings, and that the test itself was judged to be difficult.

A dataset for 4K/UHD video was presented in the work of Zhu *et al.* [40], using twelve different source sequences and encoding conditions with HEVC-compressed video at different target bitrates.

Sotelo *et al.* [41] gave an overview of different subjective studies related to video compression and UHD video and also conducted their own test. They noted the limited availability of high-quality UHD video content for test purposes. Ten source sequences were encoded with HEVC at different bitrates. Viewers were seated at two

distinct viewing distances (1.5 H and 2 H). Their results are inconclusive and reveal a large number of outliers. Furthermore, a comparison with another test conducted in a different country [40], but using the same source material, revealed large differences in MOS.

In a 2018 paper, Mackin *et al.* [42] presented a video database containing source material from the authors and third-party content. Using different downsampling algorithms, the authors created test sequences that were subjectively evaluated in a single-stimulus method (i.e., no direct comparisons were made). The authors found that, generally, users preferred UHD-1 over all other (lower) resolutions. However, depending on the resampling algorithm used, no significant differences could be detected between UHD and HD quality. Hence, the authors recommend that for transmission chains with limited bandwidth, the use of HD video instead of UHD may be viable.

To summarize, it has been shown that UHD video can significantly improve the visual quality compared to HD video. However, these differences strongly depend on viewing distance, the possibility of direct comparison against an original sequence, and — most importantly — the chosen content. In many cases, even within the very strict context of a subjective lab experiment, the differences between HD and UHD are not significant.

2.2.3 QoE Models for UHD Video

Of the many algorithms that exist to predict video quality from input signals, few have been specifically designed for UHD video. There are multiple reasons for that, which relate to *how* video resolution is used in those algorithms, if at all.

Most metrics do not use resolution information for predicting quality. A few of the commonly used image quality estimation metrics such as PSNR or SSIM [43], which are also used for video quality estimation, are of this kind. They can be used for images of any resolution, as they do not use resolution as a factor and have not been trained on subjective ratings of human viewers. In principle, the use of such metrics is valid for a video of any resolution as long as no inference on perceived quality is made without proper empirical data to support this conversion. Hence, for example, comprehensively connecting PSNR scores to MOS cannot be done without conducting a subjective study in a pre-determined application context.

There are also metrics that have been trained with subjective rating data on videos of fixed resolutions (or resolution sets), all shown at fixed viewing distances. If these resolutions were smaller than UHD, extending the metrics to support UHD — or any resolution other than the ones they were designed for — would require gathering

new subjective data or empirically supported inference about how a higher resolution might change quality ratings. This usually requires a re-training of the entire metric to achieve good prediction accuracy. For example, a video quality metric developed by Netflix, VMAF [44], was recently extended to predict quality for 4K resolution videos using this approach, with the support of new subjective test data.

ITU-T Study Group 12 is currently conducting a follow-up work on the ITU-T Rec. P.1203 family of standards[3] in a joint initiative with VQEG. The standards define an audiovisual quality model for the prediction of HTTP adaptive streaming quality. Its video component has been developed for HD resolution only. The work item is expected to be finished in late 2019, yielding new video quality models for up to 4K resolution that will be internationally standardized. Subjective tests conducted within the scope of this work have made use of an extensive library of pristine UHD content from different sources.

In a 2013 paper, Hanhart *et al.* [45] evaluated the performance of several common image quality metrics (PSNR, VSNR, SSIM, MS-SSIM, VIF) as well as VQM (a video quality metric) against MOS values obtained from a subjective study. The authors concluded that the accuracy and prediction performance of the metrics was mostly content-dependent, and that all metrics performed equally well once fitted to the per-content subjective results, except for VIF, which performed well for all contents.

In general, it has to be stated that the usefulness of any video quality prediction model — particularly when it comes to UHD — strongly depends on the assumed video consumption scenario. As the viewing distance may play a critical role in whether a human can distinguish quality differences between two clips, one has to know for which assumed viewing distance a given model was developed, and whether it takes viewing distance into account at all.

2.3 The Labeling Effect

2.3.1 Introduction of the Phenomenon

In the most simple terms, the labeling effect is the result of a process during which the information (a label or multiple labels) regarding an entity alters the way the entity is perceived or experienced. The generalization regarding the senses is important, as it may not only affect the handling of visual data, but it can affect, e.g., hearing, smelling, tasting, and the overall sensation of experience as well. Also, it does not

[3]Rec. P.1203: Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport

need to happen real-time, as the modification of the memory of an experience can have an equivalent gravity.

The labeling *effect* is not to be confused with the labeling *theory* of sociology [46, 47]. The labeling theory focuses on the alteration of the individual's self-identity, self-perception and behavior, caused by the labels given by the majority of society, in order to classify deviations from the standard social and cultural norms. In the sole aspect of perception, the difference between these two phenomena is apparent: the labeling theory refers to the change in the way individual I perceives individual I through the labels provided by group G, and the labeling effect refers to the change in the way individual I perceives entity E through the labels provided by group G. As an instance of entity E may include or be equivalent to individual I, the labeling theory strictly in this context could be considered as a real subset of the labeling effect. However, the fundamental difference is that while the labeling theory mainly aims at the analysis of altered behavior (evoked by the changes in self-identity), the labeling effect focuses on perception and experience.

There are specific types of cognitive bias connected to the labeling effect which may greatly enhance its overall impact. During the phenomenon known as confirmation bias [48, 49], an individual holds an idea, a belief, a hypothesis, that affects the acquisition of new information and experience, and also how the existing ones are recalled, remembered. In other words, information is selectively searched for and recalled in order to satisfy, to confirm certain personal ideas. In the context of the labeling effect, it means that the preconceptions generated by the label may affect the perception and the memory of the labeled entity. In this sense, the labeling effect is a special case of the confirmation bias. In many cases, the labeling effect can surface as a misinformation effect [50], especially in the presence of post-event information. New information and new memories can easily lead to retroactive memory interference; changing the way experience is remembered. The labeling effect is also connected to the framing effect [51], as the presentation of the information can significantly influence its processing, and thus, the corresponding decisions. In marketing and commerce, in general, labels advertising discounts create frames for information processing, especially regarding the price. This can also tap into loss aversion [52], as the label may suggest a negative context, like the fear of missing out on a great sale. Finally, the first information — including visual information (i.e., quality) — to affect the individual in a given situation may serve as a point of reference, and thus, it can result in the form of the cognitive bias know as anchoring [53]. In QoE methodology, particularly during quality degradation assessment, this is one of the reasons

why tests tend to begin with the reference quality, which is the best the experiment is able to provide. However, when it comes to the labeling effect, this also allows the personal interpretation of relevant information to become more entrenched in the mind of the individual, as the label becomes a sort of an anchor to which conscious decisions (i.e., quality assessments, in the context of this thesis) shall be tied to.

A classic example of the labeling effect is when a man walks into a classroom, gives a brief guest lecture, and leaves. Then the class is asked about the height of the man. If the man at the beginning was introduced as an internationally recognized expert or as a professor of the field, he could be perceived taller, than if he was introduced as a mere assistant or a fellow student [54].

Pricing is a crucial form of the labeling effect. Price tags fundamentally affect the way the quality of an item is perceived and influence monetary decisions, such as buying the item. The general concept is that the more expensive something is, the better it must be, as there must be a reason why a given item costs more than a different one. This is necessarily present in situations of financial investments and purchases of any level, as the phenomenon builds on our trust in the commercial world. However, a label in such scenarios is not limited to the price tag, but the brand alone can be sufficient to affect perception and experience. Also, the post-purchase experience highly depends on the cost, and the experience itself can become a tool of post-purchase justification. A demonstrative example of consumer price consciousness and the association between price and quality is the work of Sinha *et al.* [55], particularly dealing with private label brands.

Regarding the framing effect, the work of Gachter *et al.* [27] is a great example, as the framing of the registration fee of a scientific conference was investigated. In the experiment, the early-registration fee was presented to half of the Ph.D. students as a discount, while to the other half, a late-registration penalty was communicated. The results show that in case of a discount frame, only 67% registered before the decisive deadline, but 93% registered when a penalty frame was empathized.

As for the perception of information and its quality (i.e., credibility), the work of Rieh *et al.* [56] addressed the domain suffix (.org, .gov, .com and .edu) as an influential factor among scholars. The findings indicate that in many cases, credibility was clearly attributed to the given suffix; for example, when a test participant was asked about the credibility of the information presented on a website, the response was "Absolutely [I trust it] because its an .org" [26].

Generally speaking, the labeling effect itself is enabled by cognitive dissonance reduction between conflicting cognitions [57]. In case of this phenomenon, one of the

cognitions is perception (genuine experience) or the memory of an experience, and the other one is the collection of thoughts, feelings, and memories that can be associated with the label(s). When these cognitions contradict each other to an extent, their dissonant state is reduced or eliminated by changing one of these cognitions, which in case of the labeling effect is the genuine experience or its corresponding memory. It needs to be added that positive reinforcement is possible as well, when cognitions share the same direction (e.g., all associated cognitions agree that a given item has a good quality), but they differ in extent.

Labels in the context of items and services should not be looked at as something inherently bad. Although they do affect our perception and experience, but they also help us navigate in a sea of information. A very common type of label is the list of capability parameters, which describes the most important factors of modern electronic devices, such as household utilities or items of entertainment. With such labels, we can directly compare the capabilities of devices, before making a financial decision.

Let us take, for instance, televisions at a shop. They are usually turned on to show some looping demo content. In case of a UHD/4K-capable TV, high-quality contents are shown in the appropriate resolution in order to "show off" what the display can achieve. In a way, this is actually a tricky subject, as in a regular use case scenario, the user will not use the display to play such demo content all the time. First of all, many contents may have lower spatial resolutions, and even if a video was shot in UHD/4K, a slight noise or defocus may hinder the potential utilization of this capability.

In a typical scenario, a person wishes to purchase precisely one of the given televisions at the shop. There is a specific time constraint for the decision, as time is not unlimited, especially if having an inconclusive visit to the shop is not an option. The final monetary decision is influenced by the perception of quality and by the available information, in forms of labels (capabilities, brands, prices, etc.). Having a large variety of available televisions naturally creates a greater dependency on the labels, beyond the initial filtering. However, there is a two-way cross-influence between the factors of perception and labels: as discussed earlier, perception is affected by labels, but also, even though labels are not practically changed in any way, the processing of labels can be affected by the visual experience (e.g., "For such a nice picture, this price is not that high after all." or "I guess I should not be looking at that parameter in the list as it does not make any difference."). It is important to add that labels

can in fact affect each other as well (e.g., "This is not so expensive for this specific brand.").

In this chapter, the influence of labels over the perceived quality is addressed. More precisely, in the research, it was investigated how the perception of HD and UHD contents shown on a UHD/4K-capable TV was affected by labels indicating their spatial resolutions. The concept of dealing with the labeling effect in QoE-related researches is not novel, as there is already a vast scientific literature on the topic.

2.3.2 The Labeling Effect in Marketing

The E in the abbreviation QoE stands for "Experience". The definition of QoE narrows down the concept of experience for a given "application or service" [58], yet through the generic nature of this word, the labeling effect can be considered for a broader sense of experience. For example, drinking a glass of cold beer may be an experience, and its properties that are not directly linked to the actual taste may affect the drinker's satisfaction as well. The work of Jacoby *et al.* [59] involved price, brand name, and the composition of beer as labels. Verbeke *et al.* [13] investigated the labeling of beef, and Burton *et al.* [10] focused on nutrition reference information in the context of product evaluation. Generally, the intrinsic and extrinsic properties of different purchasable goods as labels were studied by Szybillo *et al.* [9] and Richardson *et al.* [60]. Moreover, brands were also addressed by the work of DelVecchio [61], and Heisey [62] investigated the role of a minimum information environment (involving manipulated information) with regards to the perceived quality of identical clothings; the same sweater was presented with different information cues, resulting in altered perceptions of quality.

The research of Johansson *et al.* [63] focused on the "Made in" labels in the aspect of consumer information processing. Such an experimental aim can be particularly relevant, as consumers tend to associate different levels of product quality with given countries. The work of Hamzaoui *et al.* [19] separated the country of origin into country of design and country of manufacture, and thus, the authors investigated consumer behavior and perceived quality towards bi-national products. The obtained results indicate that the perceived quality of durable goods is influenced more by the country of manufacture, and less by the country of design. The country of design can be influential when the overall value of the product is highly dependent on design (e.g., cars), but even in such cases, the country of manufacture still remains dominant. The findings of Ahmed *et al.* [64] pointed out the correlation between the country of

design and product complexity, and Batra *et al.* [65] highlighted the social signaling value of such products.

Another example of the labeling effect in marketing can be a bottle of wine, with an attractive label and an overall presentation suggesting excellence. It may indeed enhance the experience; however, the opposite can occur as well, in case the margin between the expectations and the actual taste is too large, leading to disappointment [66]. In the particular case of alcoholic drinks, the alcoholic content itself (given in percentage) can also influence the experience. In the work of Masson *et al.* [67], the same wine was provided to test participants in different bottles, and the sole variable was the alcoholic content presented on the label. The study found that the ones suggesting higher alcoholic content were more favored, even though the wine itself was the same. In case of such prominent products, the region of origin (even within a country) can play a particularly significant role, especially in case of well-known vineyards [68]. Going beyond written information, the experiment of Lick *et al.* [20] addressed the connection between the color of the label and the assumed taste. The authors conclude that certain wine tastes are generally associated to very specific colors, e.g., in the study, from the perspective of the test participants, orange labels suggested sweet and fruity flavors, while black labels were associated with woody and earthy flavors.

It is not only the color of the label, but the color of the consumable good itself may also affect expectations. In a rather broad sense, as we now live in an era where the use of artificial food colors is common practice in the production of daily consumables — and therefore, the color of the food may be manipulated without changing the original taste — food color can be a sort of a label as well. Garber *et al.* [69] addressed the effects of food color on the perceived flavor, and also involved misleading and ambiguous labels in the study. The results of the research confirmed the significant effect of both the color and the label of the food (fruity beverage in the instance of the experiment). As the meaning attributed to specific colors may greatly vary between cultures [70, 71], the work, originally carried out in the United States, was extended to a different cultural setting (with different culinary traditions), namely, to India [30]. The obtained subjective data indicated similar findings, emphasizing the effect of the color and the label. One discrepancy compared to the original work was that the Indian test participants were less affected by misleading colors, although the effect was still statistically significant; in fact, the proportion of those who incorrectly identified the purple-colored orange drink as grape was greater than of those who correctly realized that it was orange.

2.3.3 The Labeling Effect in QoE Studies

In the more conventional sense of QoE, the dissertation of Schöffler [72] investigated listening experience. In the experiments, test participants were asked to take "everything" into consideration when assessing the overall audio quality, and they were explicitly told that the stimuli differed in quality. Emphasizing the difference between the quality levels of the stimuli can lead to a preconception stating that "there should be a difference", inducing variations in the listening experience, and thus, in subjective ratings of stimuli which would have none otherwise. Music excerpts of various genres were used as audio stimuli, and test participants also had to self-assess the impact of the songs (and their performers) on their own quality ratings. The song-related information served as a label, especially since test participants were specifically given the task to consider it for the overall subjective assessment. Beyond the presence of the labeling effect, it is noteworthy that certain participants rated lower-quality stimuli higher than undistorted, high-quality stimuli, due to their prior experiences (e.g., the low-quality music excerpts reminded a participant of the pleasant memories of concerts and festivals in the past).

As shown by the previous example, in subjective tests of multimedia quality, the labeling effect may take a foothold, as almost any information can influence quality ratings provided by the test participants. However, certain studies particularly aim at this phenomenon, in order to discover the magnitudes of achievable distortions; how much the labeling effect can distort subjective test results. Many researches with such goal involve mock-up scenarios and stimuli, in which the exact same multimedia quality is provided through a given content, but the associated labels differ.

The experiment of Lamm *et al.* [8] evaluated simulated search engines biased with — as the authors phrase — "manipulated" user expectations. One group of the test participants were informed prior to quality assessment that the search engine was actually an expensive professional search system, while a different group was told that it was only a mere student project. This separation was repeated for two distinct levels of objective system performance, therefore, the participants were clustered into four groups. The results conclude no significant difference based on user expectations, but note that a given test participant was only provided one specific label, and had nothing to compare to. In a follow-up work [73], the authors extended their methodologies with direct service comparison, resulting in eight groups, as all combinations were investigated: a group was first provided a search engine with either an objectively good or bad quality, labeled with one of the previous descriptions, and then another search engine was provided with either good or bad quality — thus, half of the test

conditions included identical stimuli — but it was given the other label. Again, the labels always differed, so one was labeled expensive and professional, while the other one was the work of a student with unknown quality. Although the results do indicate the significant influence of the labeling effect, it is also shown that expectations maybe be overwritten by performance experience over time. These findings correlate with the conclusions of Szanja *et al.* [15], stating that certain expectations may fade as time progresses.

Bouchard *et al.* [1] investigated the sense of presence for virtual reality. Although test participants were immersed in a synthetic environment, they were informed before the test that they would be immersed in a real-time replica of an actual room, containing a real mouse in a cage. A different group of the test participants was told the same thing, but without the real-time component of immersion. In reality, every test participant was immersed in the same synthetic environment. The subjective results indicate a significant difference in the sense of presence of the two groups. Furthermore, the study was repeated with the use of simultaneous functional magnetic resonance imaging (fMRI), indicating significant differences in brain areas that are related to immersion and presence, and thus, concluding that the misleading information resulted in a genuinely higher sense of presence.

In the contributions of Sackl *et al.* [16, 74] and Kara *et al.* [75, 76], the label was the type of connection. In both researches, the perceived quality was measured in a mock-up scenario, where the performance of wireless and wireline connection did not differ at all. In fact, in some of these works, there was not even any multimedia transmission, as the stimuli were played from the local storage of the device. By doing so, identical quality was ensured, yet the subjective scores significantly differed. It needs to be noted that the direction of such distortion (whether it enhances or degrades user experience) is not evident; it depends on the test participant. While many test participants had notably lower degrees of QoE in the wireless test cases, others actually perceived the wireless to be much better. Sackl *et al.* also addressed the Willingness to Pay (WTP) [77,78], as labels indeed affect the customers' monetary decisions. User expectations were also in the direct focus of the research [79], due to the socio-psychological reasons mentioned earlier.

In a joint work of the authors [80], the label was the brand of the mobile end-user device. Although each device played the same stimuli locally in an unimpaired quality, most of the test participants perceived visual degradations (i.e., playback jitter, tearing, blurred pixel zones, black/missing pixels, etc.) on the unfavored devices. The

effect of smartphone brands on user experience was also investigated by the thesis of Ebbing [81].

In the work presented in this chapter, subjective tests were carried out with and without labels. When labels were used, certain test conditions were akin to the mock-up methodologies of the prior works, as identical video stimuli were compared with different labels. For example, both stimuli were either identical HD or UHD videos, but the labels stated that one of them was HD while the other one was UHD. These were compared with test conditions where the stimuli genuinely differed, and where the stimuli were identical but the labels reflected this fact rightfully. Also, each and every test condition was subjectively assessed without labels as well. The test conditions and all the important parameters of the experimental setup are detailed in the following section.

2.4 Experimental Setup

The main scientific aim was to investigate the impact of the labeling effect and the rating scale on the perceived quality of UHD services. In this section, the common attributes of the four studies are presented, and their differences in the utilized test protocols and in the questionnaires are highlighted. The four studies are:

- Labels shown to subjects, 7-point rating scale
- Labels shown to subjects, 3-point rating scale
- Labels not shown to subjects, 7-point rating scale
- Labels not shown to subjects, 3-point rating scale

All studies used the same contents but were run with a different set of subjects. This way, it could be ensured that subjects would not learn about how the test paradigms differed, which could have introduced an avoidable bias. The remainder of this section will detail the experimental setup from a technical and experimental design perspective.

2.4.1 Research Environment and the UHD Display

All the subjective tests were carried out on a Samsung 55-inch JU6400 6 Series Flat UHD/4K Smart LED TV[4]. The display, and thus, the tests themselves were located

[4]http://www.samsung.com/uk/tvs/uhd-ju6400/UE55JU6400KXXU/

in an isolated laboratory environment of Kingston University, in which test participants suffered no audiovisual distractions. Based on the guidelines of the ITU-R Rec. BT.2022, the test participants were seated at a distance of 1.6 H from the display, which in case of the aforementioned 55-inch TV, corresponded to 110 cm. The angle of vision was zero; test participants viewed the display precisely from the middle.

Test subjects participated individually, separately, as a single position of observation was defined. This is also in alignment with the approach to provide a unique, randomized stimulus order for each and every test participant, making the scenario of multiple observers unavailable.

The series of subjective tests at hand solely focused on visual quality, hence audio was excluded from the research. This means that no stimulus contained audio data, no sound was generated by the speakers of the television during the tests, and no external audio gears were worn by the test participants. The general lack of audio applies to each and every experiment presented in this thesis.

2.4.2 Rating Scales

In all experiments, test participants compared video stimulus pairs, in terms of the overall visual quality. In order to take the "expressive power" of subjective comparison scales into consideration, two scales were used with identical rating concept but with different level of detail. One was a simple 3-point ("*Worse*", "*Same*", "*Better*") comparison scale, which enabled a basic discrimination of video quality. The other one was the ITU-R Rec. BT-500.13 7-point ("*Much worse*", "*Worse*", "*Slightly worse*", "*Same*", "*Slightly better*", "*Better*", "*Much better*") comparison scale, which also lets test participants express the magnitude of the experienced difference.

2.4.3 Investigated Test Conditions

The comparisons between HD and UHD videos were either resolution transitions or self-comparisons. During transitions, the first video was HD, and the second one was UHD (or vice-versa). Self-comparison means that both stimuli in the pair had the same resolution (i.e., HD and HD, or UHD and UHD). Since there were two resolutions, this means that there were a total of four possible comparisons.

In every video pair, test participants were shown the same original video content. Each test condition was applied to multiple sources. This also means that the videos

Table 2.1: Investigated test conditions.

ID	Video 1	Video 2	Video 1 label	Video 2 label
1	HD	UHD	HD	UHD
2	UHD	HD	UHD	HD
3	UHD	UHD	HD	UHD
4	HD	HD	HD	UHD
5	UHD	UHD	UHD	HD
6	HD	HD	UHD	HD
7	UHD	UHD	UHD	UHD
8	HD	HD	HD	HD

in self-comparisons not only had the same resolution, but they were in fact exactly the same.

For the tests with labels, these conditions were paired with the four possible combinations of labels attached to the sequences. This means that the label could either indicate the correct resolution of the clip to follow or purposely deceive the user into thinking that another resolution than the one actually shown would be presented. In the latter test conditions, the displayed stimuli in the pair had the same resolution, however, the labels suggested transitions.

Table 2.1 introduces the 8 test conditions that were investigated. Conditions 1 and 2 were transitions with correct labels, conditions 3, 4, 5 and 6 were the possible combinations of self-comparisons with misleading labels, and 7 and 8 were self-comparisons with correct labels.

In case of the experiment where no labels were present, the same test conditions were included but without labels. This means that conditions 3, 5 and 7 were practically identical, and 4, 6 and 8 did not differ in any way either.

2.4.4 Source Sequences and Test Stimuli

In the experiments, 8 different source videos (SRC) were used, 2 sequences each from the 4 original contents: "*Big Buck Bunny*", "*Sintel*" and "*Tears of Steel*" from Blender, and "*El Fuente*" by Netflix, all at UHD resolution (3840 × 2160). The "*El Fuente*" video sequences are licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License[5] and the videos from Blender are licensed under the Creative Commons Attribution 3.0[6]. They are distributed within the scientific community as test sequences for evaluation via Xiph.org Test

[5]http://creativecommons.org/licenses/by-nc-nd/4.0/
[6]http://creativecommons.org/licenses/by/3.0/

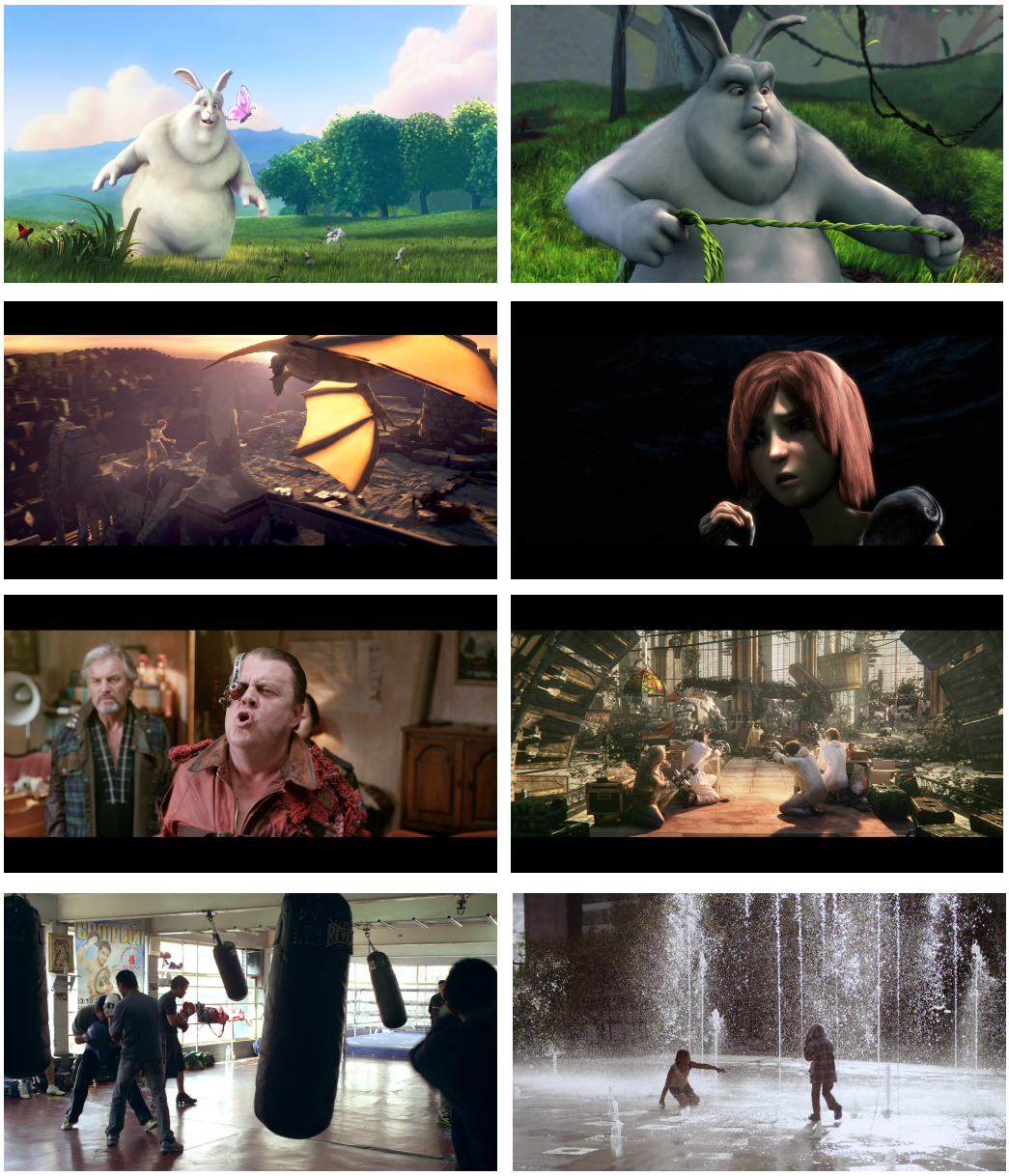


Figure 2.1: UHD: Source videos used in the experiments.

Media[7]. In Figure 2.1, one row refers to one content and shows one representative frame from each of the selected 10-second parts of the videos.

The primary aim of choosing the original contents and cutting the videos to particular scenes was to achieve diversity in content genre, motion descriptors, saturation, brightness and level of image detail, which was one of the most important parameters. Two contents were CGI animation ("*Big Buck Bunny*" and "*Sintel*"), which was found crucial to investigate in this study, as computed graphics can enable a high

[7]https://media.xiph.org/

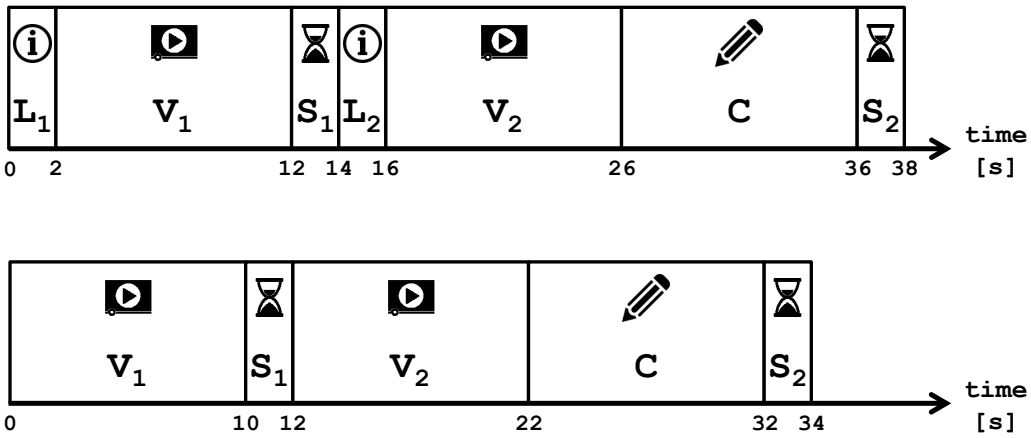


Figure 2.2: UHD: Visualization order of the experiment with labels (top) and without labels (bottom).

level of visual detail when rendered at the target resolution. As an example, the hair of the character in the zoomed-in shot of SRC-4 from "*Sintel*" was highly detailed. At the same time, SRC-2 from "*Big Buck Bunny*" mainly had smooth rendered surfaces.

The test stimuli were created by using the available uncompressed frame sequences, which were merged into video files. A frame rate of 24 was used in every video. This applied to HD videos as well, along with the spatial resolution. However, they were not identical to the UHD ones, as they were first downscaled to HD (1920×1080), and then they were upscaled back to UHD. These two samplings were both performed using a Lanczos filter.

One could immediately argue that excluding the usage of practical lossy compression methods may reduce the realism, the real-life validity of the study, and therefore, this decision can make the research feel more theoretical. This argument is, of course, valid in a certain way, and it needs to be noted that this valid point has been considered during the phase of experimental design. The reason why compression was not included in the final design is that its potential visual impact would have added another variable that could have influenced the subjective ratings. As the sole focus of the experiment was on the influence of labels, using compression is more of a next step, a future continuation of this line of research, rather than a fundamental component from tile one.

2.4.5 Test Protocols with and without Labels

As mentioned before, the four experiments differed in terms of the label: two experiments explicitly identified the spatial resolution of the next video stimulus through

labels, while the others did not.

Figure 2.2 shows the chronological structure of the stimuli as they were shown to the subjects; it indicates when and what was shown to every test participant. Every stimulus pair included two videos, referred to as Video 1 and Video 2, and the task was to compare Video 2 against Video 1.

In case of labeled videos, first the label of Video 1 (L_1) was displayed for 2 seconds, followed by the 10-second Video 1 itself (V_1). The same was shown for Video 2 (L_2 and V_2), but before that, a blank separation screen was on for 2 seconds (S_1). The stimuli were followed by a 10-second period for the subjective comparison (C), during which a short text was displayed on the screen, asking the test participant to cast the vote on the evaluation sheet. The protocol for each comparison ended with another 2-second blank separation screen (S_2), creating a brief pause between the different stimulus pairs.

In the tests where no labels were present, the experimental protocol was the same, but without the 2-second resolution identifiers before the stimuli. This resulted in a minor difference between the total test durations of the experiments. With labels, the test took roughly 40 minutes to complete, while it was approximately 5 minutes less when no labels were present.

In both experiments, the 8 test conditions were applied to all 8 sources. This means that each test participant made 64 comparisons, and thus, observed 128 video stimuli.

As it has already been stated earlier, the order of the test stimuli was randomized; it was unique for every test participant. Randomization was performed in a way that avoided content repetition so that adjacent stimulus pairs always had different sources.

2.4.6 Pre- and Post-Experiment Questionnaires

At the beginning of the experiment, before the training phase, the test participants had to fill out a pre-experiment questionnaire, which was the same for both types of experiments (with and without labels). First, the basic demographic information was gathered, such as age and gender. This was followed by three questions on prior experience and familiarity with UHD/4K, as shown in Table 2.2. Also, the test participants were subject to screenings based on the Snellen charts and Ishihara plates, in order to ensure the validity of the research; as stated in Chapter 1, these procedures applied to all the experiments presented in this thesis.

Table 2.2: The pre-experiment questionnaire.

Have you ever heard of “Ultra HD”, “UHD” or “4K”? — Yes, and I could explain what it means. — Yes, but I could not explain what it means. — No, never.
Have you seen a video in UHD / 4K resolution yet? — Yes. — No. — I do not know.
Do you possess a device with UHD / 4K resolution? — Yes. — No. — I do not know. — I do not wish to answer.

Table 2.3: The post-experiment questionnaire.

Common: — How mentally demanding was the task? — How physically demanding was the task? — How hurried or rushed was the pace of the task? — How successful were you in accomplishing what you were asked to do? — How insecure, discouraged, irritated, stressed, and annoyed were you?
Tests with labels: — After having participated in the test, would you say that 4K video is better than HD video? (Yes. / No. / I don’t know.) — When comparing HD and 4K, what is the main difference for you?
Tests without labels: — In case the videos in the pairs differed, what was the main difference for you?

The post-experiment questionnaire, as shown in Table 2.3, included five questions that were asked in both experiments. These were to be answered on a quasi-continuous scale ranging from -10 to 10 (without 0), where positive numbers (right part of the scale) represented high mental and physical demand, rushed test pace,

lack of confidence in ratings, irritation, stress and so on and so fourth, while negative numbers (left part of the scale) were used to express the opposite. In this context, the opposite of "rushed" is "not rushed", and not "too slow".

In all experiments, the test participants were asked whether they considered the UHD stimuli to be generally better than the HD, and more importantly, they were asked about what they thought the source of the difference was. Their answers were collected in written fashion.

2.5 Results

2.5.1 Panel and Pre-Experiment questionnaire

2.5.1.1 Tests with Labels

A total of 30 people took part in the experiments with a label shown before each video stimulus. The test participants were from an age range between 18 and 39, and the average age was 25. The subjective test was completed by 23 males and 7 females.

From the 30 test participants, 8 knew what UHD is, 16 heard about the term and 6 had not heard about UHD prior to the experiment.

The number of participants who had seen UHD videos before the subjective test was 8, while 13 had not, and 9 were unsure about the answer.

At the time of the research, no test participant possessed a UHD-capable device. To be more precise, according to the questionnaire, none of them could state owning such device, as 6 were unsure whether what they had were UHD-capable or not, and 24 were certain that their devices were not UHD-capable.

2.5.1.2 Tests without Labels

Similarly to the tests with labels, a total of 30 people took part in the experiments that did not contain labels regarding the resolution of the video stimuli. The test participants were from an age range between 20 and 40, and the average age was 25. The subjective test was completed by 23 males and 7 females.

From the 30 test participants, 10 knew what UHD is, 16 heard about the term and 4 had not heard about UHD prior to the experiment.

The number of participants who had seen UHD videos before the subjective test was 17, while 10 had not, and 3 were unsure about the answer.

According to the questionnaire, 7 test participants possessed UHD-capable devices, while 20 did not, 2 were unsure whether what they had were UHD-capable or not, and 1 person did not wish to answer the question.

While the experiments with labels were conducted in 2016, the experiments without labels were carried out a year later in 2017. Although there was no notable difference in "UHD awareness" among the test participants of the sets of experiments, there was an apparent rise in prior UHD video experience and in the possession of UHD-capable devices.

2.5.2 Tests with Labels

The results of the tests where labels were present during the experiment are shown on Figures 2.3a and 2.3b, with histograms of the ratings for the 3-point and the 7-point scale, respectively. There are 8 investigated test conditions, as defined in Table 2.1.

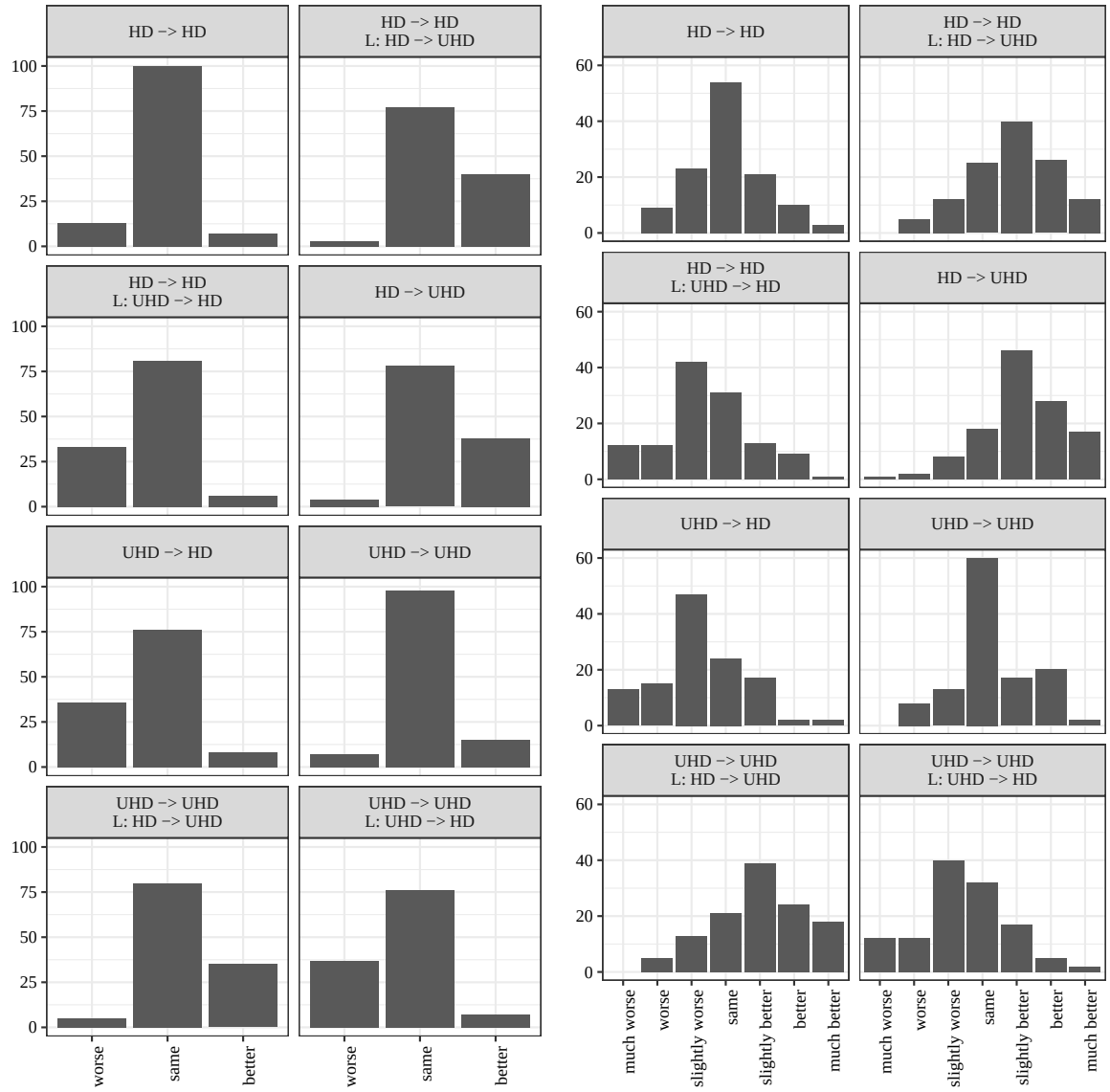
2.5.2.1 3-point Scale

When identical video stimuli were shown to the test participants, accompanied by identical labels, in case of both UHD and HD videos (conditions 7 and 8, respectively), the provided ratings clearly reflected the lack of perceived difference. However, when misleading labels were introduced for these identical pairs (conditions 3, 4, 5, and 6), roughly a third of the given scores indicated that the stimulus with the UHD label was better. On its own, this already implicates a strong presence of the labeling effect in the obtained results. However, when these scoring patterns are compared with the results of genuinely different stimuli with correct labels (conditions 1 and 2), a peculiar similarity is revealed.

These observations are reinforced by the statistical analysis of the investigated conditions, as shown in Table 2.4. In order to evaluate whether there were statistically significant differences between the ratings depending on the shown conditions, first an ANOVA was calculated using condition as an independent variable and the ratings of the test participants as dependent variables. The ANOVA ($df = 7$, $p = 0.00$) indicated a significant impact of the conditions.

To then investigate individual differences between two conditions c_1 and c_2, the Tukey HSD (T), Holm (H), and Bonferroni (B) multiple comparison tests were conducted. A condition pairing was considered to have a significant influence on the ratings if the Tukey HSD p-value was below 0.05.

First of all, there is no significant difference between conditions 7 and 8, as they both show that the test participants found the identical videos with identical labels to

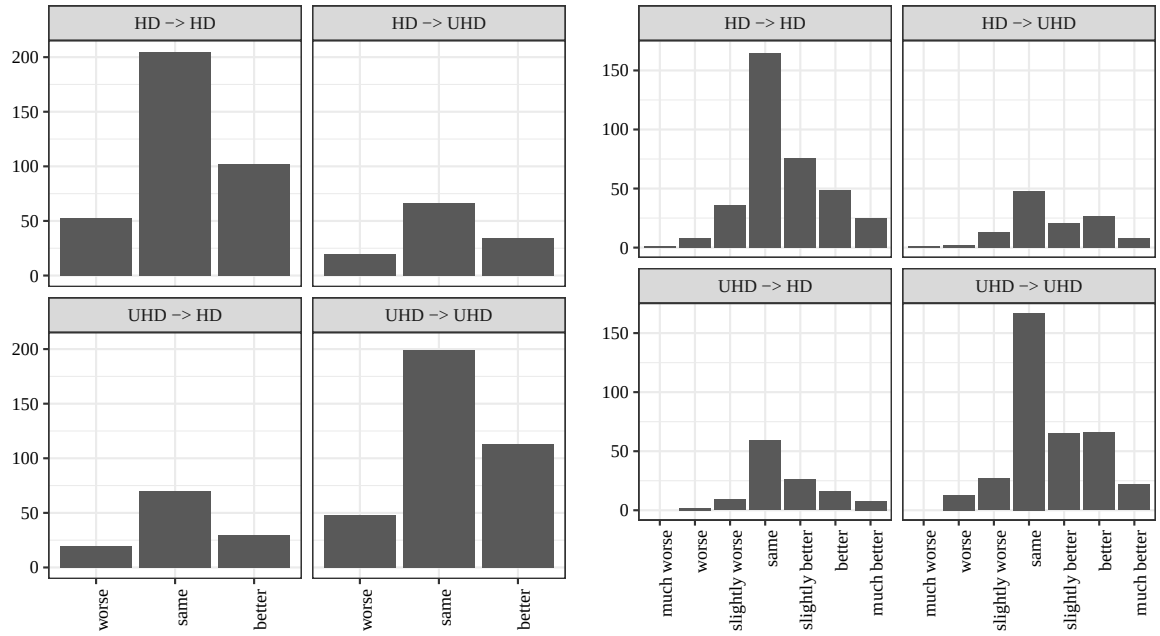


(a) Test with labels, 3-point rating scale.

(b) Test with labels, 7-point rating scale.

Figure 2.3: UHD: Histogram of test ratings with labels.

be perceptually identical. Second, condition 5 received significantly different ratings than condition 7, and the same applies to conditions 4 and 8. These obtained results mean that even though the stimuli did not differ in between these conditions, the ratings were still heavily influenced by the labels. Although the differences between conditions 3 and 7, and conditions 6 and 8 are measurable and also seem apparent from the histogram, they are not statistically significant. Third, there is no statistical difference between conditions 1, 3, and 4, and between conditions 2, 5, and 6. Therefore, it can be stated that the test participants perceived the identical videos the same



(a) Test without labels, 3-point rating scale. (b) Test without labels, 7-point rating scale.

Figure 2.4: UHD: Histogram of test ratings without labels.

way as the ones with actual visual differences. This indicates that the influence of the labeling effect on the subjective scores was evidently greater than the cognition of perception.

2.5.2.2 7-point Scale

The rating tendencies for the 7-point scale were similar compared to the results obtained for the 3-point scale. However, the main difference here was that as test participants were given a greater freedom in the expression of quality comparison, which resulted in more scoring deviation. Furthermore, the usage of "slight" ratings consistently dominated the quality assessment for the test conditions where resolution change was indicated through the labels.

The statistical analysis for the 7-point scale is provided in Table 2.5. Here as well, first an ANOVA was conducted to check the general impact of conditions, which turned out significant ($df = 7$, $p = 0.0$). The Tukey, Holm, and Bonferroni tests were conducted in the same fashion as with the 3-point scale.

Similarly to the 3-point scale, there is no significant difference between conditions 7 and 8. There are statistically significant differences between conditions 3, 5 and 7, and between conditions 4, 6 and 8, resulting in stronger corresponding conclusions compared to the 3-point scale. Finally, there is no statistical difference between

Table 2.4: Statistical analysis of the investigated test conditions (c_1 and c_2), rated by a 3-point scale, in the presence of labels. The p-values of Tukey HSD (T), Holm (H), and Bonferroni (B) multiple comparisons are given, along with significance (S).

c_1	c_2	**T**	**H**	**B**	**S**
1	2	0.000	0.000	0.000	yes
1	3	1.000	1.000	1.000	no
1	4	1.000	1.000	1.000	no
1	5	0.000	0.000	0.000	yes
1	6	0.000	0.000	0.000	yes
1	7	0.022	0.012	0.027	yes
1	8	0.000	0.000	0.000	yes
2	3	0.000	0.000	0.000	yes
2	4	0.000	0.000	0.000	yes
2	5	1.000	1.000	1.000	no
2	6	1.000	1.000	1.000	no
2	7	0.000	0.000	0.000	yes
2	8	0.096	0.052	0.145	no
3	4	0.987	1.000	1.000	no
3	5	0.000	0.000	0.000	yes
3	6	0.000	0.000	0.000	yes
3	7	0.096	0.052	0.145	no
3	8	0.000	0.000	0.000	yes
4	5	0.000	0.000	0.000	yes
4	6	0.000	0.000	0.000	yes
4	7	0.006	0.003	0.007	yes
4	8	0.000	0.000	0.000	yes
5	6	1.000	1.000	1.000	no
5	7	0.000	0.000	0.000	yes
5	8	0.047	0.025	0.064	yes
6	7	0.000	0.000	0.000	yes
6	8	0.132	0.061	0.213	no
7	8	0.632	0.524	1.000	no

conditions 1, 3 and 4, and between conditions 2, 5 and 6, just as in case of the 3-point scale.

2.5.3 Tests without Labels

The results of the tests where labels were not present during the experiment are shown on Figures 2.4a and 2.4b, with histograms of the ratings for the 3-point and the 7-point scale, respectively. There are 4 distinct test conditions, since labels were not shown. All test conditions given in Table 2.1 were separately assessed, but they

Table 2.5: Statistical analysis of the investigated test conditions (c_1 and c_2), rated by a 7-point scale, in the presence of labels. The p-values of Tukey HSD (T), Holm (H), and Bonferroni (B) multiple comparisons are given, along with significance (S).

c_1	c_2	**T**	**H**	**B**	**S**
1	2	0.000	0.000	0.000	yes
1	3	0.971	1.000	1.000	no
1	4	0.726	0.709	1.000	no
1	5	0.000	0.000	0.000	yes
1	6	0.000	0.000	0.000	yes
1	7	0.000	0.000	0.000	yes
1	8	0.000	0.000	0.000	yes
2	3	0.000	0.000	0.000	yes
2	4	0.000	0.000	0.000	yes
2	5	0.951	1.000	1.000	no
2	6	0.962	1.000	1.000	no
2	7	0.000	0.000	0.000	yes
2	8	0.000	0.000	0.000	yes
3	4	0.999	1.000	1.000	no
3	5	0.000	0.000	0.000	yes
3	6	0.000	0.000	0.000	yes
3	7	0.000	0.000	0.001	yes
3	8	0.000	0.000	0.000	yes
4	5	0.000	0.000	0.000	yes
4	6	0.000	0.000	0.000	yes
4	7	0.006	0.002	0.007	yes
4	8	0.000	0.000	0.000	yes
5	6	1.000	1.000	1.000	no
5	7	0.000	0.000	0.000	yes
5	8	0.003	0.001	0.003	yes
6	7	0.000	0.000	0.000	yes
6	8	0.002	0.001	0.002	yes
7	8	0.906	1.000	1.000	no

are clustered in the analysis (3, 5 and 7; 4, 6 and 8), as they were identical not only in content, but from the perspective of the test participants as well.

2.5.3.1 3-point Scale

For all four investigated test conditions, the histograms of the ratings indicate similar tendencies: the lack of visual difference is the dominant score, followed by assessing the second stimulus better (with roughly half as many ratings), and the preference of the first stimulus received the fewest scores (again with roughly half as many ratings

as the previous option).

Based on an ANOVA conducted between conditions and ratings, we could see no significant impact for the 3-point scale ($df = 7$, $p = 0.829$). All Tukey, Holm, and Bonferroni values consequently indicate the lack of any statistically significant difference between any given two test conditions, hence the detailed results table is omitted. This means that there was no evident visual difference between the UHD video stimuli and the upscaled HD videos. This conclusion is reinforced by the similarity in scoring between conditions 1 and 2, which are technically the opposites of each other.

As for the repeated scoring distribution, it can be linked to a simple assessment bias due to the lack of clear visual differences. Although it was not emphasized during the training phase that the stimuli *will* visually differ, with 64 paired comparisons, it is not difficult for a test participant to get the feeling that there *should* be a difference. Furthermore, as the visualization on the large UHD TV was generally pleasing and there were no additional impairments implemented, it was easier to rate the second stimulus to be the better one, via memory bias targeting the first one.

2.5.3.2 7-point Scale

The quality assessment of the test conditions using the 7-point scale resulted in similar but more deviating tendencies compared to the usage of the 3-point scale. Here, the ANOVA also shows no significant impact of the condition on the ratings ($df = 7$, $p = 0.693$). The Tukey, Holm and Bonferroni statistical analysis also conclude that lack of significant differences in the subjective scores and hence are not shown in detail.

2.5.4 Content Dependency

In this statistical analysis, similar multiple comparisons were carried out, using source as an independent variable and the ratings of the test participants as dependent variables. The analysis was first run on the data grouped by experiment type and scale type (112 comparisons) and then also grouped by condition (896 comparisons). The results generally indicate that the content did not play a significant role in the subjective assessment, as in the first analysis, only 2 out of 112 comparisons concluded significant differences, and for the second one, this was 1 out of 896. These differences were both measured for the experiment without labels, using the 3-point scale. Without grouping by conditions, the ratings of SRC-5 (first clip of *Tears of*

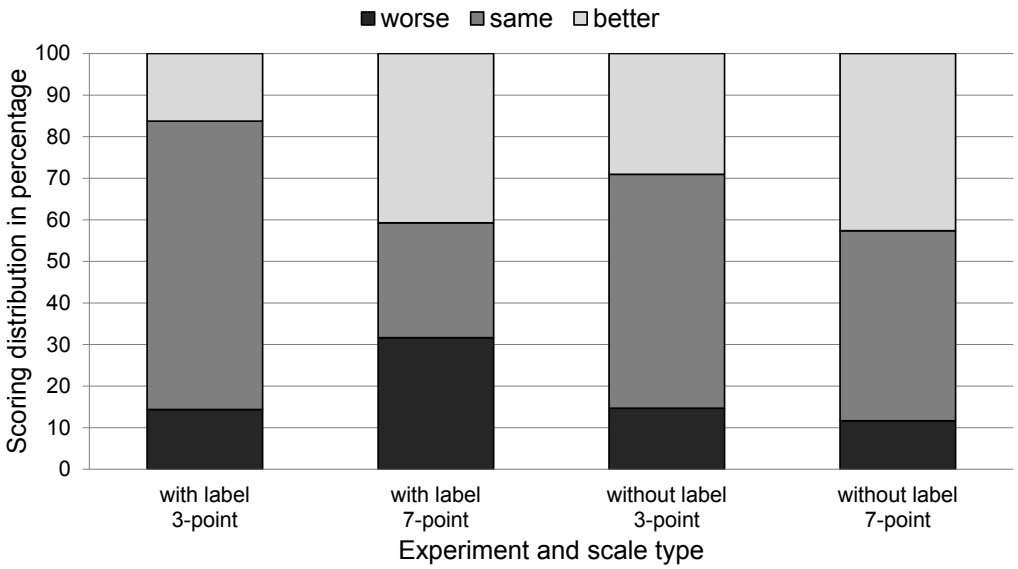


Figure 2.5: UHD: Rating distributions mapped onto the 3-point scale.

Steel) significantly differed from SRC-6 (second clip of *Tears of Steel*) and SRC-7 (first clip of *El Fuente*). When the sources were separately analyzed for each test condition, SRC-5 and SRC-7 showed a statistically significant difference.

2.5.5 Rating Scale Correspondence

For the analysis presented in this subsection, the ratings obtained via the 7-point scale were collapsed and mapped onto the options of the 3-point scale, in order to demonstrate the differences in scale usage. This means that the three positive and three negative comparisons of the 7-point scale were converted into one option each (e.g., both "*Slightly better*", "*Better*" and "*Much better*" become "*Better*"). The resulting scoring distribution is presented on Figure 2.5. It is clearly shown that there are strong differences that were observed previously on Figures 2.3a and 2.3b; during the experiment with labels present, while the 3-point scale produced 69.37% of the ratings to indicate the lack of visual difference between the stimuli, the corresponding value with the 7-point scale was only 27.6%. A similar tendency is present for the results of the experiment without labels, but the extent is less intense. Here, only the positive scores increased due to the aforementioned assessment bias.

2.5.6 Per-subject Rating Behavior

Let us now have a look at the ratings of the test subjects individually. The histograms for the tests with and without labels are shown on Figures 2.6a, 2.6b, 2.7a, and 2.7b.

Table 2.6: Criteria of rating correctness (based on the resolution of V) and compliance with labels (based on label L).

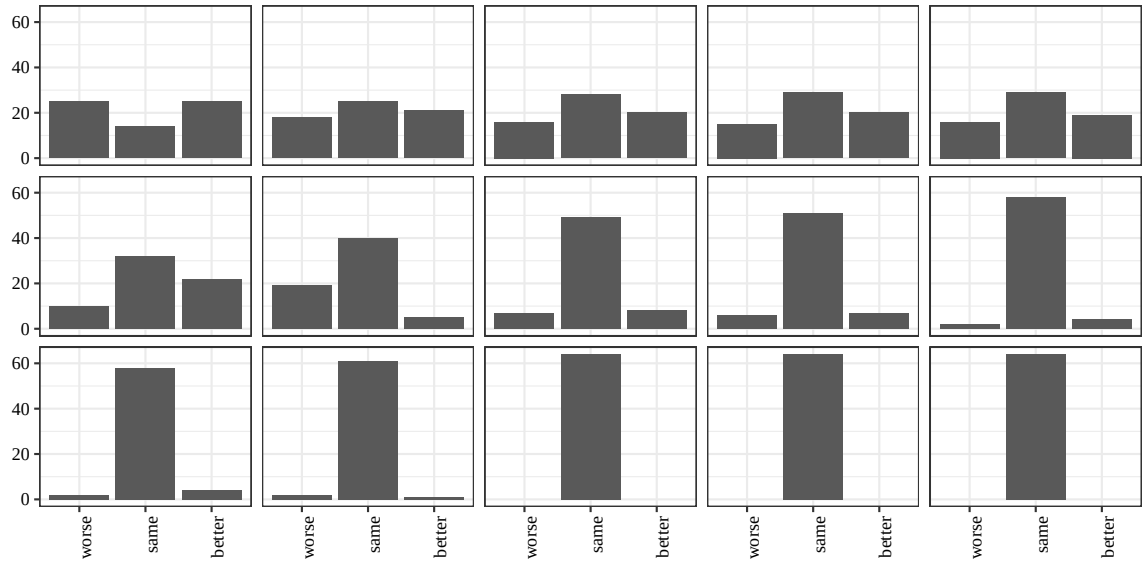
V_1/L_1	V_2/L_2	**Rating**
HD	HD	Same
HD	UHD	Slightly better, Better or Much better
UHD	HD	Slightly worse, Worse or Much worse
UHD	UHD	Same

Certain rating behavior extremes stand out at first glance, such as the scores of test participants who did not distinguish *any* stimuli in the pairs, and therefore, provided 64 identical ratings. This applied to seven individuals from the entire pool of test subjects (more than 10%). The opposite is worth mentioning as well, where test participants avoided this specific rating option (no stimuli pair was assessed as the same). This was only present for the 7-point scale, which provided three times as many options to rate visual differences.

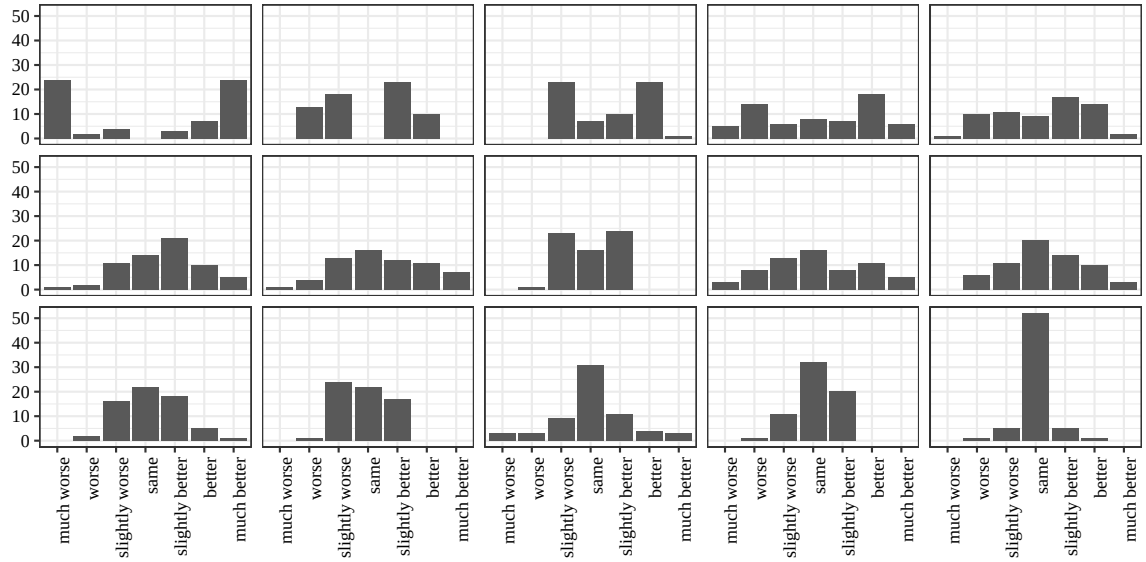
These individual results can be matched with the test stimuli and with the labels. Matching subjective ratings with the stimuli tells us the achieved rating correctness, that is, the correlation between what resolution was used and how it was reflected in the scores. Matching subjective ratings with the labels indicates a sort of obedience to the labels, as it shows how much the test participants agreed with what the labels suggested. More formally, we can classify a comparative rating as correct if the relation between the objective indicators of the stimulus pair (i.e., the actual resolutions of the videos) is essentially the same as the subjective score. For example, if the first video in the stimulus pair (V_1) was HD, and the second one (V_2) was UHD, then the rating options "*Slightly better*", "*Better*" and "*Much better*" were correct in this sense. Regarding compliance, we can classify a comparative rating as compliant if the relation between the labels of the stimulus pair (i.e., the suggested resolutions of the videos) is essentially the same as the subjective score. For example, if the label of the first video in the stimulus pair (L_1) was HD, and the second one (L_2) was UHD, then the rating options "*Slightly better*", "*Better*" and "*Much better*" indicate compliance. The full set of criteria of rating correctness and compliance is shown in Table 2.6.

2.5.6.1 Rating Correctness

The results of the rating correctness analysis are presented on Figure 2.8. As detailed in Chapter 1, a given column represents the calculated value of rating correctness for



(a) Test with labels, 3-point rating scale.

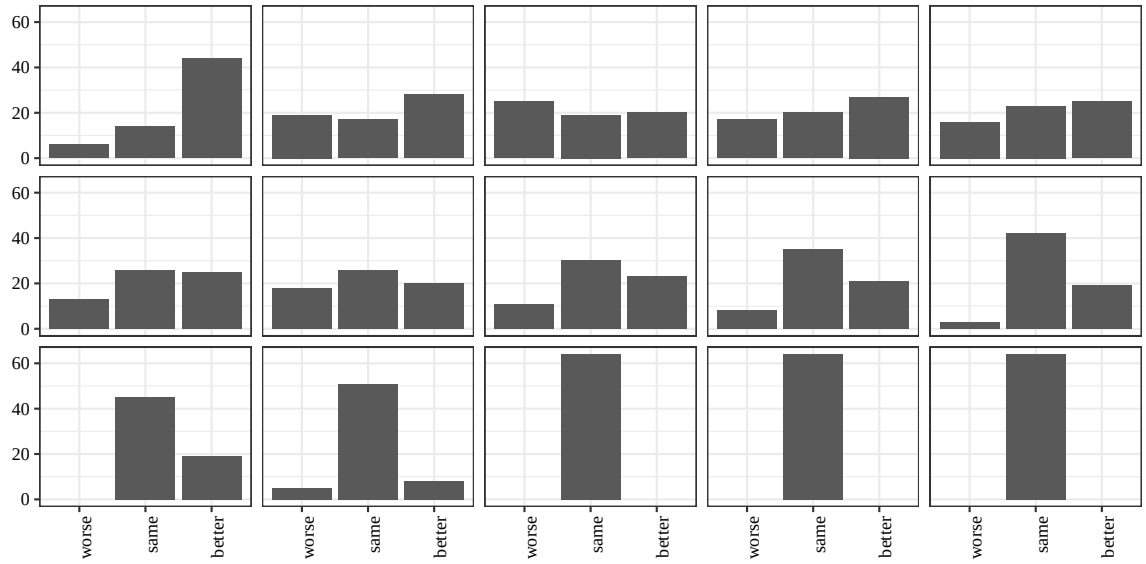


(b) Test with labels, 7-point rating scale.

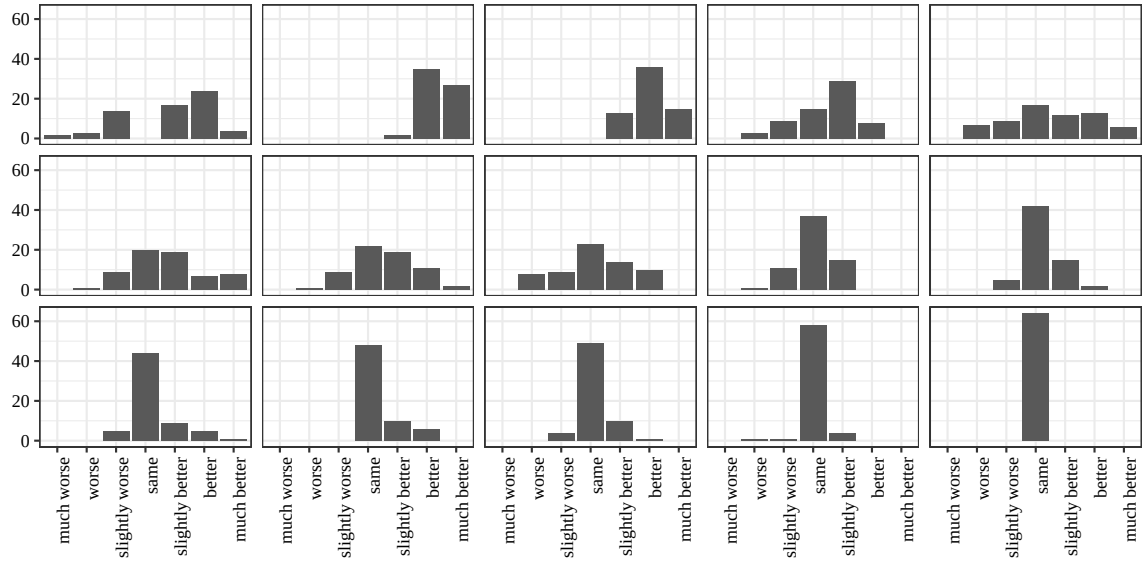
Figure 2.6: UHD: Histogram of test ratings with labels, per subject.

the subjective scores of one specific test participant. These values are categorized by experiment and scale type (15 per category), and they are shown ascending order from left to right.

When labels were present during the experiment, the 3-point and the 7-point scale produced average values of 61.04% and 40.52%, respectively. When labels were not included in the experiment, the corresponding 3-point and 7-point scale averages were 47.71% and 41.56%. In both experiments, the 3-point scales achieved higher percent-



(a) Test without labels, 3-point rating scale.



(b) Test without labels, 7-point rating scale.

Figure 2.7: UHD: Histogram of test ratings without labels, per subject.

ages of rating correctness. This is partially due to the fact that it did not enable the rating freedom of the 7-point scale, and thus, more subjective assessments deemed the stimuli to be the same, as shown on Figure 2.5. As 75% of the stimulus pairs contained identical videos, those test participants who used only the corresponding option in the scales evidently achieved a rating correctness of 75%. Furthermore, this was, in fact, the highest level of measured rating correctness for both experiments and both scale types.

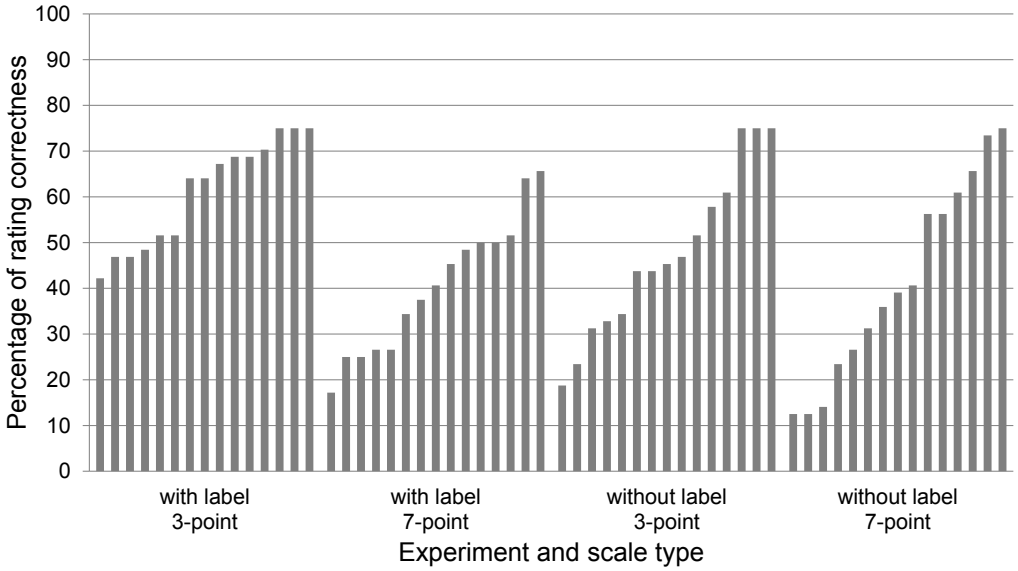


Figure 2.8: UHD: Percentage of rating correctness.

2.5.6.2 Compliance with Labels

The results of the analysis on the compliance with labels are presented on Figure 2.9. The same method of data visualization is used here as for rating correctness. However, only half of the test participants were involved in this analysis, as the other half participated in tests without labels. While the average compliance for the 3-point scale was 43.43%, the corresponding value for the 7-point scale was 59.27%. Again, the 7-point scale made it possible for the test participants to indicate smaller differences via the options "*Slightly worse*" and "*Slightly better*". This fact is quite relevant to this analysis since it supports the marking of the perceived differences evoked by cognitive bias. Furthermore, two participants achieved 100% compliance, which means they never disagreed with the labels. As for those participants who only used the middle option in the test, their compliance value was 25%, since only 25% of the presented labels suggest the lack of difference.

2.5.7 Post-Experiment Questionnaire

2.5.7.1 Common Questions

In this analysis, negative and positive scores (zero was not an option) are represented by dark and light columns, respectively, and the results are separately shown for the two experiments. Extreme scores are towards the edge of the figures; for example, the number of ratings indicating the lowest level of mental demand is represented by

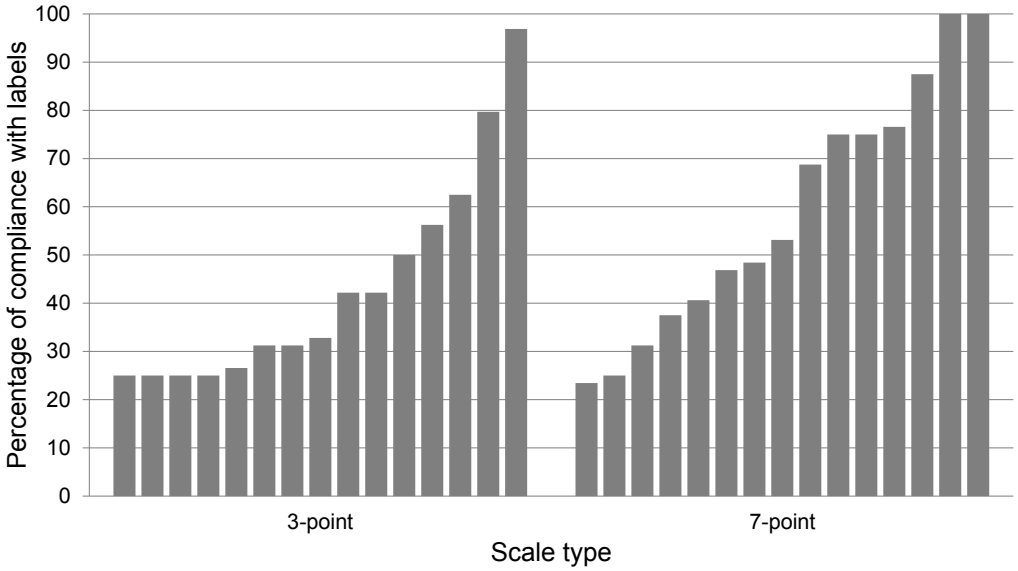


Figure 2.9: UHD: Percentage of compliance with labels.

the dark column on the left end of the figure, while the corresponding highest level can be found on the right. The height of a given column represents the number of times a specific rating option was used in the post-experiment questionnaire. As there were 30 test participants per experiment type and every test participant filled out the questionnaire, the 5 figures show the rating distribution for the 5 common questions.

Both experiments were assessed similarly regarding mental demand (see Figure 2.10). One could expect that labels mentally support subjective quality evaluation as their presence may guide the observer, yet it did not result in any significant difference. When labels were shown, the average rating was -0.33, with 11 negative and 19 positive scores, and without labels, the corresponding values were -0.16, 12 and 18. The distribution of scores was similar as well, with many test participants indicating either very low or slight mental demand.

According to the results, the experiments were less demanding physically than mentally (see Figure 2.11). The ratio of positive and negative values was roughly the opposite, with 10 positive and 20 negative scores for both experiments. The average values were -2.96 and -2.4, with and without labels, respectively. Approximately a third of the test participants expressed deficient physical demand, and cases of high demand were rarely registered.

The pace of the experiment was not deemed to be hurried or rushed. Again, the obtained results did not differ in a statistically significant extent (see Figure 2.12). With labels, the average was -3, with 10 positive and 20 negative scores. Without

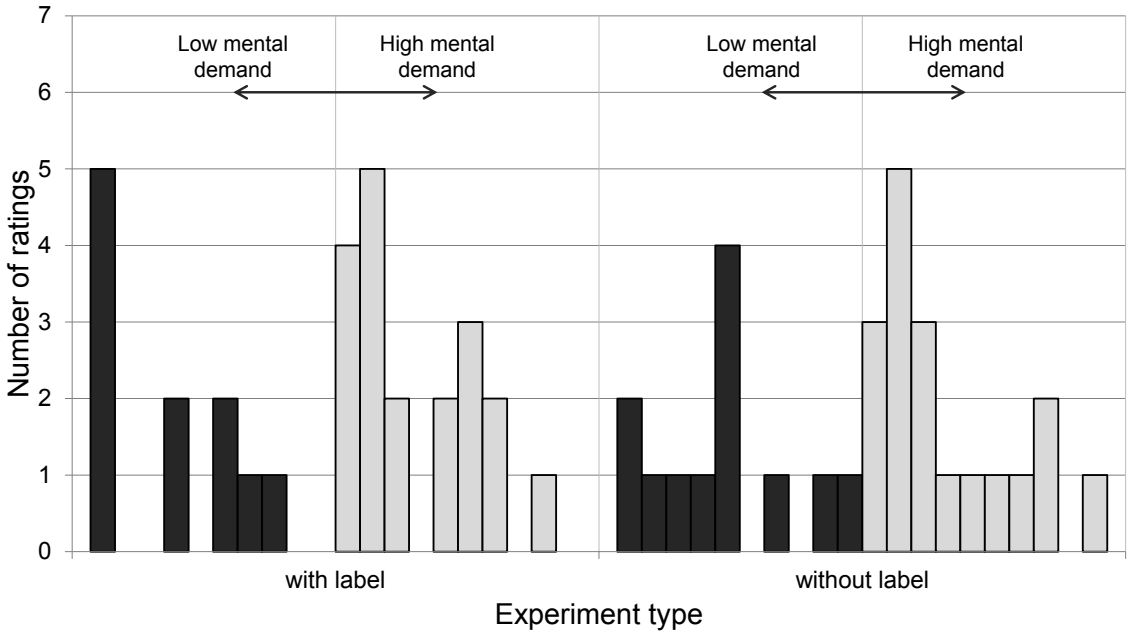


Figure 2.10: UHD: Results on mental demand.

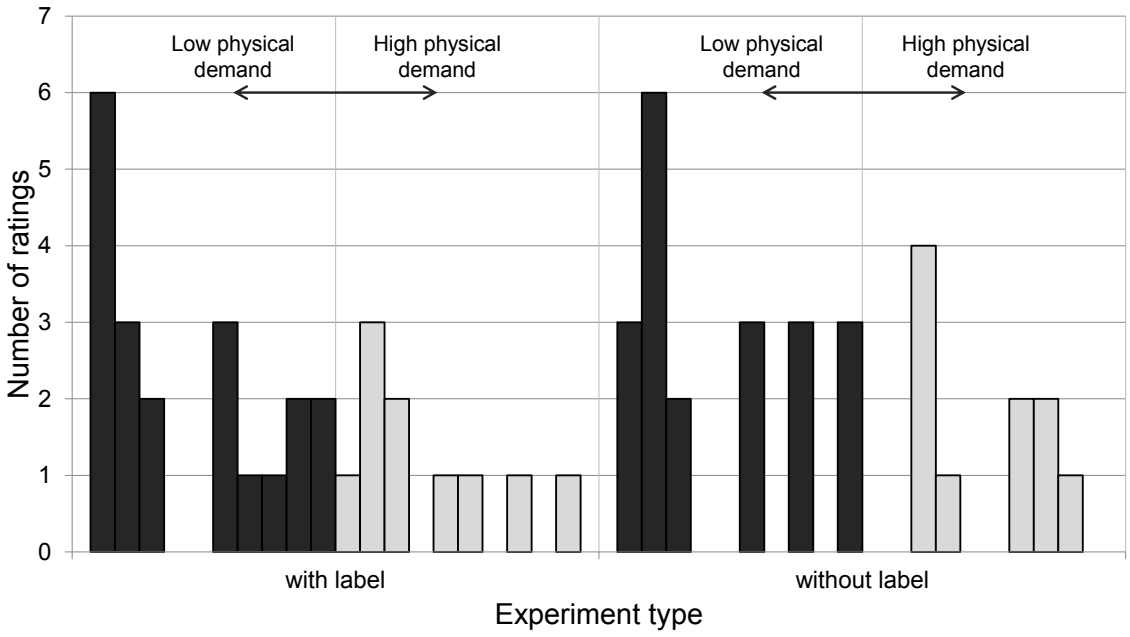


Figure 2.11: UHD: Results on physical demand.

labels, these were −2.83, 21, and 9. For the two experiments combined, there were only a total of 2 scores between 5 and 10 on the scale, indicating that only 2 out of 60 test participants considered the test structure (see Figure 2.2) to be too rushed.

Compared to the previous components of the questionnaire, the results on the self-

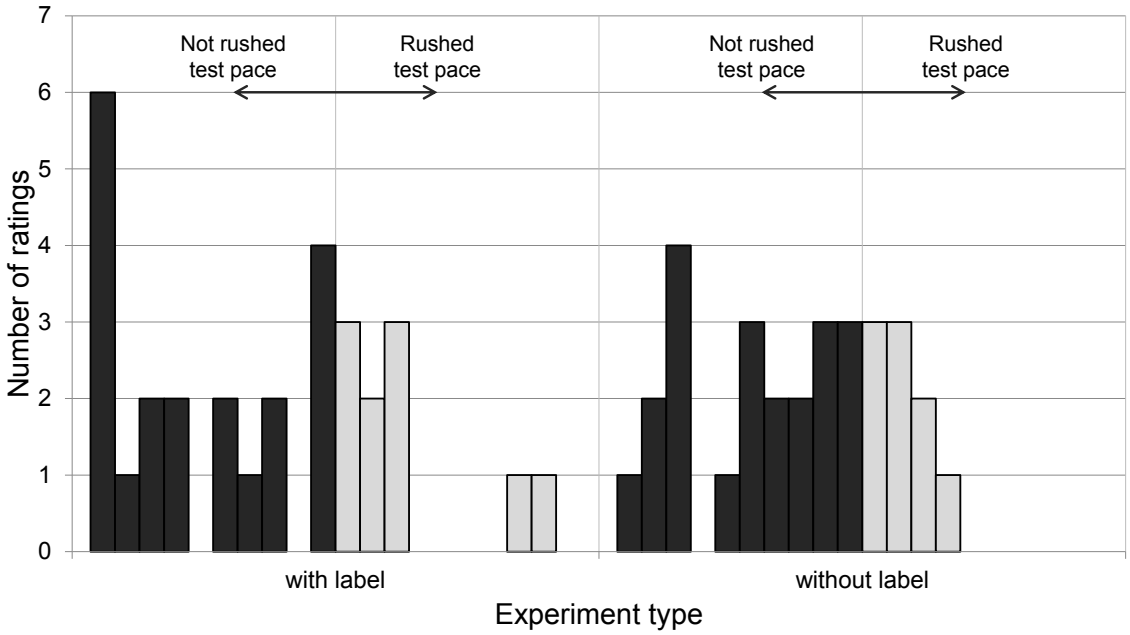


Figure 2.12: UHD: Results on test pace.

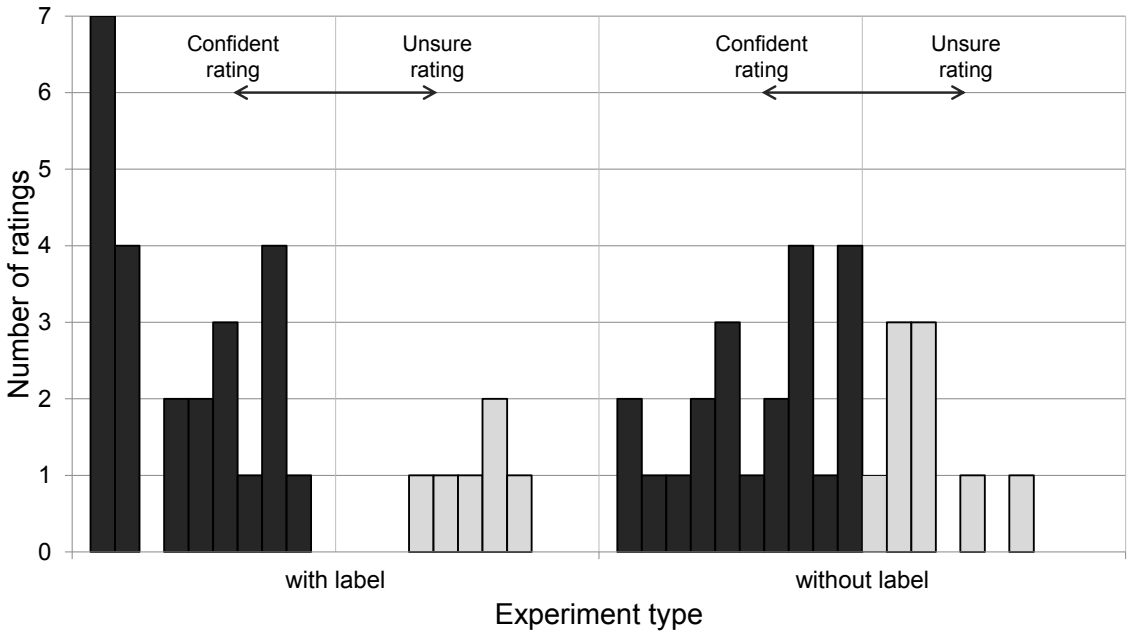


Figure 2.13: UHD: Results on task success.

assessment of scoring task success show much greater differences (see Figure 2.13). Yet the differences even for these ratings are not statistically significant, due to the high deviation of scores. The averages were -4.26 and -2.4, with 24 and 21 positive, and 6 and 9 negative ratings, for the experiments with and without labels, respectively.

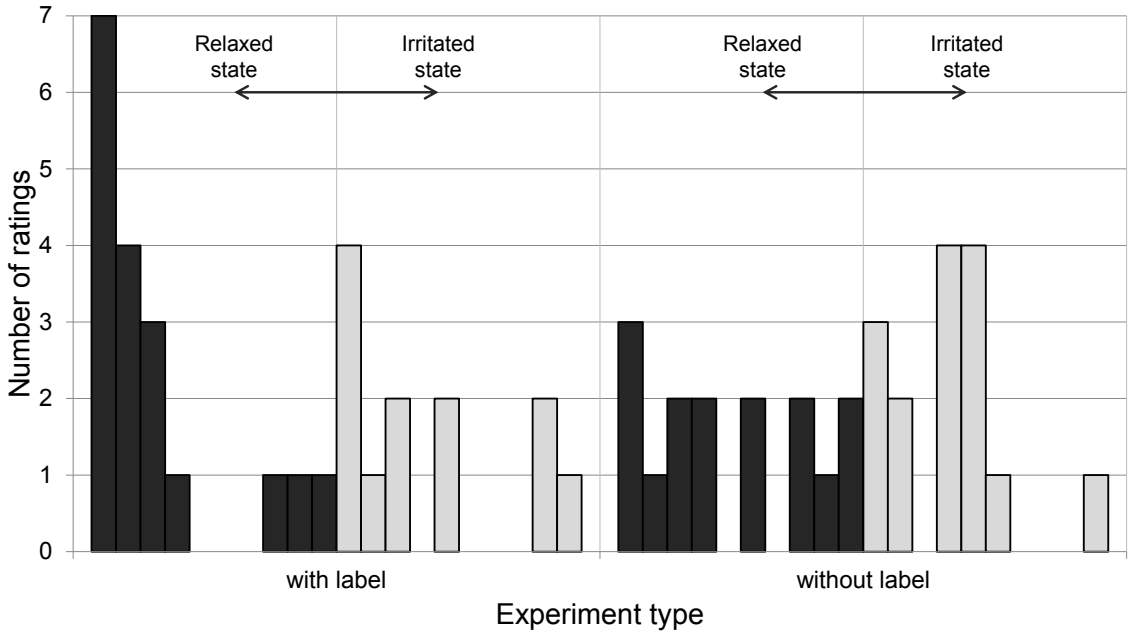


Figure 2.14: UHD: Results on irritation.

Such results can point out that the presence of labels can improve the overall rating confidence of the test participants. Therefore, ironically, it needs to be noted that distorted, biased ratings are submitted with more confidence.

The greatest scoring difference in the post-experiment questionnaire was achieved for the last common component, regarding the irritation of the test participant (see Figure 2.14). Although no statistically significant difference was found, the averages were -3.1 and -1, with 18 and 15 positive, and 12 and 15 negative scores, for the experiments with and without labels, respectively. With labels, 14 test participants gave a score of -8 or lower, indicating the lack of annoyance, while without labels, only 6 did. Apparently, the guidance provided by labels reduced the overall level of irritation during the experiment.

2.5.7.2 Preference Statement with Labels

When labels were presented, test participants were asked whether they found UHD/4K to be better than HD video or not. The results were balanced, as 11 test participants claimed 4K to be better, 10 stated the opposite, and 9 could not come to a conclusion based on the observed video stimuli and their labels.

2.5.7.3 Claimed Source of Perceived Difference

At the end of both post-experiment questionnaires, test participants were asked about the "source of difference" between the video stimuli in the pairs, with or without labels. The indicated reasons were diverse, but they could be categorized by a simple list of keywords (see Table 2.7). The first important category is "no difference", where the test participants stated that no visual difference could be perceived between the stimuli. These are also well-reflected in the per-subject rating analysis (see Figures 2.6a, 2.6b, 2.7a, and 2.7b). When labels were present, 7 out of 30 test participants could not explain what the difference was or whether there was a difference at all. This number was only 1 when no labels were provided. From these 7 test participants in the experiment with the labels, 4 of them could not explain what UHD was and 3 had never heard about UHD before; 3 had never seen UHD before and 4 did not know whether they had seen UHD before or not; 4 did not possess a UHD-capable device and 3 were unsure about the resolution of their equipment. Summa summarum, none of those who could not determine the source of the perceived visual difference knew what UHD was, had seen UHD videos prior to the experiment, and possessed a UHD-capable device — or at least was not aware of it.

The relevance of this information is that, generally, people refrain from providing an answer rather than providing a wrong, incorrect answer. These 7 test participants perceived differences between the video stimuli — as reflected by their scores — and were aware of the suggested reason via the labels, but they did not have prior experience with the given resolution, and thus, avoided answering this question of the post-experiment questionnaire in order to prevent a technically false statement. Although each and every test participant was precisely informed that data was handled confidentially and that no registered rating or answer could be linked to any individual, many of them still were afraid of being judged based on their lack of knowledge on the subject. As an illustrative example, one of these test participants particularly commented after the test,

> "I saw that there was some sort of a difference, but I just didn't want to write something stupid."

Again, when no labels were present, this applied to only one test participant. There were, however, 24 test participants who experienced differences and provided feedback on the matter. Generally, all keywords appeared more frequently in this experiment, compared to the test with labels. It is important to note that when no

Table 2.7: Keywords of the "source of difference" in the post-experiment questionnaire.

Keyword	**with labels**	**without labels**
No difference	6	5
No idea	7	1
General differences (small)	0	4
General differences (notable)	2	4
Visual details	7	9
Colors	3	8
Frame rate	5	6
Luminance	1	4
Content-related difference	1	1
Combinations of the above	1	7

labels guided the test participants, 7 out of the 24 identified multiple types of visual differences — e.g., frame rate and colors — while with labels, it only occurred once.

General differences varied a lot in ways of phrasing, formulating. Small differences indicate that the test participants managed to detect visual differences, but they were either difficult to perceive and/or did not have a significant impact on the experience. Notable differences include cases when test participants made a general remark about the visual quality, such as "4K is nicer" or stating that "everything" is different. Answers regarding visual details were the most relevant to the actual perceivable differences, yet in numerous cases, test participants noticed differences in colors, frame rate and luminance (brightness).

Frame rate is a particularly interesting aspect, as multiple test participants reported in the test with labels that HD had a better frame rate. For each and every stimulus, the frame rate was constant, unvarying, yet differences between the stimuli in the pairs were experienced, due to the concept of the trade-off between frame rate and other quality aspects. For example, one of the aforementioned participants stated the UHD had better visual details, but HD had a better frame rate. In the experiment without labels, changes in frame rate were indeed indicated as well, but not in such manner.

It must be clearly stated that the video stimuli did *not* differ in colors, frame rate and luminance at all; the stimuli only varied between the two resolutions, according to the investigated test conditions (see Table 2.1). Yet the test participants experienced differences in these aspects. This phenomenon can be explained through the process of cognitive dissonance reduction [57]. The test participants were presented 128 short videos (64 pairs), where the stimuli in the pairs did not differ significantly, if they even

Table 2.8: Significantly different options (o_1 and o_2) in the statistical analysis of the first question of the pre-experiment questionnaire. The experiment types (e), scale types (s) and conditions (c) are indicated, the p-values of Tukey HSD (T), Holm (H), and Bonferroni (B) multiple comparisons are given, along with the option achieving higher rates (O).

e	**s**	**c**	o_1	o_2	**T**	**H**	**B**	**O**
1	3	3	1	2	0.029	0.032	0.032	1
1	3	8	1	2	0.007	0.008	0.008	2
1	3	8	1	3	0.013	0.010	0.015	3
1	7	3	1	3	0.047	0.054	0.054	1
1	7	6	2	3	0.004	0.004	0.004	3
1	7	8	1	3	0.020	0.022	0.022	3
1	7	8	2	3	0.045	0.035	0.052	3
2	3	5	1	3	0.029	0.033	0.033	3
2	3	8	1	2	0.043	0.049	0.049	1

differed at all. In fact, 3 out of 4 pairs showed identical videos in both experiments. Among many other factors, the sole number of video stimuli can evoke a cognition that suggests that "there should be a difference". When labels where presented, the theoretical difference was indicated, but there was no information on how that would manifest in the perceived quality. Without labels, the only hint a participant could have extracted was from the pre-experiment questionnaire, asking three questions about UHD. In many of the cases, the cognition "there should be a difference" was matched with the perception "there is no difference", and the latter was overruled in order to eliminate this dissonant cognitive state.

2.5.7.4 Correlation between Ratings and Questionnaire Results

Beyond providing a comprehensive insight into the pool of test participants, the questionnaire results can also be used to enhance the understanding of the obtained quality ratings via correlation analysis. The first item of the pre-experiment questionnaire (see Table 2.2) had three possible options, aiming to distinguish the familiarity of test participants with UHD visualization. The quality ratings were clustered by these answers, and they were compared separately for experiment type, rating scale, and test condition, resulting in 96 statistical tests. The results are shown in Table 2.8. Only 9 out of 96 multiple comparisons indicated statistically significant differences. As an example, the first line of the results tells us that when labels were involved (as 1

Table 2.9: Significantly different options (o_1 and o_2) in the statistical analysis of the second question of the pre-experiment questionnaire. The experiment types (e), scale types (s) and conditions (c) are indicated, the p-values of Tukey HSD (T), Holm (H), and Bonferroni (B) multiple comparisons are given, along with the option achieving higher rates (O).

e	**s**	**c**	o_1	o_2	**T**	**H**	**B**	**O**
1	3	1	1	2	0.048	0.056	0.056	1
1	7	3	1	2	0.013	0.014	0.014	1
1	7	3	1	3	0.022	0.017	0.025	1
1	7	6	1	3	0.031	0.035	0.035	3
1	7	8	1	3	0.034	0.038	0.038	3
2	3	5	1	2	0.023	0.017	0.026	2
2	3	5	1	3	0.009	0.010	0.010	3
2	7	1	1	2	0.023	0.023	0.023	1
2	7	5	1	2	0.048	0.048	0.048	1
2	7	8	1	2	0.001	0.001	0.001	1

stands for the first experiments with labels, and 2 for the second experiment without labels) and the 3-point scale was used, and both video stimuli were UHD but the first one was labeled as HD, those who were the most familiar with the technical term of this visualization technology favored the second stimulus (with the UHD label) significantly more compared to those who only heard about it.

The results for the second and the third item of the pre-experiment questionnaire are shown in Table 2.9 and 2.10, respectively. In case of the third question, the fourth answer type — which is technically a lack of information — was not included in this analysis, as in practice, it may be any of the first three answers, and thus, its inclusion does not provide any meaningful conclusion.

For the first question, the majority of significant difference was found in the experiment with labels, for both scale types. The second question was similarly balanced for scale types, but also for experiment types. Significant differences in the third one were more for the experiment without labels, dominantly for the 7-point scale. One of the most notable phenomena in the analysis is that for the third question, the scores of those who confirmed owning a UHD-capable device in the experiment without labels, using the 7-point scale, were always significantly higher for the second stimulus in every test condition than the scores of those who did not have a such a display.

The results for the first question of the post-experiment questionnaire in the study

Table 2.10: Significantly different options (o_1 and o_2) in the statistical analysis of the third question of the pre-experiment questionnaire. The experiment types (e), scale types (s) and conditions (c) are indicated, the p-values of Tukey HSD (T), Holm (H), and Bonferroni (B) multiple comparisons are given, along with the option achieving higher rates (O).

e	**s**	**c**	o_1	o_2	**T**	**H**	**B**	**O**
1	7	1	2	3	0.017	0.017	0.017	3
1	7	2	2	3	0.032	0.032	0.032	2
1	7	4	2	3	0.000	0.000	0.000	3
1	7	5	2	3	0.048	0.048	0.048	2
1	7	7	2	3	0.028	0.028	0.028	3
2	3	5	1	2	0.028	0.021	0.031	2
2	3	5	1	3	0.005	0.005	0.005	3
2	3	7	1	2	0.031	0.035	0.035	2
2	7	1	1	2	0.006	0.006	0.006	1
2	7	2	1	2	0.003	0.004	0.004	1
2	7	3	1	2	0.021	0.023	0.023	1
2	7	4	1	2	0.003	0.003	0.003	1
2	7	5	1	2	0.001	0.001	0.001	1
2	7	6	1	2	0.010	0.011	0.011	1
2	7	7	1	2	0.014	0.015	0.015	1
2	7	8	1	2	0.000	0.000	0.000	1

with labels are shown in Table 2.11. With two types of scales, eight conditions and three answer types, there were a total of 48 statistical comparisons, from which 28 resulted in significant differences. The most important finding here is that 25 out of 32 comparisons involving the first answer — stating that UHD is better than HD — show statistically significant differences, commonly preferring the stimuli with the UHD label. Furthermore, 6 out these 7 comparisons where no significant difference was found either belonged to conditions 7 and 8, where the labels suggested no difference. Note that the ratings linked to the first answer differ from those associated not only with the second answer — stating that UHD is not better than HD — but with the third one — stating the lack of a confident answer — as well. Wherever there were significant differences between the second and the third answer or all three answers, the first one favored the stimulus with the UHD label the most, then the third one, and finally the second.

Table 2.11: Significantly different options (o_1 and o_2) in the statistical analysis of the first question of the post-experiment questionnaire in the study with labels. The scale types (s) and conditions (c) are indicated, the p-values of Tukey HSD (T), Holm (H), and Bonferroni (B) multiple comparisons are given, along with the option achieving higher rates (O).

s	**c**	o_1	o_2	**T**	**H**	**B**	**O**
3	1	1	2	0.000	0.000	0.000	1
3	1	1	3	0.000	0.000	0.000	1
3	2	1	2	0.000	0.000	0.000	2
3	2	1	3	0.000	0.000	0.000	3
3	3	1	2	0.000	0.000	0.000	1
3	3	1	3	0.000	0.000	0.000	1
3	4	1	2	0.000	0.000	0.000	1
3	4	1	3	0.000	0.000	0.000	1
3	5	1	2	0.000	0.000	0.000	2
3	5	1	3	0.000	0.000	0.000	3
3	6	1	2	0.000	0.000	0.000	2
3	6	1	3	0.000	0.000	0.000	3
3	8	1	3	0.036	0.041	0.041	3
7	1	1	2	0.000	0.000	0.000	1
7	1	2	3	0.035	0.026	0.039	3
7	2	1	2	0.000	0.000	0.000	2
7	2	1	3	0.000	0.000	0.000	3
7	3	1	2	0.002	0.002	0.002	1
7	3	1	3	0.031	0.023	0.034	1
7	4	1	2	0.000	0.000	0.000	1
7	4	1	3	0.001	0.001	0.002	1
7	4	2	3	0.003	0.001	0.003	3
7	5	1	2	0.000	0.000	0.000	2
7	5	1	3	0.001	0.001	0.002	3
7	5	2	3	0.039	0.015	0.044	2
7	6	1	2	0.000	0.000	0.000	2
7	6	1	3	0.002	0.001	0.002	3
7	7	1	2	0.008	0.009	0.009	1

2.6 Discussion

The results presented in the previous section highlighted the potential magnitude of the labeling effect in the context of perceived quality. As detailed in the beginning of the chapter, the labeling effect is unavoidable, simply inevitable in real-life use case scenarios. In fact, labels are actually desired by manufacturers and content providers,

as such information serves important commercial purposes, especially regarding user decisions. One could then think that QoE studies should integrate labels into their methodologies, in order to reduce the gap between the experience measured in the lab and what is actually experienced in real life. However, that would induce unnecessary cognitive bias for many types of subjective tests, where the inclusion of labels would do more "harm" to the collected results than the realism it could bring. Therefore, it is not recommended to provide further information to test participants that may bias their QoE ratings, particularly if the impact is not measurable.

2.6.1 Labels in QoE Studies

The topic of the labeling effect in QoE studies cannot simply be dealt with by saying "avoid such information in the experimental configuration, period". First of all, information that may bias the test participants should be categorized based on the type of information and how the test participants encounter it. Both can be explicit and implicit (or direct and indirect). While explicit information is straightforward and requires no additional effort to process it, implicit information needs to be derived from the experimental environment. Information can be directly, explicitly provided to the test participants, or it can be left for the test participants to discover themselves. The study introduced in this chapter involved explicit information, which was provided to the test participants in an obvious manner. Labels are always explicit information, but they can be supplied implicitly. For example, if the label is the brand of the display, but the test calls no direct attention to it, then it depends on the test participants whether they notice it or not. Let us now imagine a QoE study that uses a laptop as the apparatus of the test participants, and the test involves a service or an application that requires Internet access (e.g., streaming, browsing, online gaming, etc.). Having no Ethernet cable attached to the laptop is implicit information, since test participants may derive the fact that the laptop must be using some sort of wireless access. Both explicit and implicit information may be used in QoE studies, but their handling should commonly be explicit if it is assumed that there is an impact of that test factor on the measured variables. In general, implicit handling should be avoided — unless the research question particularly demands it (e.g., if the study aims to find out whether the test participants notice certain information or not), or the goal of the test is to deceive participants and steer their attention to other seemingly more important aspects of the study.

Regarding the explicit handling of information, the frequency of information provisioning is also worth mentioning. The experiment at hand used a high-frequency

notification technique, i.e., the test participants were notified of video resolution right before every single stimulus. This study could be repeated with a single notification in the beginning — e.g., analogous to conditions 1, 3, and 4 (see Table 2.1) — which applies to all stimulus pairs, but the test is conducted without the repetition of this information. In this case, the influence of the labeling effect could diminish over time.

What experimental methodology, which combination of the aforementioned techniques and parameters is the closest to certain real-life scenarios? Let us examine two cases. The first situation is a person buying a UHD-capable display in a shop. In a QoE study, this could correspond to an experiment on the WTP, as the financial decision at hand fundamentally depends on the quality. The labels would be explicit and emphasized, covering technical capabilities and price. The labels, especially the price, would either be frequently (explicitly) or constantly (explicitly or implicitly) shown to the test participants. The test could be performed with either a single apparatus or multiple; although having multiple displays is more realistic, a single one is sufficient as well, if the different stimuli (with different quality parameters) represent different displays, and the one used in the experiment has the sufficiently high capabilities to accommodate the stimuli properly.

The second situation is a person watching a UHD-capable display in a home scenario. When modeling such a situation in a QoE study, it is actually important to distinguish whether the display has been recently purchased or it has been used for a while. The reason of its relevance is post-decision dissonance [82,83]. It is a process of decision justification (e.g., "buying this TV was a good choice"), in which — similarly to other forms of cognitive bias — the perception of quality may be affected. If this effect is not excluded from the experimental configuration, then the subjective test needs to include a user decision prior to quality assessment, and the labels play a more significant role as well. If a study does not take the post-decision dissonance into consideration, it is sufficient to explicitly present the labels at the beginning of the test. The relevance of labels in the modeling of such scenario is that a person is usually aware of fundamental information regarding his or her TV. While some people may know their displays better than others, and such knowledge may easily fade, one tends to remember at least the property that convinced him or her towards the decision of purchase. Furthermore, in many parts of the world, people change their displays more frequently (handhelds and larger screens alike), and therefore, more decisions are made, and knowledge on display properties has less time to fade.

2.6.2 Experimental Design and Source Contents

One could argue that the experiments presented in this chapter did not utilize the full potential of UHD visualization, as the source content used in the subjective tests were not as visually appealing as the typical UHD demo videos. It is indeed true that the test participants were not shown slow-motion close-up macro shots of the human eye or the wings of an exotic butterfly. However, as it has already been discussed earlier in the chapter, it is not a typical or even realistic user behavior to use a UHD display solely for such demo materials in a home scenario of multimedia consumption.

The source videos of this experiment were diverse in the sense that the contents included CGI animation, live-action clips and CGI-enhanced live-action scenes as well. When the experimental configuration was being put together, there was a reasonable thought that the choice of the source videos would affect the results, i.e., that there would be a statistically significant impact of the content, or that the content could lead to no visual differences being apparent. As it has been presented in the analysis, when grouped by the test conditions (and therefore, any other factor was ruled out), only 0.11% of all the comparisons differed significantly.

Regarding CGI videos, in general, as they are commonly rendered in the target resolution, they normally provide "crisp", detailed visuals (unless the artistic intention is to do the opposite). The only limitation here is the resolution of the 2D textures used on the video, but other than that, content can be rendered at any resolution in a straightforward (perhaps time-consuming) manner.

Professional contents usually intend to artistically exploit the technology they use. For example, if a movie is shot in stereoscopic 3D, it is expected to have at least one scene where the added value of 3D is justified through the visuals. The same idea can be applied to UHD contents as well. However, it needs to be noted that not all professional content shot in UHD considers the resolution during production, as the term "professional content" is not limited to high-budget feature films, and in fact, it is not even limited to movies. Furthermore, as a quality feature integrates into the use case scenarios and becomes the de facto standard, the will of the content creator(s) to emphasize it diminishes.

Unlike the professional content of the movie industry, where the cameras are handled by individuals with the necessary expertise and the scenes are adjusted in a way to provide the desired visual quality, user contents have no such criteria. The handling of the camera is highly emphasized here, as the captured noise and blur can easily degrade the added value of a higher resolution, as other studies have found.

2.6.3 Test Subject Behavior

In the analysis of the results, test subject behavior was addressed in the form of correctness and compliance with labels. Again, the metric of rating correctness describes how much subjective scores are in alignment with the objective quality of the stimuli. It applies to paired comparison tests if and only if the stimuli differ in a single parameter, which provides a clear differentiation in quality; in case of multiple variables, objective comparison becomes ambiguous. It is possible to utilize objective metrics and then measure how much the ratings agree with their results. The idea of rating correctness can be applied to ACR scales as well, which means that the results of every test condition must be paired and then compared. Although it is uncommon to use this approach in QoE studies, it has the potential of providing valuable insights for experiments that involve test stimuli with just noticeable differences (JNDs) or even less.

For certain studies, correctness could also be used as a weighting factor for subjective results. Let us assume that an experiment uses three variables, with all the different combinations of their selected values. In this case, the study could involve several test condition pairs, in which only one variable changes. Then the correctness of these results could be assessed, and based on their values, weights could be assigned to test participants. Therefore, a test participant with a higher correctness rate would have a more significant impact on the overall evaluation of the multivariable test conditions. This can be particularly useful when visual accuracy is more important in the study than personal preference.

The compliance with labels is a metric that shows how much the subjective scores are in accordance with the labels. If a QoE study aims to include labels in its experimental configuration, then the rate of compliance may prove to be useful in the task of understanding test subject behavior. This measure is applicable to both explicit and implicit labels; however, in case of implicit labels, it needs to be recorded in a post-experiment questionnaire whether the test participant noticed and considered the label or not.

2.7 Chapter Summary

The chapter has presented a series of experiments addressing the influence of the labeling effect on the perception of HD and UHD video. The obtained results indicate that the labeling effect had a significant impact on the subjective scores, regardless of test condition and source content. The corresponding study without labels concludes

the lack of statistical difference between the two video resolutions for either rating scale. However, the choice of rating scale greatly affected the test with the labels, as the more fine-grained 7-point comparison scale enabled the expression of the slighter perceived — and/or cognitively induced — differences, in contrast to the 3-point scale.

In the experiments of this chapter, the impact of the labeling effect was straightforward in the sense that the positive label was aimed at the overall visual experience. In the next chapter, a similarly positive label is used to identify the test stimuli, however, user experience is divided into different quality aspects, and they are addressed separately, some of which are not related to the visual appearance of the content, e.g., frame rate and stalling event duration. Furthermore, while the video sequences were genuinely different in certain stimulus pairs of the tests presented in this chapter, all stimuli in the paired comparisons of the next chapter are identical (they are exactly the same in every single aspect), in order to have a clear focus on the labeling effect.

Chapter 3

High Dynamic Range Visualization

3.1 Introduction

The Human Visual System (HVS) is the axis around which modern visualization technologies revolve. Whether we talk about resolution, frame rate or any parameter of a display or a visual content, the primary goal in research and development is to reduce the gap between digital visualization and the capabilities of the HVS. A common scenario is that the human observer can perceive more than what can be digitally provided, and the long-term scientific objective is to match the limits of such a biological sensory system. However, the opposite case is frequent as well, when certain limitations of the HVS can be exploited, i.e., for data compression, thus, making perceptual coding possible.

The dynamic range of the HVS is one aspect where conventional capture and visualization technologies — referred to as Low Dynamic Range (LDR) — are heavily outranked. It is the typical case of an evident, undeniable difference between what can be captured and displayed, and what can be perceived by a person in real life.

The recent advances of High Dynamic Range (HDR) motion imaging enable a contrast range in digital visualization that is close to the capabilities of the HVS, and may even exceed them. The high difference between the peak brightness and the black level, combined with the richer range of colors, enables HDR to deliver a much more life-like visual appearance, compared to the technological predecessors.

At the time of this thesis, HDR visualization already appears in multiple forms, and related standards emerge continuously. In the commercial sector, the standards of HDR10, its update HDR10 Plus and Dolby Vision are competing, but the ETSI SL-HDR1 standard (originating from Technicolor, STMicroelectronics, and Philips)

should be mentioned as well. The related ITU-R standard is Rec. BT.2100[1], which uses wide color gamut according to the Rec. BT.2020 color space, supports 10-bit and 12-bit colors, covers both HD and UHD spatial resolution, and also both Perceptual Quantizer (PQ) and Hybrid Log-Gamma (HLG) transfer function.

Although HDR visualization may greatly enhance the Quality of Experience (QoE), manufacturers boldly labeling their displays as "HDR" could severely damage the perception of the added value, and thus, result in disappointed users. In order to prevent this from happening, the Ultra HD Alliance defined strict requirements for what can be certified as HDR with the Ultra HD Premium logo. The given display needs to have a peak brightness of at least 1,000 nits, and black level must be lower than 0.05 nits, where the non-SI unit nit is defined as the luminous intensity per unit area of light traveling in a given direction, and 1 nit can be approximated as the light emitted from a single candle. As OLED televisions cannot reach a peak brightness of 1,000 nits, their requirements were specially reduced to 540 nits for peak brightness, and 0.0005 nits for black level, in order to compensate. In comparison, the peak brightness of LDR LED televisions are usually in an interval between 300 and 500 nits.

Needless to say, the difference in visual experience between watching a regular LED TV and an HDR10-certified TV displaying HDR demo (eye candy) content can be staggering. In the upcoming years, numerous further developments in manufacturing and standardization are expected in the field of HDR visualization. It is sufficient to consider the fact that on the level of specifications, Dolby Vision is superior to the current HDR10 and HDR10 Plus standards. As an example, while HDR10 only supports peak brightness up to 4,000 nits, this value is 10,000 in case of Dolby Vision. Also, not every manufacturer is aiming to achieve the HDR certification of Ultra HD Premium for their displays — as it is not mandatory — and some use their own terminology, e.g., "HDR Pro" in case of LG.

The potential future diversity in HDR visualization technology and its corresponding labeling shall provide a wide variety in both visual quality and device information. Let us consider the classic scenario where the end-user goes to a shop, observes the displays and their lists of parameters (including their prices), and selects one for purchase. There is an inevitable 2-way effect between the perceived visual quality and the listed parameters: (a) the list of information may affect the way the user perceives the quality, e.g., being aware of a higher bit depth may enhance the experienced difference regarding the actual difference in colors, and (b) the perceived

[1]Rec. BT.2100: Image parameter values for high dynamic range television for use in production and international programme exchange

quality may affect the interpretation of the information, e.g., seeing something that is very appealing visually may justify a higher cost or increase the subjective relevance of a technical parameter.

In the research present in this thesis, the first case is investigated, where the related information affects the perceived quality. More precisely, experiments were carried out where the research focus was on the effects of a single, non-technical word on the selected QoE aspects of HDR video quality. Test participants were shown identical video stimulus pairs of different source contents in a paired comparison, but one of the stimuli was labeled as a simple "HDR" video, and the other one as a so-called "Premium HDR"; the latter suggested the provision of superior visual quality in comparison to the other. The research question addressed in this thesis targeted the cognitive bias evoked by this specific label.

As stalling duration is one of the most important quality indicators of on-demand real-time video streaming, the effect of the chosen label on the perceived stalling duration of HDR videos is investigated as well. Furthermore, the research is extended by two works on subjective stalling detection: one that measures the perceptual thresholds of the stalling events used in the research on stalling duration, and one that analyzes the same in the context of conventional LDR visualization, in order address the topic of cognitive load and visual attention.

3.2 Related Research on HDR QoE

As HDR visualization is considered to be the next big step in consumer-grade home TV entertainment, its QoE aspects have been and currently are still being extensively investigated. The works of Narwaria *et al.* [22, 84, 85] address HDR QoE, taking into consideration immersion, the natural feeling of the visualized content, visual attention and many more aspects, while also discussing subjective measurement methodologies. The authors particularly investigated tone mapping operators (TMOs) and how they affect the perception of HDR content, and also proposed a novel objective video quality metric for HDR [86].

Trivially, the major added value of HDR visualization from a QoE perspective originates from the *high* dynamic range itself. However, measuring the dynamic range perceived by test participants is quite far from being a trivial task. The work of Hulusic *et al.* [32] introduces a subjective measurement methodology for the perceived dynamic range. The authors carried out a series of subjective tests with 20 test participants, in which HDR images (photographs and video frames) from various

sources (e.g., Fairchild's HDR Photographic Survey [87], the Stuttgart HDR Video Database [88], etc.) were assessed on a Full HD (1920×1080) SIM2 HDR display, namely the HDR47ES4MB. All still image stimuli were converted to grayscale, as the research solely focused on the perceived dynamic range. The test participants had to evaluate "the overall impression of the difference between the brightest and the darkest part(s) in the image" using a variation of the Subjective Assessment Methodology for Video Quality (SAMVIQ) [89]. The ratings were collected on a continuous scale (from 0 to 100), which was divided into 5 labeled, uniform intervals ("*Very low*", "*Low*", "*Medium*", "*High*" and "*Very high*"). The findings highlight the importance of content characteristics, such as the relative surface of bright areas and the distance, the separation between dark and bright areas.

Although one of the key features of HDR visualization is the higher level of brightness, having a screen that is too bright might not be preferable by the end-user. The work of Bist *et al.* [25] proposes a content-based method for brightness control, based on subjective studies of brightness preference. The algorithm operates on a pixel-level; the "bright" pixels of the visualized content are taken into consideration during brightness adjustment, which means that the larger the portion of bright areas on the screen is, the lower the level of brightness that shall be set. In their experiment, 16 test participants viewed static images on a SIM2 HDR47ES4MB HDR display, the brightness of which they had to re-adjust in case they found the images too bright.

Korshunov *et al.* [17] published a database of HDR images, compressed at different bitrates, and with different compression profiles. The resulting stimuli were evaluated using the same SIM2 HDR display as the previous experiments. During the subjective tests, multiple test participants (three in each session) were simultaneously seated in an arc configuration at 3.2 H distance. A DSIS method was chosen with simultaneous side-by-side stimuli (left and right side of the screen), and the difference between the reference and the compressed image was registered via a 5-point Degradation Category Rating (DCR) scale ("*Imperceptible*", "*Perceptible but not annoying*", "*Slightly annoying*", "*Annoying*" and "*Very annoying*"). The obtained subjective ratings validated the database through an even distribution of mean scores.

Using physiology in QoE studies is a very well-known approach within the scientific community [90]. Depending on the methodology, subjective tests may provide an immense amount of useful information regarding the personal quality preferences and the specific perceptual thresholds of the test participant; however, opinion scores do not report anything about the internal physiological levels of the individual. The

work of Al-juboori *et al.* [21] used electroencephalogram (EEG) to analyze the correlation between the perceived quality of HDR images and the different bands of brain activity. Four tone mapping algorithms were applied to 5 source HDR images, and the 20 stimuli were shown to the 28 test participants on an iPhone 6. The results highlight the emotions that were induced by the visualized content, as they correlate with the acquired EEG signals. EEG and peripheral physiological signals were also used by Moon *et al.* [23, 91], who found statistically significant differences in physiological signals between test scenarios of LDR and HDR visualization. EEG was also used by Darcy *et al.* [92], and the experiment of Daly *et al.* [93] studied pupil behavior during HDR video.

Similarly to Chapter 2, the research on the labeling effect is very relevant to the work presented in this chapter. The related experiments and contributions in the scientific literature is given in section 2.3, in the previous chapter of this thesis.

3.3 Displays and Research Environments

3.3.1 HDR Research

The subjective tests were performed in an isolated, controlled laboratory environment, with dimmed lighting conditions. The ambient luminance was nearly 10 lx and not lower in order to avoid visual discomfort [25]. The test participants viewed the HDR videos on a SIM2 HDR47ES6MB HDR display[2], with peak brightness over 6000 nits.

The viewing angle was zero degrees (center view) during the entire test, and the viewing distance was a fixed 3 H (1.75 meters) according to the recommendation[3], as Full HD (1920×1080) content was displayed on the full screen of the 47-inch Full HD display.

3.3.2 LDR Research

The LDR research was carried out under similar environmental conditions and experimental methodology. The only notable difference was the display itself. For these subjective tests, a Panasonic TX-P42S10E was used, which is a 42-inch Full HD plasma television. Similarly, the Full HD content was displayed on the entire screen, but as the display had a smaller screen compared to the one used in the HDR tests, the 3 H distance in meters (1.57 meters) was adjusted accordingly.

[2]SIM2 HDR47ES6MB display:
http://hdr.sim2.it/hdrproducts/hdr47es6mb

[3]Rec. BT.710: Subjective assessment methods for image quality in high-definition television

Table 3.1: Source video contents used in the HDR experiments.

ID	Video content name	Starting frame
1	Beerfest Lightshow	102351
2	Bistro	091397
3	Carousel Fireworks	097209
4	Cars Longshot	092355
5	Fireplace	092341
6	Fishing Longshot	060033
7	Poker Fullshot	045787
8	Poker Travelling Slowmotion	033800
9	Showgirl 1	235636
10	Smith Welding	248520

3.4 Source Videos

The contents were selected from the Stuttgart HDR Video Database [88], which is free for academic and educational use[4]. Table 3.1 shows the list of the 10 chosen contents (see also Figure 3.1), their associated IDs and the starting frames, from which the subsequent 500 frames were cut into 10-bit videos with 24 fps. Source video 2 ("*Bistro*") contains one cut and 5 ("*Fireplace*") fades from one camera image into another, while the other videos are continuous shots, either with a fixed-position or a panning camera.

The stalling events of the experiments on stalling detection and duration were implemented as frame freezing without visual indicators (e.g., rotating rebuffering icon); the selected frame was shown multiple times (12 times for 500 ms and 24 times for 1000 ms of stalling) before continuing with the next frame. For each content, 3 stalling event positions were selected, based on their TI values, which is a good estimation of the changes between frames. Figures 3.2 and 3.3 depict the TI charts of the 10 contents defined in Table 3.1, as well as the positions where frame freezing starts. The first stalling event in every video is denoted as A, the second as B and the third one as C. Note that, in all three experiments containing impaired videos, a given stimulus always contained exactly one stalling event. One stimulus is identified by the naming convention of either $\{Source\_ID\} + \{Stalling\_event\}$ or $\{Source\_ID\} + \{Stalling\_event\} + \{Stalling\_duration\}$; e.g., in the research on stalling detection, where only one given stalling duration was used, $5C$ denotes the third stalling event

[4]https://www.hdm-stuttgart.de/vmlab/hdm-hdr-2014/

Figure 3.1: HDR: Source videos used in the subjective tests.

in content 5, and in the research on stalling duration, this is extended with either an S (short duration) or an L (long duration) character, thus, the identifiers *5CS* and *5CL* were used.

The stalling events were particularly positioned on local and global minima and maxima in the TI chart, but also addressed near-identical TI values (even within a content, e.g., *6B* and *6C*). Some of these events were extreme cases, such as *3C*,

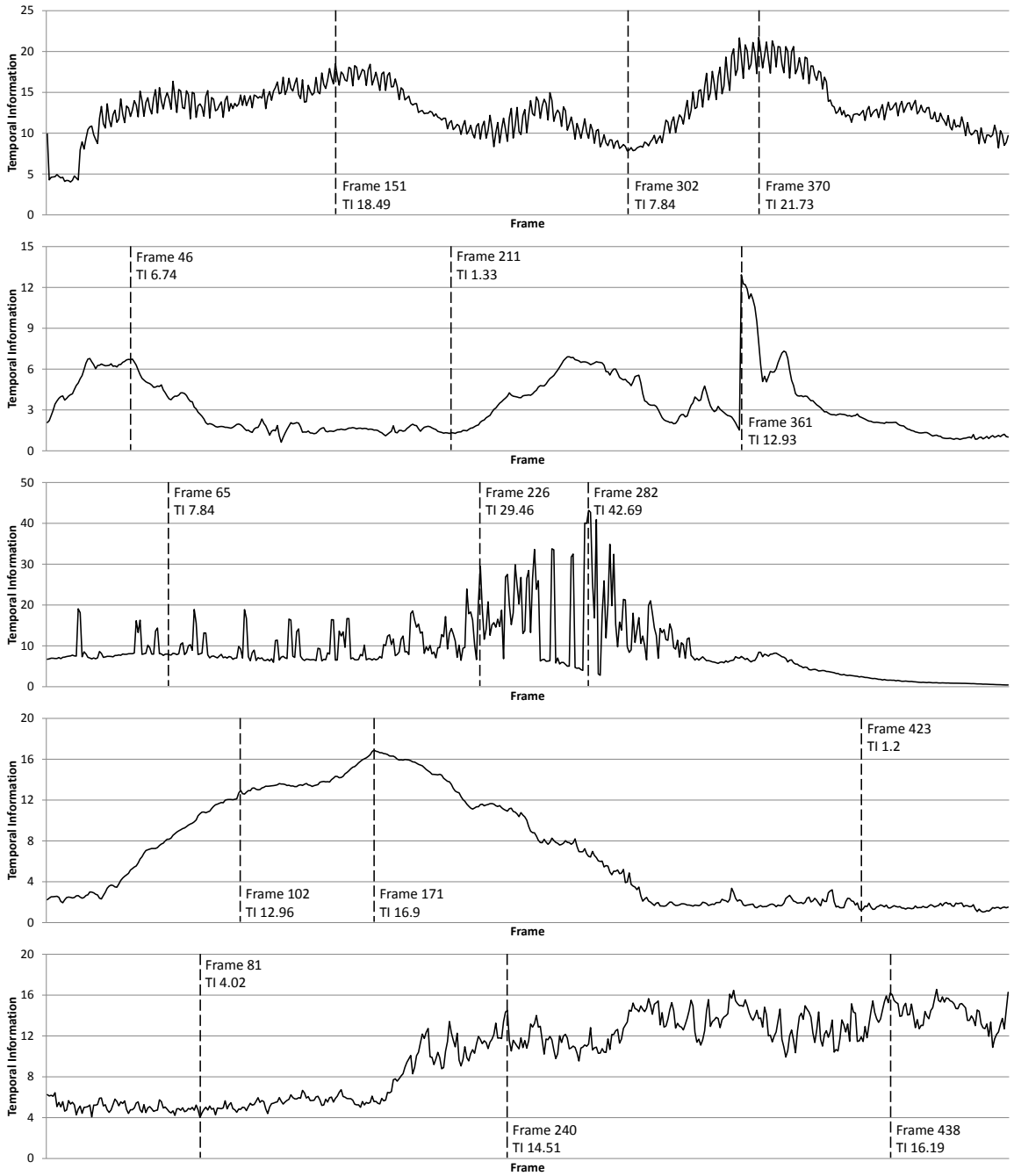


Figure 3.2: HDR: Temporal Information of contents 1 to 5, presented in a top-down order. The stalling events are denoted with dashed lines.

shown on Figure 3.4; the frozen frame (282) was a 1-frame flash of light. The first and last nearly 2 seconds of the content were kept clear of stalling.

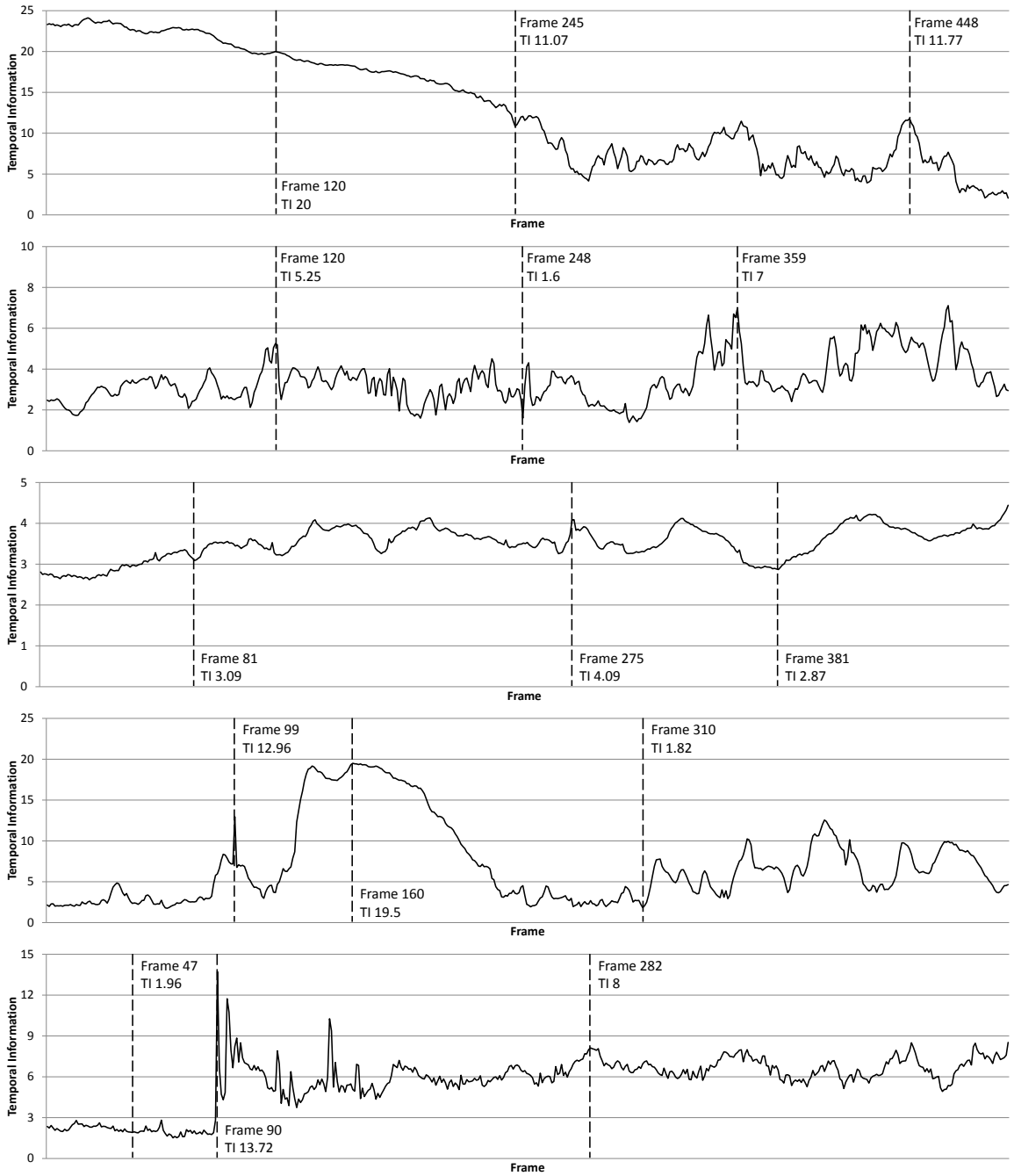


Figure 3.3: HDR: Temporal Information of contents 6 to 10, presented in a top-down order. The stalling events are denoted with dashed lines.

3.5 Research on HDR Quality Aspects

3.5.1 Research Aim

The aim of the research was to assess the impact of the labeling effect on the selected visual quality aspects [94].

3.5.2 Test Conditions

The test itself was a paired comparison, which compared video stimuli on a 7-point comparison scale ("*Much worse*", "*Worse*", "*Slightly worse*", "*Same*", "*Slightly better*", "*Better*", "*Much better*"). In order to gain a more detailed insight into the cognitive bias created by the labeling effect, instead of comparing the overall QoE, the test participants had to assess four aspects of HDR video quality: *luminance*, *frame rate*, *color* and *image quality*.

Before the subjective test, the test participants received training, during which the four aforementioned aspects were interpreted and demonstrated. *Luminance* was described as the perceived difference between the brightest and the darkest portions of the screen; greater differences were to be evaluated better. Although *frame rate* was considerably self-explanatory, it was still explained to every participant in order to avoid confusion and possible misunderstandings. *Color* was interpreted as the richness, the depth of the colors on the screen. Lastly, *image quality* was approached from the angle of spatial resolution and classic coding artifacts, independently from the other three aspects.

During the training phase, test participants were informed about the two labels. However, unlike the case of the quality aspect, they did not receive any specific interpretation of the labels. Doing so could have compromised the fundamental goals of the research, as test participants were to build their very own preconceptions regarding the label "Premium" in the context of HDR visualization. This consideration, of course, applied to the research on HDR stalling duration as well due to the same method of labeling.

The double stimulus method was used, with the stimuli in a pair shown after each other. They were separated by a 5-second blank screen, and comparison was performed directly after each pair, in a time window of 10 seconds. The stimulus pairs were also separated by a 5-second blank screen.

As detailed on Figure 3.5, for a given content i — where i is a content identifier between 1 and 10, corresponding to the source order randomized for each participant — the first instance of the content (VA_i) is played, followed by the stimulus separation (S_i), and then the identical second instance (VB_i) is shown. After this, VB_i is compared to VA_i in the comparison period (C_i), and finally the separation screen between the pairs is displayed (P_i). As this given structure is repeated over the duration of the subjective test, if i is at least 2 but at most 9 (i.e., neither the first not the last pair), then VA_i occurs directly after the comparison period and the

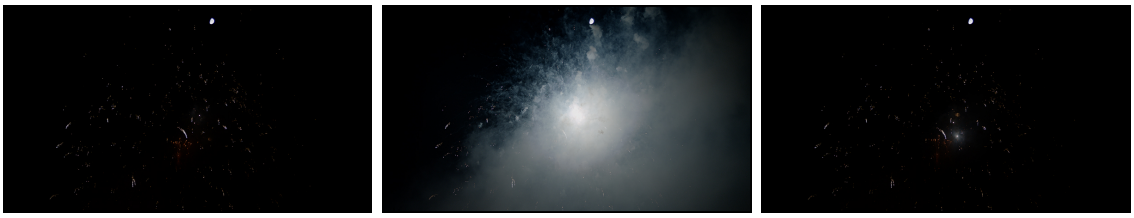


Figure 3.4: HDR: Frame 281, 282 and 283 of content 3.

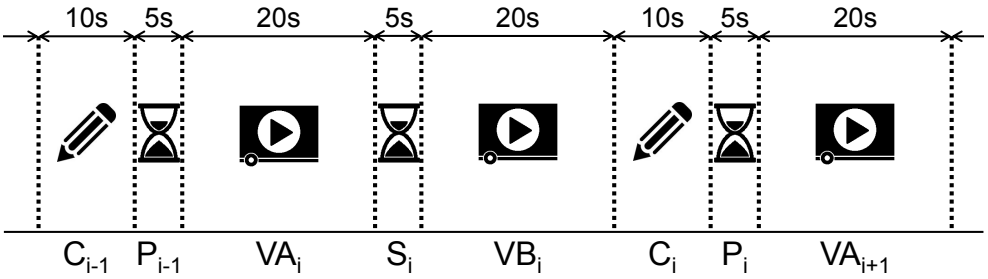


Figure 3.5: HDR: Temporal structure of the subjective test on quality aspects.

separation screen of the prior content $i-1$ (C_{i-1} and P_{i-1}, respectively), and P_i is followed by the first instance of the subsequent content $i+1$ (VA_{i+1}).

The order of stimuli in the subjective quality assessment varied among test participants. For half of the participants, the "Premium" video was always the first one in the pair (VA), and for the other half, it was the second one (VB). Again, this means that for each and every test participant, the assignment of the label was consistent and did not change during the test. As the labeling effect can influence both perception and the memory of perception, this given division between the test participants was included in order to investigate the role of label order.

3.5.3 Results

A total of 40 individuals participated in the tests (30 males and 10 females). The age range was from 20 to 56, and the average age was 30. 10 participants had prior HDR video experience, and the rest had never seen any HDR video before the experiment.

The obtained subjective scores are represented by their numerical counterparts, ranging from -3 to $+3$. During the subjective tests, the test participants were presented a combination of the available qualitative tags for stimulus comparison — defined in the previous section — and these values, emphasizing a uniform distance between the values of the scale. In this analysis, positive values favor the "Premium HDR" stimulus, while negative values indicate that it was deemed to be worse in the given aspects.

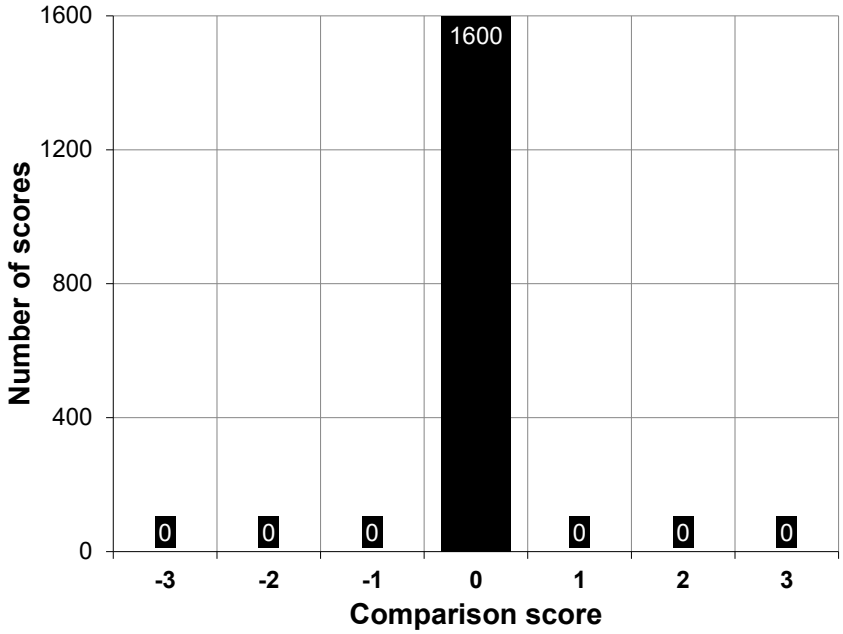


Figure 3.6: HDR: Ideal distribution of scores of the subjective test on quality aspects.

Each of the 40 test participants compared 4 quality aspects of 10 stimulus pairs, thus, 1600 subjective scores were collected in the experiment. In an ideal scenario without the presence of cognitive bias through the labeling effect, all these 1600 scores would have reported the given aspects to be the "*Same*" (see Figure 3.6). However, according to the scoring distribution, only 356 (22.25%) of them were zero, and 1244 (77.75%) assessed a certain level of either positive (1089 scores) or negative (155 scores) difference (see Figure 3.7).

The most frequent quality comparison score was "*Slightly better*", followed by "*Better*", "*Same*", "*Much better*", "*Slightly worse*", "*Worse*" and "*Much worse*". This order took all of the investigated aspects into consideration. If we separate them, we can observe rather similar mean values for *luminance*, *color*, and *image quality* (see Figure 3.8). In fact, the aforementioned order in score frequency applied to all three of them (see Figure 3.9), and there was no statistically significant difference between them.

However, *frame rate* was assessed differently. The mean score was significantly lower compared to the other aspects, as the number of positive scores was the lowest, while it received the most zero and negative scores. Moreover, the number of negative scores *frame rate* received was near to the number of negative scores received by the other three aspects together.

It needs to be noted that more than half (201 out of 400) of the scores for *frame*

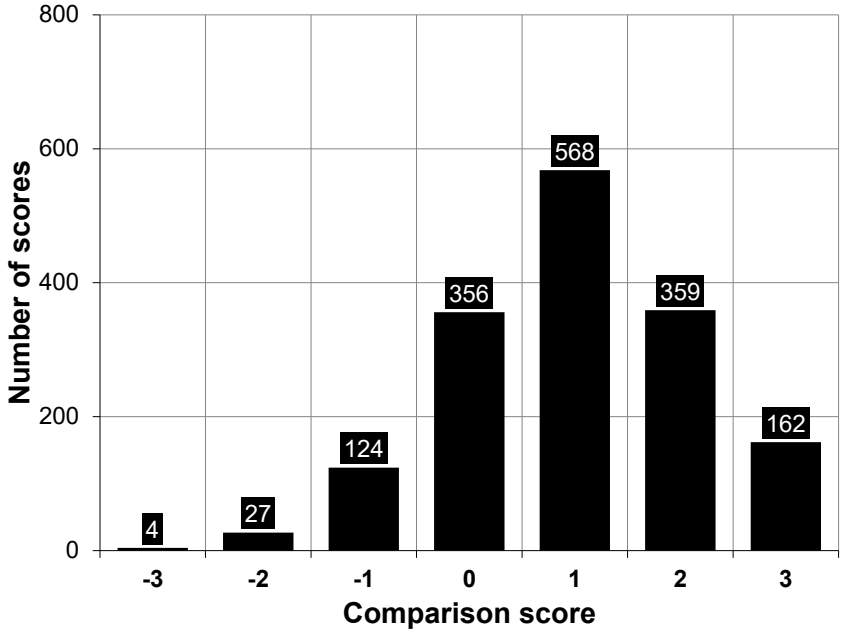


Figure 3.7: HDR: Scoring distribution of the subjective test on quality aspects.

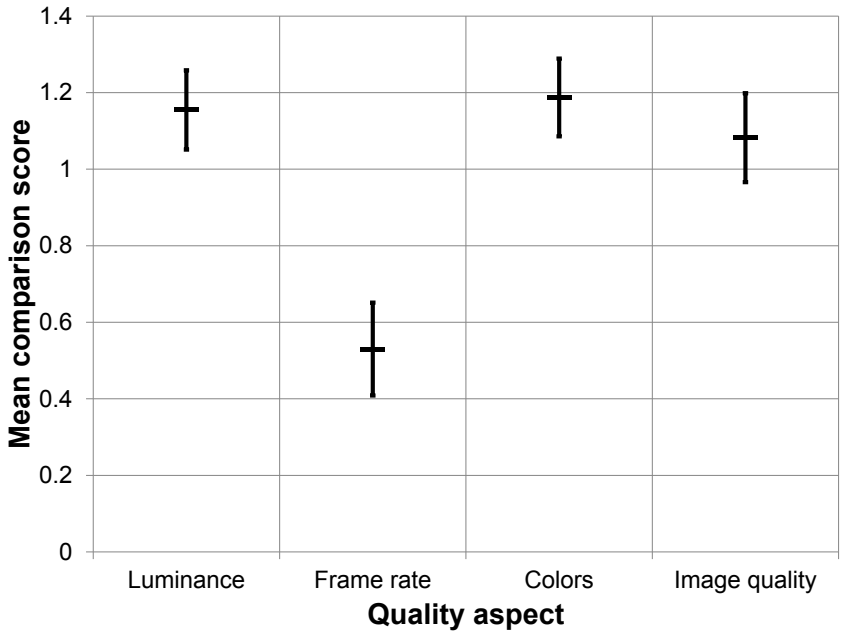


Figure 3.8: HDR: Mean comparison scores of the subjective test on quality aspects.

rate were indeed positive, meaning that the test participants providing those scores experienced an improvement in this aspect for the stimuli with "Premium HDR" quality. Yet there were many who either did not perceive a change in *frame rate* or experienced degradation.

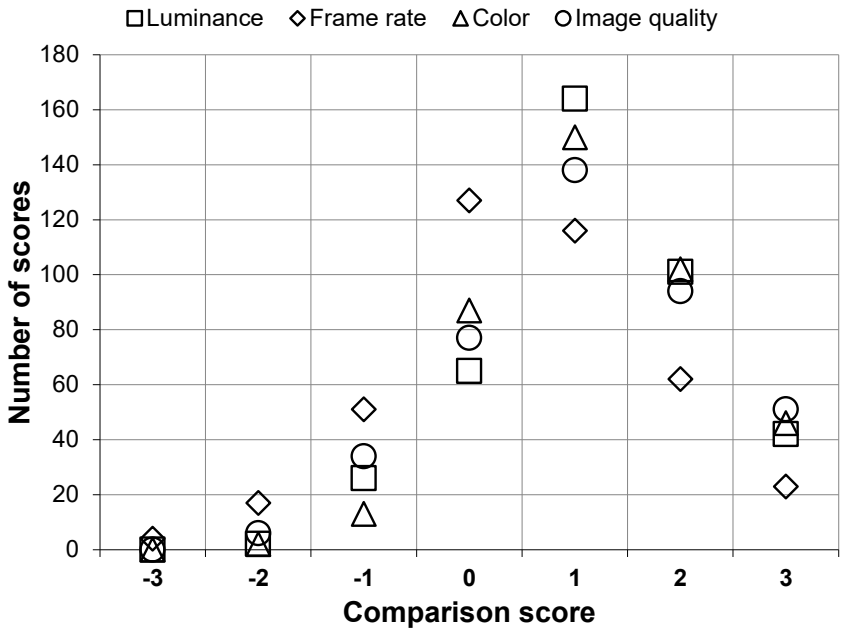


Figure 3.9: HDR: Scoring distribution of the quality aspects.

Although the experimental setup did not define any feedback beyond the comparison scores, some test participants provided us valuable insights into their visual experience. One of the test participants, who works in the movie industry, claimed that

> "The first version (Premium HDR) is always more pushed to the limits; it's actually more magical, but less controlled. The second one (HDR) feels more controlled, less magic. Personally I would go for a middle path. The frame rate doesn't seem to improve significantly."

There were also test participants who consistently experienced frame rate drops in the "Premium HDR" videos, while perceiving improvements in the other aspects. Their comparison patterns can be summarized by the following feedback:

> "It is such a pity that these incredible visuals come at the expense of frame rate. Yet to be fair, it is most certainly worth it."

The cognition originated from the concept of compensation, the idea of balance; if certain aspects become better, then their improvements negatively affect the performance of others. One could suggest that such bias might be limited to test participants with educational backgrounds of engineering or computer science, but these

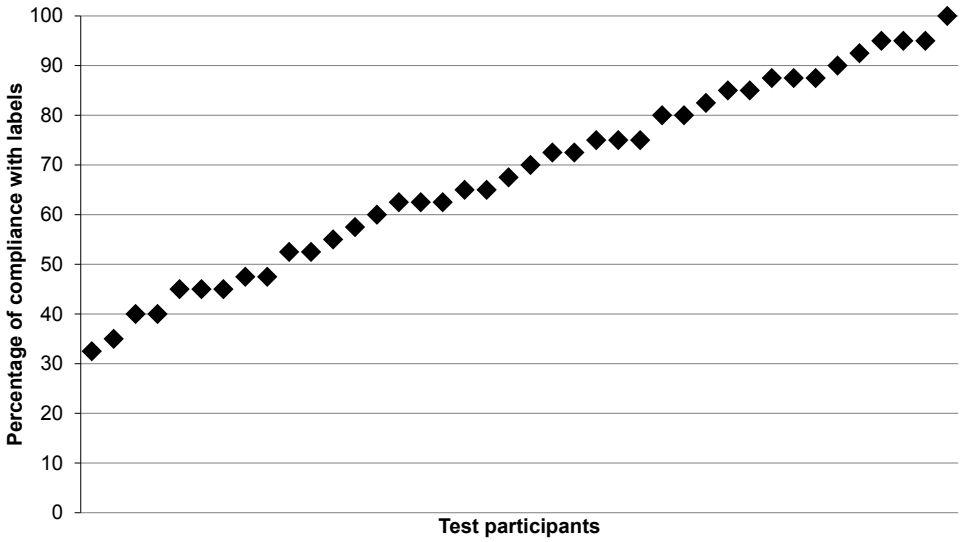


Figure 3.10: HDR: Percentage of compliance with labels in the subjective test on quality aspects.

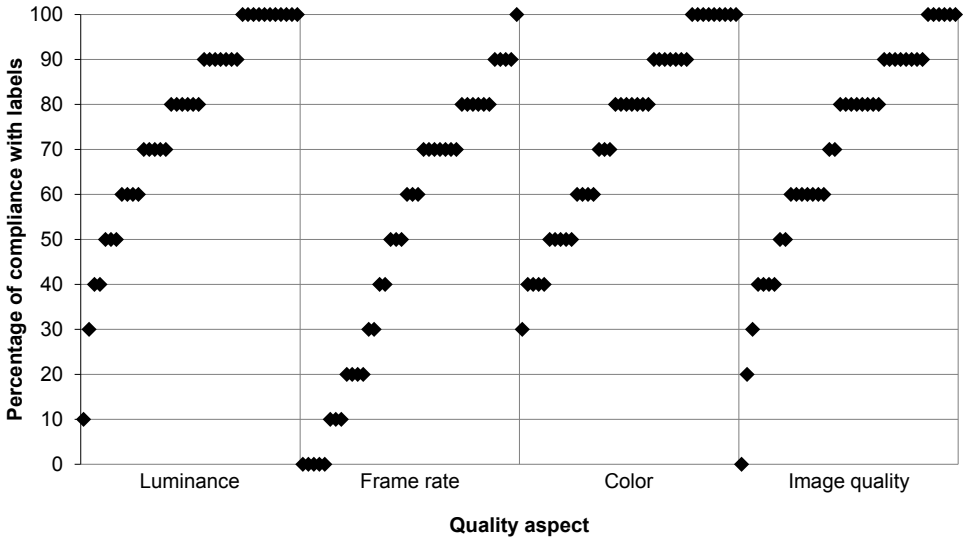


Figure 3.11: HDR: Percentage of compliance with labels per quality aspect.

patterns appeared randomly within the observer population. The impressive visuals of HDR compared to regular LDR TV experience are easier to connect with a "premium" quality when it comes to *luminance*, *color*, and even *image quality*, compared to a *frame rate* of 24 fps, when 60 fps is spreading in the everyday use case scenarios. Also, from the three highlighted aspects, *image quality* received the least positive and the most negative scores, even though it was not statistically different from the other two. Repeating the same experiment in UHD resolution is expected to boost this aspect in the positive direction.

Regarding the effect of the label order, no significant difference was found between the ratings of the two groups, and the general findings applied to this scoring separation as well. When statistically analyzing the data for each source content, the one and only case for which a significant difference was found was the *image quality* of source video sequence 1. When the "Premium HDR" was the first stimulus, the mean was 1.4, but when it was the second one, it was only 0.7. For this comparison, the p-value of the ANOVA test was 0.012. For the other 39 cases, it was above 0.05, and for 27 comparisons, it was above 0.5, even reaching 1 (e.g., *image quality* of content 3 or *color* of content 1). Therefore, based on these results, the influence of the order of labeling was not investigated in further experiments.

Finally, the compliance with labels was measured. In the context of this experiment, the decision of the test participants was considered compliant to the labels, if the "Premium HDR" stimulus was preferred (positive ratings). The overall compliance and the per-aspect compliance are shown on Figures 3.10 and 3.11, respectively. The data is visualized in the same manner as in Chapter 2; one marker corresponds to the value based on the subjective ratings of one test participant. The results indicate a rather even distribution between 30% and 100% of compliance rate, with an average rate of 68.06%. In this analysis, 100% means that the test participant preferred the "Premium HDR" stimulus for each and every source sequence and quality aspect. This applied only to a single test participant. When separated by quality aspect, we can see that a 100% of compliance was achieved by 11, 9, 6, and 1 test participants for *luminance*, *color*, *image quality*, and *frame rate*, respectively. The average rates of compliance for this order of quality aspects were 76.75%, 74.5%, 70.75%, and 50.25%. Note that in case of *frame rate*, 5 out of 40 test participants achieved a rate of 0%, which means that they either did not distinguish the stimuli or assessed the *frame rate* of the regular "HDR" stimulus to be better. All things considered, the high average rates for the three other quality aspects further reinforce the findings on the influence of the labeling effect.

3.6 Research on HDR Stalling Detection

3.6.1 Research Aim

The aim of the research was to assess the perceptual sensitivity towards a stalling event with a given duration on an HDR display.

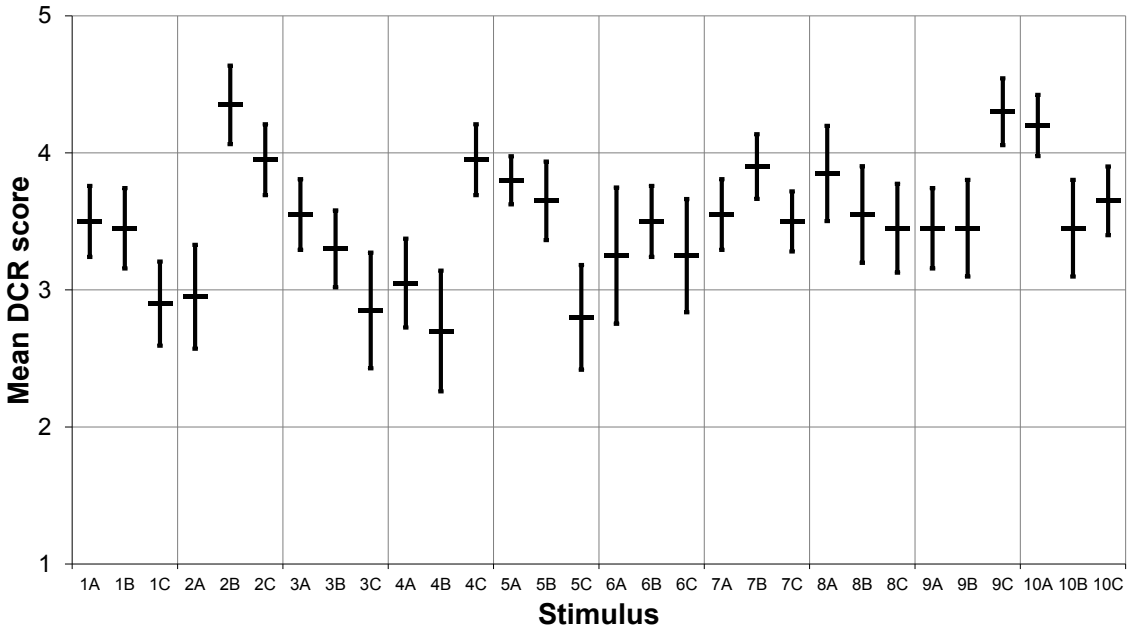


Figure 3.12: HDR: Mean DCR scores of the subjective tests on HDR stalling detection.

3.6.2 Test Conditions

The subjective test was performed using a double stimulus methodology for a paired comparison with a 5-point DCR scale. For every test condition, the test participants compared an impaired stimulus (containing a single stalling event) to the reference video. They had to assess whether the playback interruption was observable or not and, if it was, then how annoying it was (as defined by the scale).

Instead of focusing on perceptual thresholds based on stalling duration — which has already been extensively investigated in the past — the primary focus was on the content itself through TI. Thus, one single stalling duration was used for every test condition, and the stimuli only varied in content and the positioning of the event. The duration of 500 ms was chosen, which is, according to the literature, a clearly perceivable duration [95–100]. The test participants were not aware that the stalling duration was the same in every stimulus.

After the training phase, the stimulus pairs were shown in random order and were separated by 5-second blank screens. The rating task was performed directly after each impaired video stimulus. As there were 3 stalling events for 10 source videos, this means that 30 stimuli were to be assessed, each with the duration of 21.3 sec (512 frames).

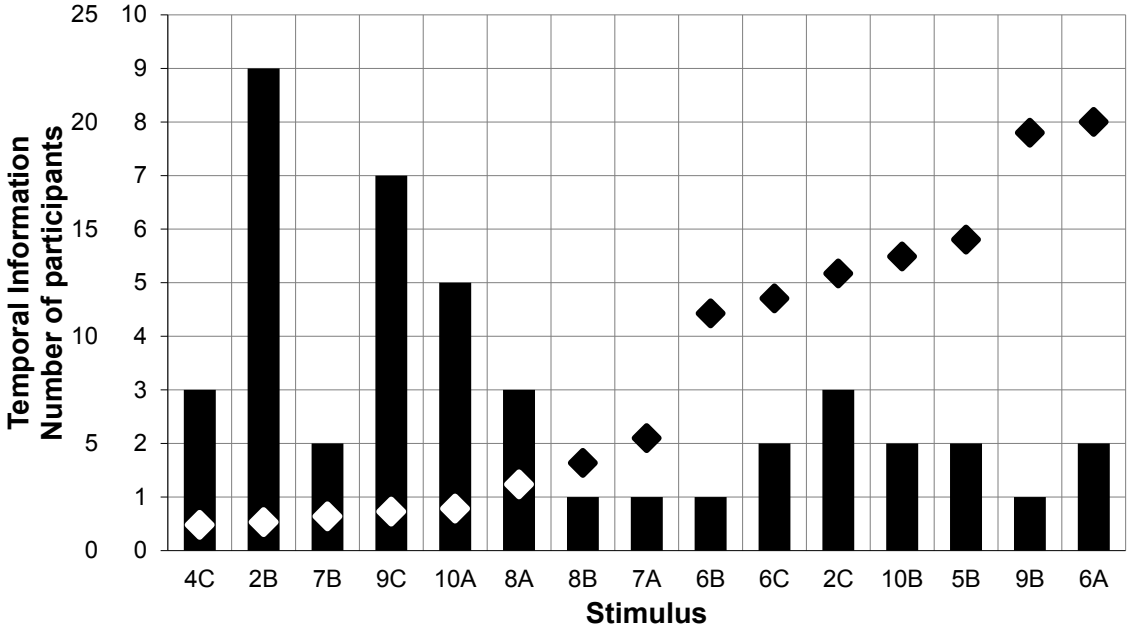


Figure 3.13: HDR: Number of test participants who assessed the given stimuli with "*Imperceptible*" ratings (bars) and the TI of the corresponding stimulus (markers).

The subjective test was followed by a post-experiment questionnaire. These questions addressed the memory bias, as test participants had to recall attributes of the stimuli they did not focus on. They were asked about the perceived variation about the aspects of *luminance*, *frame rate*, *color*, and *image quality*. Prior to the experiment, they were not informed about the questions of the post-experiment questionnaire, as these aspects would have diverted attention away from the stalling events. For each aspect, the test participants were asked whether there was a variation at all, and if there was, the number of affected contents was to be specified. The possible options were "No", "Not sure", "1–3 contents", "4–6 contents" and "7–10 contents".

As the main part of this research focused exclusively on perceptual thresholds, the term "Premium HDR" was not used during the test. The same applies to the identical test with LDR visualization.

3.6.3 Results

A total of 20 individuals participated in the tests (15 males and 5 females). The age range was from 21 to 37, and the average age was 28. 3 participants had prior HDR video experience, and the rest had never seen any HDR video before the experiment.

The mean scores are shown on Figure 3.12. Although each and every stalling event had the exact same duration (500 ms), the impact on perception varied significantly.

Figure 3.14: HDR: Frame freezing at *2B*, *9C*, and *10A*.

Table 3.2: Results of the post-experiment questionnaire.

	No	**Not sure**	**1–3 contents**	**4–6 contents**	**7–10 contents**
Luminance	2	6	8	3	1
Frame rate	0	2	7	9	2
Color	4	8	3	2	3
Image quality	6	8	5	1	0

The greatest difference can be observed in case of content 2, between *2A* (mean score 2.95) and *2B* (mean score 4.35). Again, the stalling duration was identical; however, while *2A* was a fast-paced walking motion from the right to the left, across the entire scene, *2B* was limited to subtle hand motions. As for *2C*, its TI value was nearly twice the value of *2A*, yet it received particularly high scores. *2C* was at a sudden scene change within the content, hence the spike in the TI chart. Stalling was not only well-tolerated at this frame, but also eluded the perception of 3 test participants.

Such cases, when test participants failed to perceive the 500 ms stalling event in the stimuli and provided "*Imperceptible*" as the assessment score, are summarized on Figure 3.13, displayed together with the corresponding TI values. According to this analysis, *2B* was indeed the least noticed, followed by *9C* and *10A*. These frames are shown on Figure 3.14.

Table 3.2 shows the results of the post-experiments questionnaire. The first things that really stand out from the data are that not a single test participant stated that there was no variation in *frame rate*, and that the number of unsure test participants was by far the lowest as well. In fact, nearly half, 9 out of 20 test participants stated that at 4, 5 or 6 contents contained *frame rate* variations. *Image quality* was the clearly least affected by the memory bias, followed by *color* and *luminance*.

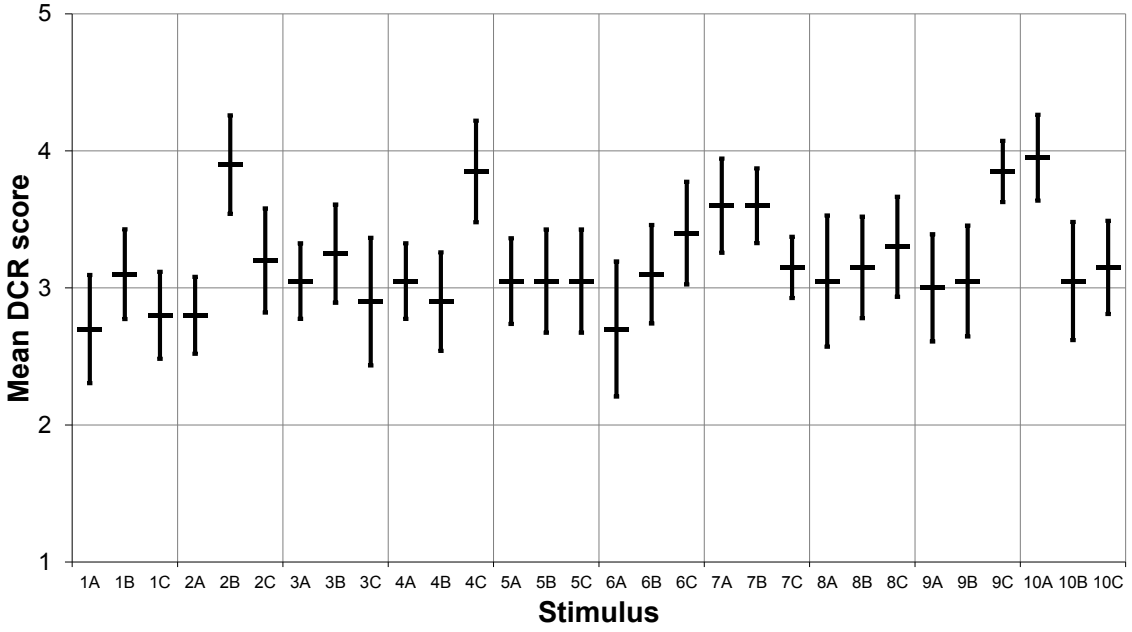


Figure 3.15: HDR: Mean DCR scores of the subjective tests on LDR stalling detection.

3.7 Research on LDR Stalling Detection

3.7.1 Research Aim

The aim of the research was to assess the perceptual sensitivity towards a stalling event with a given duration on an LDR display, and thus, serve as a comparison to the previously introduced experiment.

3.7.2 Test Conditions

The test conditions were identical to the parameters of the research on HDR stalling detection. The only differences were the display and the bit depth of the stimuli.

3.7.3 Results

A total of 20 individuals participated in the tests (18 males and 2 females). The age range was from 21 to 60, and the average age was 29.7.

The mean scores are shown on Figure 3.15. At first glance, the figure indicates that the obtained scores of several test stimuli were lower than what was achieved for HDR stalling detection, and variations were smaller as well. To be precise, while the average of all HDR scores was 3.5, the corresponding value for LDR was 3.19. This suggests that the stalling events in the HDR experiment were more difficult to

perceptually detect and/or they were more tolerable, compared to the LDR experiment. However, in order to draw any conclusion, a direct comparison with statistical analysis is required.

3.7.4 Comparison of HDR and LDR Stalling Detection

Figure 3.16 compares the scoring distributions and the aforementioned means of the two experiments. The latter indicates a significant difference, as the 0.95 CIs do not overlap. This difference is well-reflected in the scoring distributions. Since both subjective studies addressed stalling detection, the most important DCR score in this analysis is 5 ("*Imperceptible*"). While the HDR experiment produced 44 of this score, this was only 16 in case of LDR.

Does this mean that compared to conventional LDR visualization, HDR stalling events were more difficult to detect in general? Not necessarily. In order to gain more insight, let us compare the distribution of these scores particularly. Figure 3.17 shows the number of test participants using the "*Imperceptible*" rating option for the given test stimuli, separately for LDR and HDR. The results show that every HDR stalling event received as least as many "*Imperceptible*" ratings as LDR did. The greatest differences were measured for *2B* and *9C*, which were the two least detectable stalling events in the HDR study (see Figure 3.13). These findings indicate that difficult-to-perceive stalling events (with minimal amounts of variation between adjacent frames) may go unnoticed during HDR visualization, but the same is less likely to happen in case of LDR.

It is important to note that the findings presented so far do not mean that each and every test stimulus differed significantly. The statistical analysis of the conditions is presented in Table 3.3. We can see that for 9 out of 30 cases, the difference was statistically significant. In all of these cases, HDR visualization achieved significantly higher scores, thus, these stalling events were more difficult to detect and/or easier to tolerate. The differences between them on the scale from 1 to 5 were at least 0.35, but for *8A*, it was 0.8. Some of these frame repetitions were rather subtle — like the ones presented on Figures 3.13 and 3.17 — while some others were quite obvious.

The results of the comparison do not correlate with TI due to the aforementioned diversity, but they are most definitely connected to the so-called "visual awe". Let us take *1A*, *5B* and *10C* (see Figure 3.18) as counter-examples for the idea that hard-to-detect, low-TI stalling events differ more between LDR and HDR visualization technologies. All of these stalling events had high TI values, as shown on Figures 3.2 and 3.3. *1A* captured a vertical camera panning during a highly dynamic scene, *5B*

Table 3.3: Statistical analysis of the conditions (c) of the LDR and HDR stalling detection; each line compares the results of a given test condition for the two experiments. The p-value of ANOVA is given (p), along with significance (S).

c	**p**	**S**
1A	0.002	yes
1B	0.123	no
1C	0.657	no
2A	0.538	no
2B	0.059	no
2C	0.002	yes
3A	0.013	yes
3B	0.828	no
3C	0.876	no
4A	1.000	no
4B	0.495	no
4C	0.661	no
5A	0.000	yes
5B	0.016	yes
5C	0.364	no
6A	0.130	no
6B	0.080	no
6C	0.600	no
7A	0.818	no
7B	0.108	no
7C	0.033	yes
8A	0.010	yes
8B	0.131	no
8C	0.547	no
9A	0.075	no
9B	0.148	no
9C	0.011	yes
10A	0.203	no
10B	0.162	no
10C	0.024	yes

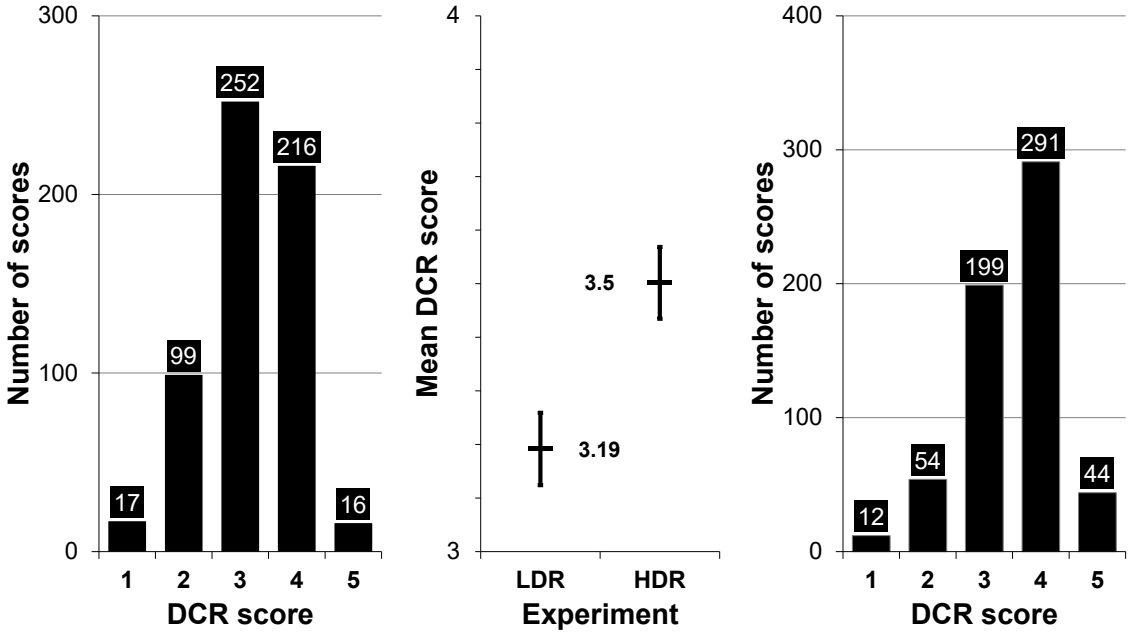


Figure 3.16: HDR: Scoring distribution of the LDR (left) and the HDR (right) experiment on stalling detection, and their mean scores (middle).

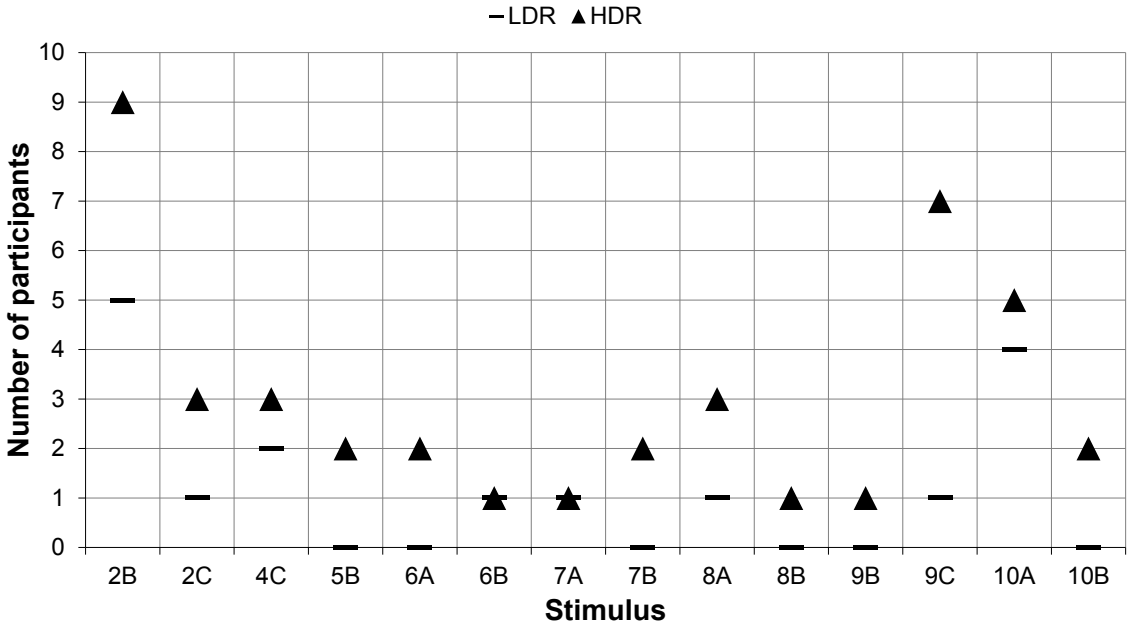


Figure 3.17: HDR: Number of test participants in the LDR and HDR tests who assessed the given stimuli with "*Imperceptible*" ratings.

was a closeup on the lit bonfire with added movement on the right, and *10C* also captured camera movement, during the visually intense moment of welding. Therefore, these stalling events were difficult to miss (yet for *5B*, two test participants actually

Figure 3.18: HDR: Frame freezing at *1A*, *5B* and *10C*.

managed to, during the HDR test, as shown on Figure 3.13), but they were all visually impressive. To be more precise, they were visually impressive when shown as HDR contents on an HDR display.

What was also common in them is that the stalling event itself was not *too* irritating. Let us now examine *3C*, with its 1-frame flash of light (see Figure 3.4). The mean scores for the LDR and the HDR tests were 2.9 and 2.85, respectively, not a single test participant deemed it "*Imperceptible*", only 6–7 found it not to be annoying, and the worst score "*Very annoying*" appeared twice in both experiments. Similar assessments were applied to *3B* as well, which also repeated the selected frame amidst sudden flashes, and the achieved means were 3.25 and 3.3. The reason why source video 3 ("*Carousel Fireworks*") is a good example for the very similar ratings in both experiments, is that it had the most significant contrast due to the pitch-black night sky and the exceptionally bright fireworks. Yet the test participants were similarly annoyed, regardless of visualization. However, *3A* — which was before the bright flashes, and therefore, the visual awe was not disturbed by a highly annoying stalling position — was rated differently for LDR and HDR (means of 3.05 and 3.55, respectively), and, in fact, the difference was statistically significant.

3.8 Research on HDR Stalling Duration

3.8.1 Research Aim

The aim of the research was to assess the impact of the labeling effect on the perceived duration of stalling events.

3.8.2 Test Conditions

For the indication of difference in perceived stalling duration, a 7-point scale was used ("*Much shorter*", "*Shorter*", "*Slightly shorter*", "*Same*", "*Slightly longer*", "*Longer*", "*Much longer*"). Based on the results of HDR stalling detection, for each source video,

Table 3.4: Selected stalling events.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	X	X	X	-	X	-	-	X	X	X
B	-	X	-	X	-	X	X	-	-	X
C	X	-	X	X	X	X	X	X	X	-

two stalling events were selected: the easiest and hardest to detect and tolerate. Table 3.4 shows these selected stalling events. Each stalling was included twice, once with a duration of 500 ms and once with 1000 ms. Therefore, each source video was assessed 4 times, and thus, 40 comparisons were made. Labeling was present in the experiment, in a similar manner as in the research on HDR quality aspects; the utilized mock-up methodology here was the same as before.

3.8.3 Results

A total of 36 individuals participated in the tests (22 males and 14 females). The age range was from 20 to 42, and the average age was 26. 8 participants had prior HDR video experience, and the rest had never seen any HDR video before the experiment.

The obtained subjective scores are represented by their numerical counterparts, ranging from -3 to $+3$. During the subjective tests, the test participants were presented a combination of the available qualitative tags for stimulus comparison — defined in the previous subsection — and these values, emphasizing a uniform distance between the values of the scale. In this analysis, positive values indicate longer perceived stalling durations for the "Premium HDR" stimulus, while negative values indicate that it was perceived as the shorter one.

With 36 test participants and 40 comparisons, a total of 1440 scores were collected. Figure 3.19 shows the distribution of these scores. It is apparent that the labeling effect had a significant impact on the perception of stalling duration. Only in 22.6% of the ratings indicated no perceived difference between the identical video stimuli, which is very similar to the scoring distribution of the experiment on quality aspects (22.25%, see Figure 3.7).

The obtained ratings are decisively positive (59.2%), which means that the stimuli labeled as "Premium HDR" was generally perceived to have longer stalling events. The most common score by far was +1 (38.5%), indicating slightly longer stalling events for the "Premium HDR" stimuli. Negative scores are present as well in the

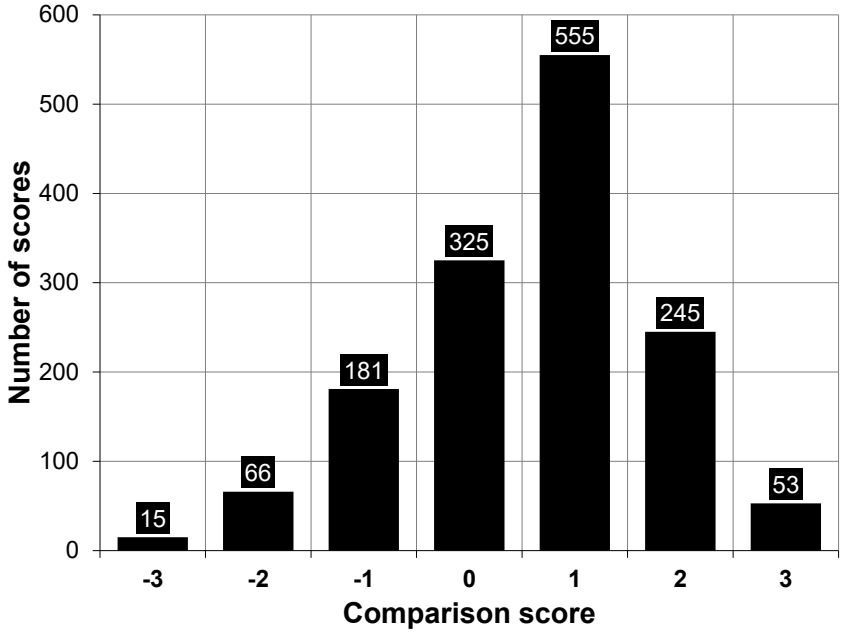


Figure 3.19: HDR: Scoring distribution of the subjective tests on stalling duration.

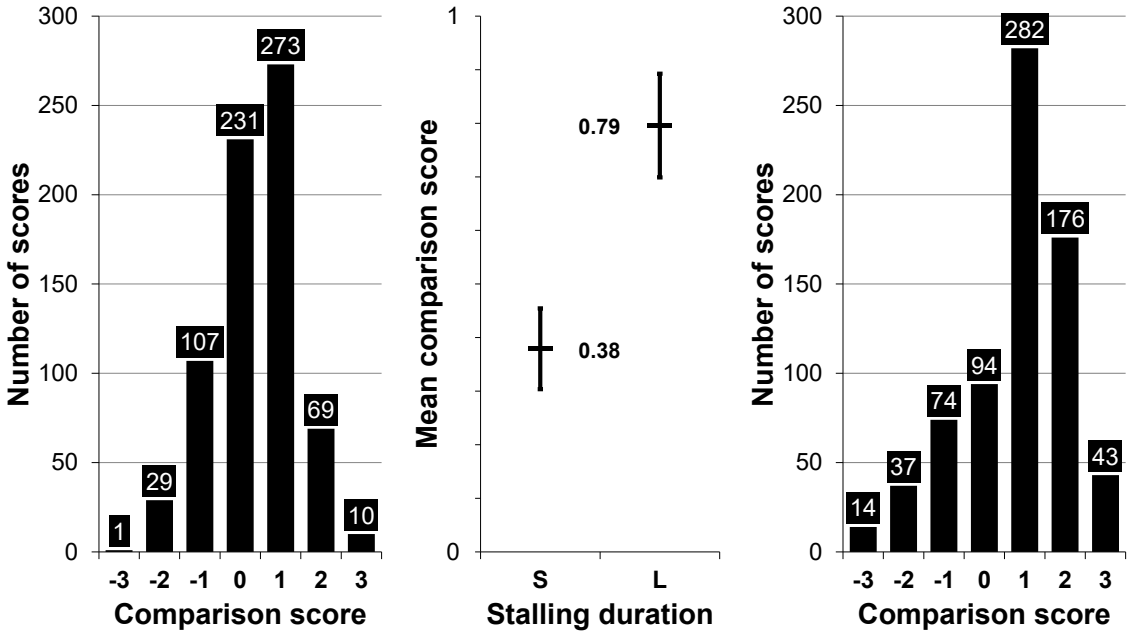


Figure 3.20: HDR: Scoring distribution of short (left) and long (right) stalling events, and their mean comparison scores (middle).

analysis (18.2%), but the number of -2 and -3 is particularly low (5.6% combined), while the same cannot be said for the corresponding positive scores (20.7% combined).

In this experiment, two different stalling durations were used. Figure 3.20 shows their separate scoring distributions and their mean comparison scores. The results

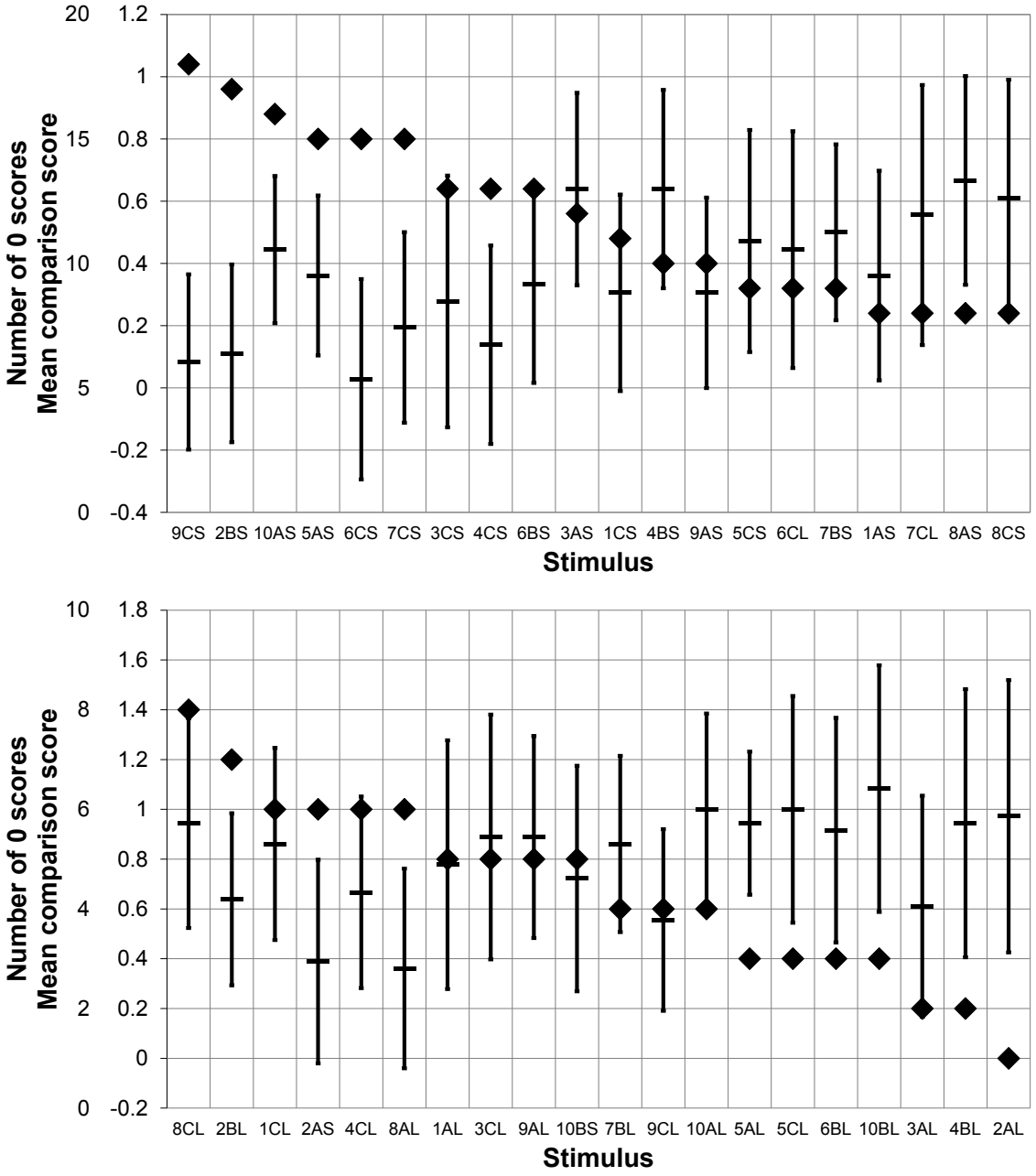


Figure 3.21: HDR: Number of 0 scores (markers) and mean comparison scores (intervals) of the research on stalling duration.

clearly indicate that the bias in perception was significantly stronger for the video stimuli with longer stalling durations. While 32.1% of the scores of the stimuli with short stallings report the lack of difference, this is only 13% for long stallings.

Figure 3.21 shows the number of 0 ("Same") scores for each test stimulus (ranging from 0 to 20 and 10 in the top and bottom part of the figure, respectively), and the mean comparison scores with 0.95 CI. The test stimuli are sorted by the number

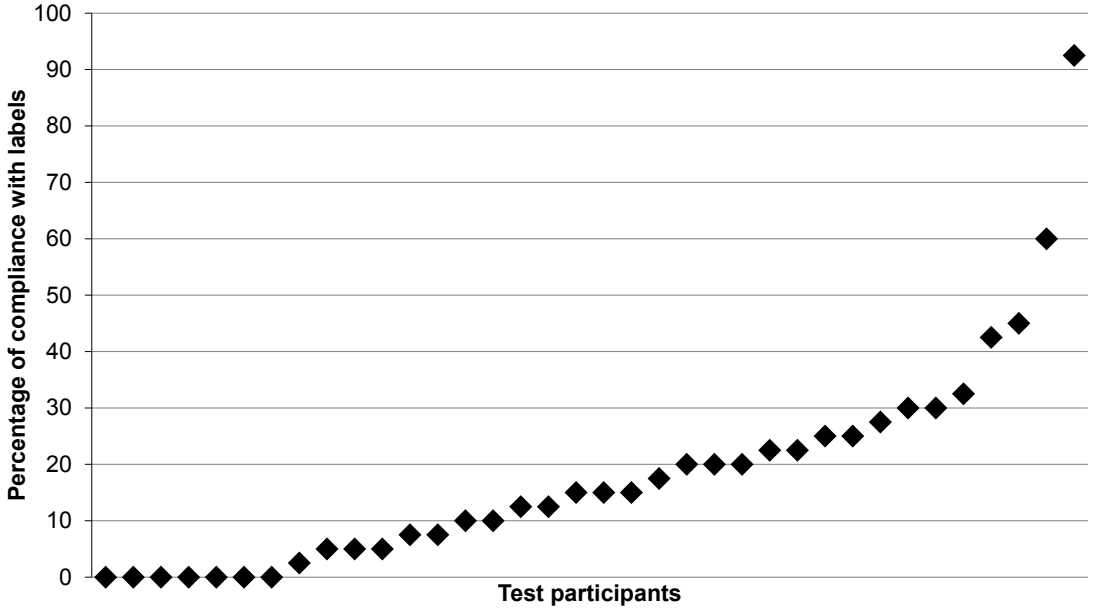


Figure 3.22: HDR: Percentage of compliance with labels in the subjective test on stalling duration.

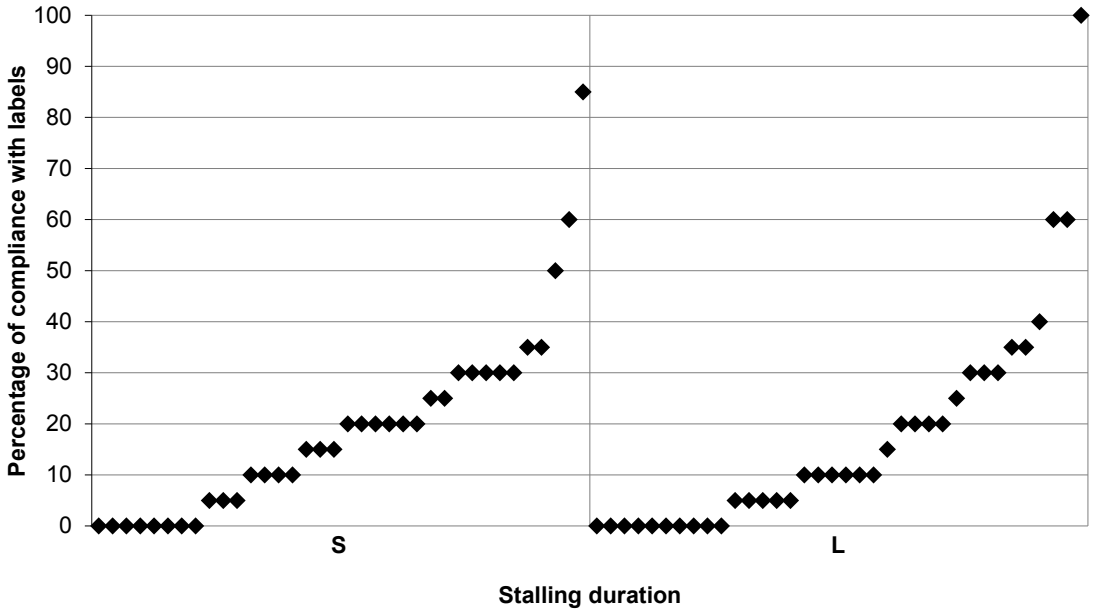


Figure 3.23: HDR: Percentage of compliance with labels for short (S) and long (L) stalling events.

of 0 scores in descending order. The highest numbers of 0 scores were achieved by *9CS*, *2BS*, and *10AS*, which were the stimuli with the least detectable and least annoying stalling events (see Figures 3.12, 3.13, and 3.14). The findings extracted from Figure 3.20 apply here as well, since the upper half of the descending order of

0 scores is dominated by short stalling events. In accordance with the distribution of Figure 3.19, mean scores rise as 0 scores get lower, however, statistical differences are difficult to find, due to the large scoring deviations. Note that it is likely in such experiment that while a specific test participant rates a given stimulus with +3, a different participant may rate it as −3. Standard deviation at the upper end of this order (highest numbers of 0 scores) is only 0.7–0.8, while at the other end, it is 1.6–1.7.

Lastly, the compliance with the labels is addressed. As "Premium HDR" is basically a positive label, the compliance rate is based on the ratings that indicate a shorter perceived stalling duration for the "Premium HDR" stimulus. The overall percentage of compliance and the compliance rate separately investigated for the different stalling durations are shown on Figures 3.22 and 3.23, respectively. The data is visualized in the same manner as earlier in the chapter; one marker corresponds to the value based on the subjective ratings of one test participant. The average compliance rate was 18.19%, with 7 of 36 test participants who never found the stalling event of the "Premium HDR" stimulus to be shorter than the other one. 31 test participants had a compliance rate of 30% or less, and only two overall rates were above 50%. In comparison, 31 out of 40 test participants had the corresponding value above 50% in the experiment on quality aspects, and not a single individual had an overall rate below 30%. Regarding the separation based on stalling duration, the average rates for the short and the long stalling events were 19.03% and 17.36%, respectively, and one test participant reached a 100% rate for the stimuli with the long stalling events. The low compliance rates, in general, indicate that the vast majority of test participants did not believe that a format with superior visualization quality should have shorter stalling durations. In fact, the common concept (or rather the common preconception) was actually the opposite, as shown by the results.

In conclusion, the obtained results indicate a strong presence of the labeling effect on perceived stalling duration. The effect is stronger for longer stalling events, and cognitive bias is less significant if the given stalling event is difficult to perceive in the first place. Both directions of distortions are represented in the collected data, but the preconception stating that "Premium HDR" should have longer stalling events is significantly more dominant.

3.9 Chapter Summary

The chapter presented four experiments on HDR video QoE. The results of the study on quality aspects concluded that an aspect that is not evidently connected to a clearly positive label may be penalized by the viewers through the preconception, the idea of trade-off, of compromise. The outcome of the experiment on stalling event duration was analogous in results to the one on quality aspects, and, in fact, showed more consistency in penalization, as the "Premium HDR" video sequences were deemed to have longer stalling events — even though there was no difference in stalling duration whatsoever.

The other two subjective tests on LDR and HDR stalling event detection demonstrated that the "wow effect" due to the novel nature of the visualization technology may result in additional cognitive load, and therefore, certain harder-to-detect stalling events may have a better likelihood to go unnoticed. In the following chapter, light field technology is addressed, which has not emerged yet on the consumer market and people, in general, are commonly unaware of this form of visualization, or at least the majority has not seen anything like it in real life. The resulting "wow effect" was taken into consideration in the experimental designs of the subjective tests on light field QoE.

Chapter 4

Light Field Visualization

As light field is gaining more attention among researchers and developers, the spelling of the term is slightly changing over time. Traditionally it is spelled as "light field", as it is a field of light rays, but "light-field" and "lightfield" are appearing in disseminations of knowledge as well. In this thesis, only "light field" is used as the consistent spelling of this technical term, although note that in some recent publications of the work with Holografika, "light-field" was used.

4.1 Introduction

The concept of light field first emerged over a hundred years ago through the work of French physicist Gabriel Lippmann on integral imaging [101]. The name of this technique originates from Lippmann's term "photographie integrale", which can be directly translated as "complete photography" or interpreted as "complete imaging". Its motivation was and still is the incomplete nature of 2D visual representation; the 3D world we live in simply cannot be fully embodied in a flat, 2D image.

Lippmann's technique uses an array of microlenses — in a layout similar to the eye of the fly — which enables the human observer to perceive the captured scene in a way that depends on direction and position. What it visualizes is actually a light field: a function that defines light radiance (intensity) in every direction — in which light flows — through every single position. The technical term "light field" was first defined by Andrey Gershun, in his publication titled "The light field" [102].

In order to describe a light field, we need to identify position and direction. Position in the 3D space can be defined by three parameters (x, y, and z coordinates), and the 2D orientation of the vector is determined by two angles (θ and ϕ). These five parameters together construct a 5D function, which is known as the 5D plenoptic function [103]. Higher dimensions of the function can also be defined, i.e., the 7D

plenoptic function, which also takes time (t) and wavelength (λ) into consideration. By doing so, every direction from every position is represented at every moment and at every wavelength. It is, of course, important to state that the plenoptic function — regardless of dimension — can only represent visual information if it is restricted to geometric optics, thus, light in this model is non-coherent (incoherent), which means that not all light rays share the same phase and their wavelengths differ as well (coherent light is, i.e., laser). Also, the physical objects in the 3D space need to be greater than any of the light's wavelengths; otherwise they would not be big enough to be visually represented.

With all this information, light field visualization brings a more complete representation of the real world, and thus, enables a better, more immersive visual experience. However, such powerful representation also comes with a significantly increased data requirement. Fortunately, there are several ways to simplify this function.

First of all, since the light field is meant to be observed by humans, the HVS shall be taken into consideration, so it is enough to visualize the three prime colors (red, green and blue). This reduces the plenoptic function by one dimension, because instead of having wavelength as a separate parameter, the intensities of the prime-color lights are given by three functions (one function for each prime color).

The dimension of time can also be excluded from the function. As an example, if we want to visualize a light field video, then similarly to the 2D scenarios, a video is technically a series of still images, whose sampling frequency is defined as frame rate. Thus, it is sufficient to describe one specific snapshot of the light field per frame, without the dimension of time.

Last — but definitely not least — comes a simplification based on the constraints of geometric optics. If we measure the radiance of the ray on the enclosing volume's surface, there is no way to differentiate optically between a light ray hitting a surface point due to a complex series of scattering, reflections, refractions and other photonic quantum effects, and a ray originating directly from a photon emitter pointing towards the surface point along the ray, if the radiance of the ray at the point of measurement is the same. Therefore, we can substitute all light rays to photons emitted from photon emitters located in the enclosed volume, as in vacuum these light rays will have the convenient property that their radiance is constant along the ray. Thus, the radiance along the ray can be substituted with a constant and we can reduce the dimensionality of the plenoptic function by one. However, this final method of simplification is not performed by simply eliminating a parameter from the plenoptic function and keeping the rest unmodified (i.e., wavelength and time). Instead of

describing a ray by defining three coordinates and two angles — like in case of the 5D function — a ray shall be determined by two positions in the 3D space. One could immediately argue that the 2 positions shall make the plenoptic function 6D and not 4D — as a position requires three coordinates in space — and also point out the redundancy in this parametrization, since any position pair along a given vector is suitable to describe it. What we actually do is define 2 parallel planes, one behind and one in front of the scene — given that the scene has a finite size — and accommodate the positions on these planes (one on each plane). This means that a position can be given by two coordinates, thus, four coordinates (s, t, u, and v) for the two positions. A position pair on the two planes shall identify exactly one vector (and no other pair can identify that same vector), and the radiance at these positions is the same on any given position of the ray between the planes [104] [105]. About the planes themselves, it is irrelevant where we place them, e.g., the distance between the farthest object in the scene and the plane behind it shall not affect the usability of this parametrization, even though the coordinates on the plane for a given ray will be different. One could think that this is the one and only way to parametrize the vector, using the two parallel planes. It is indeed a convenient and easy-to-visualize method, commonly used in the literature. However, if the scene is properly, fully enclosed by the two planes, they do not even need to be parallel; one point — defined by a coordinate pair — on each plane still uniquely determines a vector. In fact, they do not even need to be planes. Any surface can be used, as long as they fully enclose the scene and allow the unique determination of vectors by point pairs. If we wanted to imagine a surface which is unusable for this purpose, one of the easiest example is to think of a rug which is warped or partially rolled up; in this case more than one pair of points can determine a vector. On a practical level, cylinders and spheres are suitable for the implementation of this representation, when we want visualization to be position-dependent in either one way or both ways, respectively.

With all these simplifications, we obtain a four-dimensional representation via a plenoptic function, commonly known as a 4D light field [105], but also often referred to as Lumigraph [106]. A light field is basically a 3D vector field, as it is a collection of 3D vectors, where a vector represents a ray of light. As described earlier, the intensity along a vector is defined to be constant. So a light field is a 3D vector field, however, it is important to note that the plenoptic function is not a vector function, because it returns a scalar value — this aforementioned light intensity — and not a vector. It could return a vector if we defined it in a way that it returns the values for the three prime colors (RGB), but we can simply have a separate function for each color.

Using these principles, manufacturers can now build light field displays, offering a visual experience that fundamentally differs from what we can observe during everyday multimedia consumption in this day of age. This chapter of the thesis studies the QoE of the novel visualization technologies of light field. In particular, a series of subjective tests were carried out on HoloVizio displays of Holografika, which employ holographic diffusers to create the glasses-free 3D experience. There are other light field visualization technologies as well, like the holographic walls of Light Field Lab[1], however, in the scope of the thesis, lenticular displays, such as the Alioscopy 3D display[2], hogel optics, such as the FoVI3D display[3], and near-eye light field displays, such as the ones presented by Lanman *et al.* [107] and Hansen *et al.* [108], are not covered.

As a quick summary of these other visualization technologies, Alioscopy 3D displays use 720 cylindrical micro lenses, forming a so-called lenticular array, which covers the LCD panel, directing the appropriate view to the appropriate angle, and thus, the two eyes of the observer encounter different views. The hogel optics (originating from the words "holographic" and "element") of the FoVI3D display is a lens element positioned directly above a multitude of pixels, capable of visualizing roughly 50 views, but only a single pixel can be observed from a given viewing angle. In case of near-eye light field displays, the user practically wears the display itself, similarly to VR solutions. It also uses a micro lens array that transforms pixels into light rays, and thus, creates a dense light field over the eyes.

Before introducing the experiments on light field QoE, first, it is important not only to understand what a light field is, but also to review the parameters of the displays and the contents which contribute to visualization quality. The following section provides a comprehensive, detailed overview of the attributes of the technology at hand.

4.2 The Key Performance Indicators of Light Field Visualization

In recent years, multiple solutions regarding 3D visualization established a presence on the consumer market, the majority of which is fundamentally dependent on viewing devices, i.e., special glasses and headgears. Not only does such equipment limit

[1]https://www.lightfieldlab.com/

[2]Alioscopy 3D UHD 84" LV display:
http://www.alioscopy.com/en/datasheet.php?model=Alioscopy_3D_UHD_84_LV

[3]http://www.fovi3d.com/

the number of simultaneous viewers of a given content, but it also poses numerous inconveniences and issues. Light field displays offer glasses-free 3D experience, as no such equipment is required, therefore, any number of viewers can simultaneously observe the content without the inherent problems of other technologies.

Of course, this does not mean that any light field content visualized on any light field display is de facto superior in visual experience to any other 3D technology. There is a long list of parameters that directly affect the quality of light field visualization, regarding both display and content. They are Key Performance Indicators (KPIs), as these parameters define the objectively and subjectively measurable visual performance, and any of them having insufficient characteristics may severely degrade the overall quality.

In this section, the KPIs of light field visualization are systematically reviewed. The properties of displays and contents are detailed separately, but their interdependencies are taken into account as well. For each parameter, the state-of-the-art scientific results are discussed, and use-case-aware recommendations for display manufacturers and content providers are made. This section also discusses future research efforts regarding capabilities that are not even available for the high-end systems of the present day but hold notable potentials for the light field displays of tomorrow.

4.2.1 Display Parameters

This part of the section reviews the parameters of display systems that are relevant to light field visualization quality, addresses their contributions to visual performance, analyzes how they affect each other, details how insufficient properties manifest in practice, provides recommendations to achieve visual excellence and introduces the state-of-the-art related work.

A light field display can either be a front-projection system or a back-projection system. Front-projection systems have projectors on the same side of the screen as the observer, and as light rays are basically reflected from the screen to the eyes of the observer, these systems are also known as reflective displays. Back-projection systems evidently have projectors at the other side of the screen, and therefore, they are transmissive displays. Regardless of the location of the projectors, display parameters can be categorized into those that are derived from the physical setup of the system, and those that are defined by projection.

4.2.1.1 Physical Setup

4.2.1.1.1 Screen Dimensions

One may think that discussing the physical dimensions of the screen is an unnecessary triviality. However, these parameters affect the visual performance of the system much more than just how big the screen is that the observer is looking at. The size of the screen fundamentally determines most system requirements; bigger screens demand higher capabilities. One must note that the system scales up together with the screen, and it may also directly influence other parameters, such as physical depth. As an example, for reflective systems, having the same screen width but a different screen curvature (curved shape of the screen) results in a different field of view (FOV). This will be explained in detail later in this section, together with other attributes that have an effect on the FOV. For both back- and front-projection systems, the spatial requirements can be prohibitive for deployment and practical use; having a bigger screen scales up the physical size of the projection subsystem as well.

In practice, the screen of light field displays may vary a lot in physical dimensions; one could say that they appear in various shapes and sizes, and they actually do. The smallest light field display implemented and used in research was the one appearing in the work of Adhikarla *et al.* [109], with a 8.6-inch screen. The largest systems in the related literature were upscaled designs of a light field cinema [110], proposed by Kara *et al.* [111]. Different variations appeared in the publication in 450-inch, 540-inch and 630-inch sizes. Among the large-scale implementations of multi-view and super multi-view technologies today, Lee *et al.* [112] designed a system with a 100-inch screen, and Inoue *et al.* [113] worked on a 200-inch display, similarly to Kawakita *et al.* [114]. At the time of writing this thesis, the largest commercially available light field display is the HoloVizio C80 cinema system [110], with a 140-inch display.

4.2.1.1.2 Spatial Resolution

The spatial resolution of light field displays is often labeled as the 2D-equivalent resolution of the system. It is a general statement in the literature that the concept of pixels does not apply to such systems, as light rays hit irregular positions on the screen. In a way, we can indeed talk about pixels in the context of light field displays, however, it most certainly does not apply to them in the way we know it for 2D displays. Due to this aforementioned irregular nature of light ray propagation, the grid of pixels is far from being uniform. Furthermore, even though we can identify

pixels, the position, color, and intensity of a given pixel is direction-selective, which means that the perception of the pixel depends on the angle of observation.

Insufficient spatial resolution for light field displays does not result in the blockiness that is apparent for conventional 2D displays. Instead, visualization is affected by blur. It is important to note that the blur that applies to such displays is not uniform across the screen. The amount of perceived blur is determined by pixel density — measured in pixels per inch (ppi) or pixels per centimeter (ppcm) — which also depends on the screen size. The typical values of ppi for light field displays are between 10 and 50. For example, due to the large screen of the C80[4], it only has a ppi of 10.8, while the smaller screens of the HoloVizio 722RC[5] and the 80WLT[6] enable ppi values of 22.6 and 47.2, respectively.

The smallest spatial resolution that applied to a fully-implemented system was 320×240, with a screen size of $144 \times 81\,\text{mm}$. Common values in practice include 1024×768 (C80) and 1280×720 (722RC). The highest spatial resolution of a light field display at the time of writing this thesis is 1920×1080, which applies to an experimental system of Holografika, that is not commercially available yet.

Kovács *et al.* [35, 115] performed measurements regarding the spatial resolution values of light field displays. The proposed method of measuring display capabilities uses sinusoidal patterns with increasing frequency, which is displayed on the screen, captured and analyzed in the frequency domain. The procedure is fully automatic for spatial resolution and does not require any camera movement — in contrast to angular resolution, which is also addressed by the research of the authors. Recommendations regarding the general techniques of such measurements are provided by the International Display Measurement Standard (IDMS)[7]. The IDMS covers measurements related to several other parameters as well, such as angular resolution and FOV.

4.2.1.1.3 Angular Resolution

The angular — or rather angle-dependent — nature of light field displays means that one shall see a different view of the visualized content from a different angle

[4]HoloVizio C80 light field cinema system:
https://holografika.com/c80-glasses-free-3d-cinema/

[5]HoloVizio 722RC light field display:
https://holografika.com/722rc/

[6]HoloVizio 80WLT light field display:
https://holografika.com/80wlt/

[7]International Committee for Display Metrology (ICDM) and the Society for Information Display (SID): Information Display Measurements Standard (IDMS) v1.03

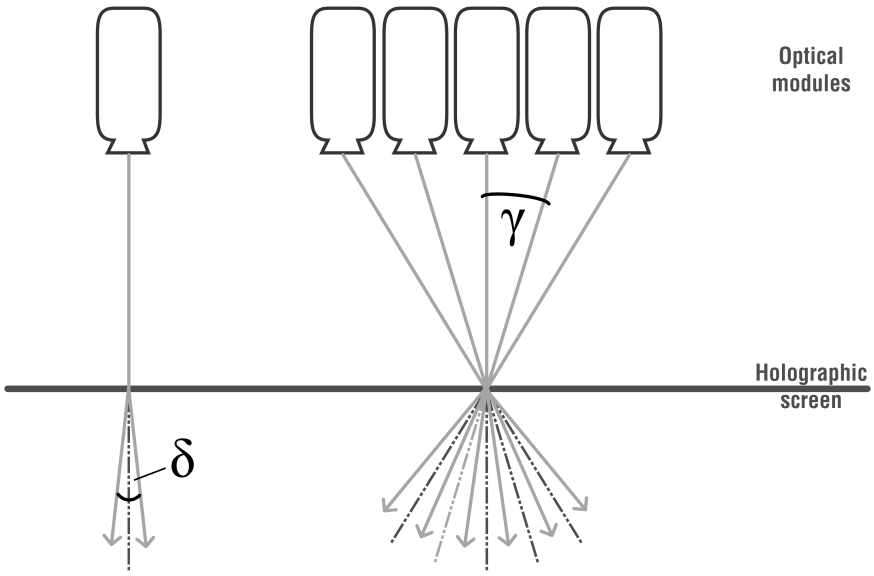


Figure 4.1: LF: Illustrations of angular resolution definitions.

of observation. This visual phenomenon is analogous to what can be seen in the real world, as light field displays aim to provide the parallax effect. It means that portions of the visualized content that are farther away from the observer change the perceived position slower than those which are closer. This effect applied to the horizontal axis is known as the horizontal parallax, and vertical change relies on the vertical parallax. The currently available systems are horizontal-parallax-only (HPO) light field displays, and future development is converging towards full-parallax (FP) displays. Displays providing only one of the two parallax components (horizontal and vertical) are also known as half-parallax displays. It is important to note that the parallax effect provided by such displays is continuous in the entire FOV, while multi-view displays show angularly repeating contents that are only observable from certain positions, known as "sweet spots".

Angular resolution is technically the resolution of angular change that only applies to the horizontal axis in case of HPO displays. To be more precise in defining this term, it is the minimal angle of change that rays can reproduce with respect to a single point on the screen [116]. Figure 4.1 illustrates the definitions of angular resolution [117,118]. γ is the common understanding of mechanical angular resolution, which is the angle between horizontally adjacent light rays targeting a given point on the holographic screen. δ is the corresponding angle for distinct light rays leaving the surface of the screen. These two angles are not equivalent, as δ depends on

the scattering characteristics of the screen. In practice, although both values are either calculated or measured, it is γ that describes the angular resolution of the display system. The calibration of δ is used to improve the homogeneity of light field visualization. For visually efficient light field displays, the value of γ is close to δ, but δ should always be slightly greater. However, if δ is much greater than γ, than this jeopardizes angular resolution, as the high extent of scattering mixes up the rays that were originally in correct horizontal order before reaching the screen.

Generally speaking, angular resolution defines the smoothness of the parallax effect. This is by definition the smoothest in the plane of the screen. Therefore, the more a given content comes out of the screen, the higher angular resolution is required to sufficiently support its angular smoothness. Furthermore, the perceived angular resolution depends on the distance of observation (measured from the plane of the screen), thus, it also determines the viewing distance supported by the display.

Therefore, a system with low angular resolution may not be able to properly display a content with greater depth in good quality, and observers viewing the display from a certain distance may not even experience the 3D-nature of light field visualization. Building a system with insufficient angular resolution may also result in artifacts and visual phenomena such as the crosstalk effect, which means that adjacent views overlap each other in a semi-transparent manner, and the total lack of parallax smoothness can create sudden jumps between given views. Furthermore, a δ that is much greater than γ also results in the crosstalk effect, due to the previously mentioned reason.

The lowest angular resolution in practice was 2.25 degrees, which belonged to an experimental system of Holografika. The commercially available HoloVizio 80WLT has an angular resolution of 1.5 degrees, and the highest angular resolution currently in use is 0.5, applying to the C80.

4.2.1.1.4 Depth Budget

The depth budget is a distance vector perpendicular to the plane of the screen, and measures how much the content can come out from this plane. It is more-or-less symmetric, which means that this distance is approximately the same for the positive (towards the observer) and the negative (away from the observer) direction. It is called a budget, as the content does not necessarily need to fully use it. The depth budget directly scales up with angular resolution and the size of pixels on the screen, therefore, for a reflective screen, it is also determined by the dimensions of the screen.

In case the parameters of the content surpass the available budget, light field visualization becomes blurry, particularly around the affected side(s) of the depth budget of the display (positive and/or negative). One method to deal with such issue is to realign, relocate the content along the z-axis (depth). This is common for contents recorded with virtual cameras, but it is possible if and only if just one of the sides of the depth budget is affected, and the distance between the other side and the part of the content which is closest to that depth budget is greater than or at least equal to the extent which surpasses the affected side (i.e., there is enough space to move the content in the other direction without surpassing the depth budget). However, particularly for contents captured by real cameras, visualization is scaled down to fit the depth budget, which unfortunately reduces the length of the baseline and the global amount perceived visual information.

The smallest depth budget in practice was credited to the 8.6-inch screen of the display used in the work of Adhikarla *et al.* [109], which was a mere 10 cm, and the greatest one currently available is 1.5 m [110]. When compared to the size of the screen, a remarkable depth budget was 1 m, which was achieved for the screen size of a regular PC monitor (unpublished work of Holografika). The greatest depth budget presented in publications is 12.5 m [111], which has not been implemented yet.

4.2.1.2 Projection

4.2.1.2.1 Field of View

The FOV is an angle that determines the area in which light field visualization takes places. While near-eye 3D technologies approach FOV from the perspective of the observer, FOV in case of light field displays apply to the display itself. One of the greatest advantages of such systems compared to sweet-spot-based displays is that they can utilize the entire FOV to visualize a given content, with a virtually continuous horizontal motion parallax.

In the current literature, FOV refers to the horizontal angle of HPO systems. For future FP displays, FOV can also encompass a vertical component. However, even though FOV appears to be well-defined already, in practice we have two co-existing definitions, that are in fact both correct, yet provide different values. Figure 4.2 demonstrates these two FOV definitions. The one on the left defines FOV to be the angle of the valid viewing area (VVA), while the one on the right measures FOV directly from the screen. The VVA takes the depth budget into consideration; the bigger the depth budget, the farther the VVA is distanced from the screen. If an observer comes closer to the screen than what is allowed by the VVA, the observer

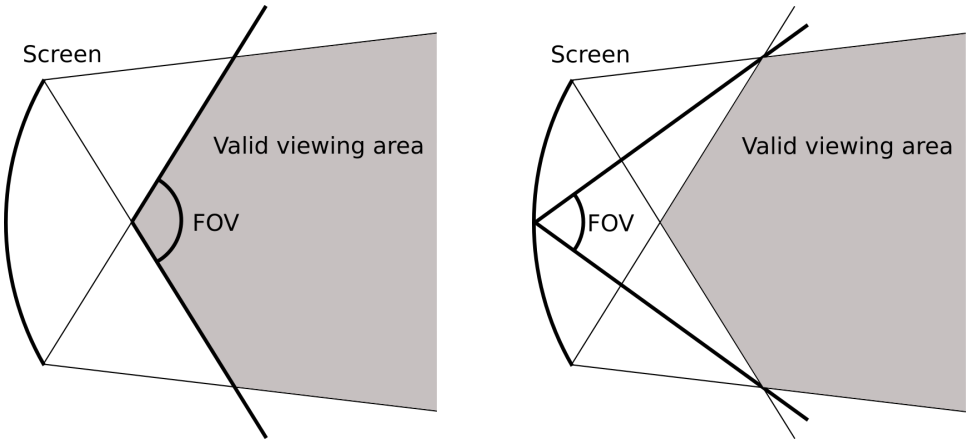


Figure 4.2: LF: Illustrations of FOV definitions.

might perceive an invalid, broken light field due to the missing visual information. Furthermore, front-projection systems create their own restriction for the shortest valid viewing distance, since having an observer occlude with the light rays that are cast onto the screen also results in missing visual information (in the form of a dark shadow). Finally, it must be added here that screen dimensions also affect the VVA, as shown in the figure.

Again, both FOV definitions are correct. However, note that the one measured at the VVA is always at least as big as the one measured at the screen. Therefore, display manufacturers and constructors of prototypes can indeed be encouraged to use the first one so that they can use a bigger number in the system specifications.

In the scientific literature, FOV and VVA are usually treated as synonyms, as equivalent terms. Sometimes VVA is also written as valid FOV, since it is a portion of the area defined by the FOV that enables the valid perception of visualization.

Having a sufficiently large FOV is required for multiple reasons. If we consider only a single observer, the shape and size of the VVA determines those positions (and orientations) from which the content can be viewed, and also sets the boundaries for observer movement. For the case of multiple observers, a bigger FOV also means the accommodation of more simultaneous observers. Lastly but evidently, a greater FOV enables more viewing angles for the given content.

The smallest FOV in research and development currently is the windshield head-up display (HUD) of an experimental vehicular system. As the display was directly designed for the driver and only the driver, the distance of observation and the horizontal deviation of head position was highly constrained. Therefore, an approximately 30-degree FOV was sufficient for the given application. This value is measured at the

VVA vertex, and the corresponding value at the screen itself is only 20 degrees. Ni *et al.* [119] proposed a 360-degree large-scale multi-projection light field display. In this system, 360 optical modules project images onto a cylindrical diffusion screen, the height of which is 1.8 m and its diameter is 3 m. In between these extremes, the FOV of the C80 is 45 degrees, the HV722RC is 70 degrees, and the 80WLT is 180 degrees.

4.2.1.2.2 Overall Resolution

The overall resolution is the number of individual pixels detectable within the FOV. This value can differ significantly from the number of pixels projected from the optical modules due to the optical loss in the display frame. The optical efficiency of the display is characterized by the ratio of the overall resolution and the total projected resolution.

With regards to visualization quality, overall resolution should be as near to the total projected resolution as possible. With every light ray lost due to the difference between them, we actually lose visualization quality, as fewer rays compose the same light field.

In practice, optical efficiency is around 80–90%. This means that roughly 10–20% of the emitted rays do not contribute to light field visualization. Well-optimized systems tend to reach 95–96%, but it does not go beyond that, although even 100% is theoretically possible. The reason why it cannot be done in practice, is that other parameters must be taken into consideration as well. For instance, in case of a wider FOV, due to the skewed projection to achieve it, we lose a certain amount of light rays.

4.2.1.2.3 Brightness

Even though in this subsection, the properties of the projection subsystem are being discussed, brightness is measured at the screen itself, and not at the projectors. The brightness value of a system is measured when a completely white image is projected onto the screen. Even though the entire technology sector refers to the phenomenon at hand as brightness, what we actually deal with is the photometric measure of luminous intensity per unit area.

The first commercial light field display, the HoloVizio 128WLD, had a brightness of 20 $\mathrm{cd/m^2}$. This value is more applicable to cinematic scenarios, as the proper perception of visualization requires more or less dark surroundings (approximately 20 lx). Any value above 1000 $\mathrm{cd/m^2}$ is suitable for general system deployment in most

scenarios. For example, the C80, 640RC and the previously mentioned automotive HUD have a brightness values around 1500 cd/m^2.

If brightness is insufficient compared to the environmental lighting conditions, visualization cannot be adequately perceived. Comparably low brightness directly affects the perceived contrast value of the display, as the environmental light degrades the perception of the darker segments of the visualized content.

With the state-of-the-art projectors — and even most of the regular projectors — available at the time of writing this thesis, brightness typically does not pose an issue when designing light field display systems. For example, if we consider a large 120-inch screen with 100 projectors and a 45-degree FOV, if we use 500-lumen projectors, then visualization is suitable for any given environment and use case.

4.2.1.2.4 Contrast

The contrast of a light field display is directly determined by the contrast of the projection subsystem. Thus, in this context, contrast is the ratio of light rays with the lowest and highest possible intensities.

The lowest contrast in practice was 100:1, which applied to the very first light field displays. Today, most of the displays available are typically between 500:1 and 2000:1.

If the contrast of a light field display is insufficient, details of visualization are lost, as only the differences between bright and dark portions of the content can be properly perceived. Contrast depends a lot on the use case and the content itself. For example, if we consider the light field HUD of a vehicle, the perceived contrast will be low due to the background of visualization, however, there is still a high requirement towards system contrast. As most advertisements apply high-contrast content, using light field for advertising comes with rather low requirements towards system contrast, and focuses more on brightness. On the other hand, in medical applications of light field technology, any loss from the necessary levels of contrast can result in diagnostic inaccuracy. For example, a false negative in tumor detection may originate from the insufficiently low contrast, as the investigated area is not dark enough compared to its surroundings. This is analogous to its 2D counterpart. For example, the accuracy of mammography ultimately depends on the visibility of small, low-contrast objects, as the goal is to distinguish malignant tissue from normal tissue [120].

4.2.1.2.5 Refresh Rate

The refresh rate of light field visualization is basically analogous to 2D technologies. Therefore, the quality requirements towards refresh rate can be considered to be the same. 60 Hz can be achieved by the projection subsystem, however, this is rather a challenge regarding the frame rate of certain use case scenarios and content types. This is detailed later in this section.

4.2.1.2.6 Color space

The color space considerations in the context of visualization quality for the projector array of a light field display are mostly equivalent to what we have for conventional 2D projection. The only aspect here that needs special attention is the color calibration of the projectors. This is typically performed via software.

Without proper calibration — especially in case of a large array with numerous projectors — the projectors do not emit light rays with a perfectly identical color space. This can easily lead to incorrect content colors and color mismatches, and what is even worse from the perspective of the user, certain areas of the projection (commonly vertical areas, but it depends on the construction of the array) can stand out from the rest of the content, degrading general user experience and the natural feel of glasses-free 3D visualization.

4.2.2 Content Parameters

Light field displays may visualize all sorts of content, like static 3D models, light field videos (including real-time transmission), interactive applications (e.g., games), and many more. In this part of the section, content parameters are clustered into common parameters, which apply to any given content, and content-specific parameters.

4.2.2.1 Common Parameters

4.2.2.1.1 Resolution

Content resolution can be approached based on the type of content. If we think of a still image or video content that is captured and stored as a series of 2D images, then the 2D resolution of these images is the spatial resolution, and the ratio of the number of images and the FOV is the angular resolution. HPO content is basically a 1D array of images, and FP content can be imagined as a 2D matrix in this type of representation. In practice, these images are processed by the converter at the input

side of the system, turning a discrete set of images into a continuous light field, by assigning the appropriate light rays to the optical modules.

If the content is directly rendered from a 3D mesh, the definition of content resolution is rather different. Methods of rasterization and ray tracing are used, and even though they differ much in implementation, they both create the final 3D content with the parameters of the display. This means that if we wish to ray trace a given scene, then the output will match the spatial and angular resolution of the display. The primary difference in usage is that rasterization is more suitable for real-time content (e.g., light field gaming) than ray tracing due to the lower computational requirements.

In case of converted content, if the values of resolution are insufficient for the given use case scenario and display, then the degradations are analogous to what applies to the display itself. Therefore, low spatial resolution results in a blur that is not uniform across the screen, and low angular resolution comes with the loss in the smoothness of the parallax effect and with the previously discussed visual phenomena, such as the crosstalk effect.

4.2.2.1.2 Frustum

Frustum in the context of light field visualization is a geometrical portion of 3D space that is defined by the cutting planes, and thus, describes the space in which the content resides. The cutting planes in the front and in the back define the depth of the content, and the ones on the sides are aligned with the properties of the projector array. The primary properties here are the projection aspect and the horizontal projection angle. This implies that the best visualization can be achieved if content generation takes into account the display it is created for.

Figure 4.3 depicts a general representation of a view frustum[8]. In the scientific literature, it is also known as the "pyramid of vision", although it is a truncated pyramid. In this visualization of the concept, the "near" bound of the frustum trivially corresponds to the frontal cutting plane.

If the depth of the frustum is too small (i.e., not deep enough), then the content is inherently limited in the depth of its visualization. However, if it is too deep in contrast to the depth budget of the display, then visualization suffers multiple issues. First of all, the portions of the content that reach beyond the limits of positive and

[8]This figure is licensed by Wikipedia (https://en.wikipedia.org/wiki/File:ViewFrustum.svg) under the Creative Commons Attribution-ShareAlike 3.0 Unported license.
https://creativecommons.org/licenses/by-sa/3.0/

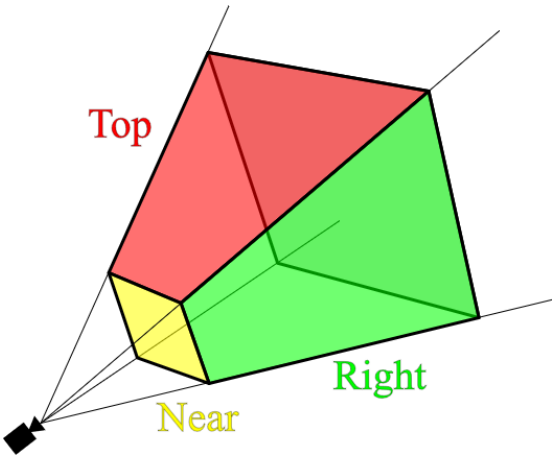


Figure 4.3: LF: A general view frustum.

negative depth budget cannot be properly addressed by the light rays of the projector array. This results in invalid light field data, which affects the rest of visualization as well.

If the frustum is too narrow compared to what projection would demand, then either the sides of visualization become invalid, or the content needs to be stretched and cropped, resulting in the loss of visual data and in skewed proportions. If the frustum is too wide, it is less problematic, since the common solution is simply cropping the data. Having a frustum with inappropriate height parameters in FP visualization is analogous to these issues.

4.2.2.1.3 Scalability

Scalability is a property that characterizes the upscaling and downscaling procedures for various content types. In case of static scenes and videos, where the source content is a series of 2D images, scalability in its concept is similar to what we have for 2D technologies. Basically, the resolution values determine how much a given content can be upscaled for a specific display. For instance, if the angular resolution of the content is much lower than what the display supports, then the conversion procedure may create major inaccuracies in visualization due to a great extent of interpolation. The word "may" is key here, as displaying a high-angular-resolution content on a display that has an even higher — in fact, much higher — angular resolution does not necessarily result in the degradation of subjectively perceived visual quality. However, objective quality metrics may still measure the effect.

Projection-based light field displays always interpolate the source content during conversion, even if the properties of the content perfectly match the capabilities of

the display. Let us now see how this applies to downscaling procedures. For example, if we have two contents, one with an angular resolution of 0.5 degrees and another one with 0.33 degrees, and we wish to convert both of them for a display with an angular resolution of 0.5 degrees, then the second one is expected to have a better perceived quality. This is because both contents get interpolated, but the second one has a higher view density to begin with. Of course, as these values get higher, the perceived differences get smaller. Furthermore, the gain in visual quality will also decrease as the distance between the angular resolution of the content and the display increases. If we take the display with the angular resolution of 0.5 degrees, and we convert contents with 0.3, 0.2, and 0.1 degrees of angular resolution, then the perceived difference between the first and the second converted content will definitely be smaller than between the second and the third. In fact, it is arguable whether there would be any perceived difference between the second and the third at all.

Point clouds are slightly different in the context of scalability. Basically, the loss in visual quality is replaced here with so-called "holes" in the content. It means that if the point cloud content is upscaled more than what it could support, then the distances between the adjacent points become too great, and the continuity of the visualized content is disturbed. For 2D view arrays, scalability is based on the resolution values, and for point clouds, it is approached by the distance between the points. To be more precise, the scalability threshold can be determined by the distance of two adjacent points, that have the greatest distance from all the adjacent point pairs in the model or the scene, since that pair will be the most vulnerable to upscaling (i.e., it will be the first to have a hole in between projected points).

4.2.2.1.4 Color space

The color space of light field visualization is analogous to 2D visualization technologies. The only main difference is within the color space of the projection subsystem — which has been detailed earlier — but the color space of the content has no special consideration with regards to visual quality.

4.2.2.2 Content-specific Parameters

4.2.2.2.1 Frame Rate

Frame rate does not apply to several types of light field content (e.g., static models). It only applies to light field video and interactive applications. Video frame rate is completely analogous to its 2D counterpart. As for interactive applications, it

depends on the computational requirements of the content and on the computational capacities of the system.

Displaying content with high frame rate can be challenging for several reasons. First of all, it requires more data that is to be stored and transmitted. Second, even though the GPUs of the system may have a solidly high frame rate on their outputs, but this does not guarantee a sufficiently high final frame rate. This is because the different GPUs may have different workloads due to the diversity of computational requirements based on the content itself; the content may very well be rather simple from a specific direction, while it may be quite complex from another. Furthermore, these GPU outputs need to be synchronized, and thus, the slowest one may become a bottleneck of the entire visualization process.

4.2.2.2.2 Compression

A light field content does not necessarily need to be compressed. Compression is, of course, desirable, as light field content usually means a tremendous amount of data. If we wish to store and transmit such data — especially in real time — then the smaller the data (file size or data rate), the better. However, there are many applications of this technology that do not involve the processes of storage and transmission. These are typically the interactive applications, where the visualized data is a collection of rendered 3D meshes.

If we do compress — for instance for static models and videos — then compression can happen in two fundamental ways. One is using conventional compression for 1D and 2D view sets, where views are separately compressed. Real light field compression relies on the redundancy between these views [121]. Although the thesis primarily focuses on wide-baseline light field displays (with an observer line measurable in meters) and the associated content, it is necessary to review the compression methods used for narrow-baseline lenslet-based light fields (with an observer line in the order of centimeters), as both techniques can exploit inter- and intra-view redundancy for light field compression.

An example of the compression of light field acquired by multiple cameras is the work by Tamboli *et al.* [122], where camera images were separately compressed using JPEG, JPEG2000, and WebP compression methods. Similarly, the array of lenslet images captured using plenoptic cameras — considered as a single image — was compressed using intra-coding methods [121, 123].

Light field compression methods that rely on inter- as well as intra-view redundancy have been shown to be better in terms of various objective quality metrics,

especially for high compression ratios [121]. For multi-camera sequences, the multi-view extension of High Efficiency Video Coding (HEVC) has been used to exploit spatial, angular (inter-view) and temporal redundancy [124, 125]. Similarly, multi-view compression methods targeted towards lenslet images arrange the content as pseudo-temporal sequences [126]. The compression of light field using point cloud codecs has also been proposed in the literature. For example, the work of Zhang *et al.* [127] maps the multi-view images to a point cloud and jointly compresses the geometry and the view-dependent colors.

Again, it needs to be noted that the size of the light field data after conversion is independent of the compression of the source content. On the one hand, compression may negatively affect the quality of visualization, but on the other hand, it may boost performance for applications relying on data transmission.

4.2.2.2.3 Render Type

The visualized light field content can either be a converted image set, or if we have a 3D mesh, then we see a rasterized or ray-traced content. Light field images and videos that are captured by a real or virtual camera array are typically the first case. Static scenes and videos can also be rendered by using methods of ray tracing for virtual content generation. Interactive contents — such as games and applications that require user input — are normally represented with 3D meshes. These meshes are either rasterized or ray traced before their visualization on the screen of the light field display. Rasterization is computationally less expensive and faster, therefore, it is the most common visualization method for such contents. Ray tracing is also possible, which is demonstrated by the work of Doronin *et al.* [128].

The methods of rasterization and ray tracing are analogous to conventional 2D visualization; ray tracing enables more life-like visuals compared to rasterization, as it traces light rays through the given 3D scene. However, these operations are not only content-specific, but they are display-specific as well, without any intermediate process. The output that they provide is directly matching the parameters of the display. In contrast to these, conversion creates intermediate visual data, which is then interpolated for the projector array.

It needs to be added that volume rendering is an increasingly emerging form of light field visualization, particularly in the field of medical imaging [129]. The primary benefactors of such rendering are CT, MRI and PET scans, as they are fully 3D by definition, and having them visualized in a proportionate volumetric 3D manner —

without the need of viewing devices — instead of the classic 2D cross-section analysis supports the work of radiologists.

4.2.3 Discussion on Current and Future Research Efforts

This section discusses quality indicators and features of light field visualization that are not fully implemented yet and are to be integrated into such systems in the future.

4.2.3.1 Super Resolution

Super resolution in the general context of light field technology is often understood as a reconstruction method of increasing the spatial resolution at the capture side [130]. It has a great potential in enhancing the quality of the reconstructed content that is to be displayed. However, in the context of light field displays, super resolution carries a rather different meaning.

When this technical term is applied for display systems, it refers to extremely high resolution capabilities. By resolution, mostly the angular component is meant, but spatial resolution is a vital part of it as well. The core concept of super resolution is that the achieved resolution of the display — and of course, the visualized content — is so high that the human eye can focus on different portions of the content. At the time of writing this thesis, it is true for every single projection-based light field display system that no matter how advanced it is, the human eye always focuses on the plane of the screen. This attribute is evidently desirable for future light field displays, as it makes visualization feel more realistic — or at least more spatially present — in general.

Let us imagine a light field display with a screen height of 70 cm that is viewed from a distance of 2 m, and let us also assume a pupil diameter of 5 mm. In such a case, an angular resolution of at least 0.14 degrees is necessary to achieve super resolution. Generally, if two distinct rays approach the eye from a given point of the screen, then both of these rays will enter the pupil, and therefore, the eye may focus on different depth levels of the content. This definition of super resolution is also known as "high-density directional display technique" in the work of Takaki [131].

As for spatial resolution, there is no specific calculation or estimation on how great it should be. The general rule here is that light field visualization should not suffer clearly perceptible degradations. It can be assumed that a minor level of blur due to insufficient display and/or content spatial resolution may be tolerated, and thus, shall not affect the perception of super resolution. Again, if the angular resolution

satisfies the previously described optical conditions, then it is up to spatial resolution whether super resolution will be achieved or not. It is important to note that an excessively high angular resolution cannot compensate insufficient spatial resolution and vice versa.

Let us now address a practical application where super resolution may significantly benefit user experience. The work of Cserkaszky *et al.* [132] introduces a novel light field telepresence system. The greatest contribution of such application of light field technology is towards the "sense of presence". If the human eye can focus on different depth levels of the communication partner via super resolution, then this can boost the sense of presence, as the scenario becomes more natural and lifelike.

Although a telepresence system may indeed benefit from super resolution, it needs to be noted that such application by definition does not have a visual depth where super resolution can truly shine. A better example could be an automotive HUD, which practically necessitates a greater level of depth. If we consider a spatial navigation application where the visualization is basically a combination of real-life visuals and projected components, then super resolution may blend visualization seamlessly — or at least less artificially — together. The importance of super resolution in this use case scenario is also reinforced by the safety concern of having a driver focus separately on real-life depth levels and on a fixed-focal-distance visualization.

Basically, any use case of light field visualization may be enhanced via super resolution, where depth values play a significant role in the overall user experience, and thus, the relevant portion of the visualized content is sufficiently far from the plane of the screen. This can both apply to industrial, medical, and entertainment purposes alike.

4.2.3.2 HDR

The technical term HDR has two widely spread interpretations. First, it may refer to HDR-color imaging. In this sense — as opposed to the commonly used 8-bit, or even 5-bit per-channel coding for LDR imaging — an HDR image can currently be coded using 10, 12, or 16 bits per each channel. Different image formats (e.g., DDS or EXR) allow an image to be encoded using float-precision values, i.e., 16 or 32 bits per channel. The latter is used mostly for image processing and professional image editing applications, but rarely by regular end-user programs.

Regarding light field, the implementation of HDR-color imaging is completely dependent on the configuration of the display system at hand. For example, the widely-used configuration of multi-view displays — based on the ultra-res LCD screen

with a microlens array on top — does not seem to have any additional issue with adapting to HDR-color imaging when compared to its 2D counterpart.

The term HDR may also refer to HDR-luminance imaging. In this context, it is usually assumed that LDR images can utilize only $[0, 1]$ range for luminance, while the HDR images may store potentially unlimited, i.e., $[0, +\infty)$ range. Even for the case of 2D visualization, there are several active research topics on HDR-luminance imaging. Conditionally, they can be subdivided into two categories. The first category is the capturing of HDR content with LDR cameras, and the second category is the visualization of HDR content on LDR devices. The research in the latter category is mainly focused on how to make a conversion from $[0, +\infty)$ range into the $[0, 1]$ range, in a way that would be plausible for the HVS. This conversion operation is commonly known as tone mapping.

In a recent study, Eilertsen *et al.* [133] summarized the most prominent modern approaches for tone mapping 2D video sequences, which are also applicable for still-image processing. Since 3D display technologies right now are under active development, there is no similar state-of-the-art dissemination of knowledge for tone mapping on light field displays. The only currently available papers on related topics are about tone mapping either for VR [134], panoramic images [135], stereo images [136], or multi-view displays [137].

For future research on tone mapping for 3D displays, we can distinguish three main directions. The first one is mimicking the existing 2D methods. For example, it seems straightforward to take the Reinhard's approach [138] of global tone mapping and apply it for a 3D display. The aforementioned approach requires global luminance estimation, which can be found per each particular viewing position separately. For multi-view displays, this problem seems trivial; for real light field displays, such as the projection-based HoloVizio-like systems, one may refer to Doronin *et al.* [139]. Second, one can define the ground truth tone mapped 2D images for the series of observer positions, and then make an effort to approximate them by altering the 3D image in a display-specific format. Such an approach would likely involve the constitution and solution of an optimization problem, which will depend both on the nature of the ground truth, and on a particular display parameterization. Third, it is possible to make volumetric tone adaptation. In this approach, for each particular point in physical 3D space, one could make a tone adaptation for its local 3D neighborhood. This approach seems valid for volumetric displays and for Lambertian scenes, for which we can assume that any point in space emits light in all directions equally. For different types of light field displays and for non-Lambertian scenes,

instead of 3D neighborhood, one would need to consider the 5D neighborhood (space position and ray direction) of each point in space, which can be both ambiguous and computationally expensive.

4.2.3.3 High Frame Rate

High frame rate (HFR) visualization is a common label for any display technology above 60 Hz. In case of the conventional 2D displays, commercial HFR screens are typically 144 Hz or 240 Hz — mostly for gaming purposes. Such displays can reach 400–600 Hz, but nearly 1000 Hz is possible as well. Commercial projectors for stereoscopic 3D visualization support 240 Hz, which means that they provide 120 Hz per eye.

HFR visualization is an absent research topic in the area of light field at the time of this thesis; in fact, no result or effort has been published so far towards HFR light field. If we consider the three future features discussed in this section, it can be stated that while the other two have similar levels of potential contributions regarding visualization quality, HFR light field has a lower scientific priority in comparison. This is due to the limitations in the use case scenarios where HFR systems could truly benefit the users. Furthermore, the aforementioned bottleneck issue would still apply, reducing the achieved frame rate of visualization to the output of the slowest GPU.

Probably the most notable contribution of HFR light field systems would occur in case of hybrid visualization, where elements of the real world are combined with light field visuals; the smaller the perceived difference, the better. The previously mentioned example of automotive HUD applies here as well. The sense of presence could also profit from such feature, making telepresence application more natural in appearance. Any utilization of HFR light field with critical user reaction time may be supported by this indicator of quality, however, it does not provide a universal benefit to visualization like the other two do.

4.2.4 Research Direction of the Thesis

From all the indicators discussed in this section, many can be bestowed with research questions worthy of scientific investigation. In fact, all of these KPIs may benefit from experiments on perceived quality. In this thesis, the primary research focus is on resolution, extended with some related parameters that affect user experience, such as the FOV. Before my work is introduced in detail, let us first review the scientific contributions in the area of light field QoE.

4.3 Related Research on Light Field QoE

The published works of Kovács *et al.* [35,115,116,140–142] address light field visualization from both the angle of perceived quality and the measured objective capabilities of systems. Spatial and angular resolution are particularly highlighted in these works.

Due to the apparent sheer importance, spatial and angular resolution enhancement efforts are spreading in the field. As a recent example, the work authored by Gul *et al.* [143] introduces a method for this purpose, which was trained through supervised learning. The results are promising, as the proposed method may provide significant improvements compared to certain conventional interpolation methods.

Tamboli *et al.* [34, 144–147] investigated the perceived quality of light field view synthesis, created a high-angular-resolution dataset for objective and subjective assessments, evaluated content features, and developed an objective light field quality metric with an angular component.

Perra [148] also proposed an objective metric in his study on decompressed light fields. Other works of Perra *et al.* [149–151] address the QoE of light field subsampling, investigate the reconstruction of point clouds based on light fields and study the use of non-overlapping tiles for generating pseudo-temporal sequences of light field images in an attempt to efficiently encode the data.

The perceived quality of light field visualization evidently depends on data compression as well. Coding will play a significant role in the delivery of future light field multimedia, which will need to balance between the extent of data reduction and the possible changes in visual quality. Results of scientific effort can already be observed in the works of Viola *et al.* [28, 152] and Paudyal *et al.* [153]. Other works of Viola *et al.* [154] address subjective test methodology, which was also investigated by Darukumalli *et al.* [29]. The works of Paudyal *et al.* [33, 155–159] demonstrate the importance of light field content and display system selection for subjective tests on perceived quality, consider watermarking, introduce a reduced reference quality assessment method for light field images and present a light field dataset captured by a Lytro camera. A database particularly created for QoE studies on the perceived quality of light field visualization was also presented by Murgia *et al.* [160].

Shi *et al.* [24] proposed a database as well and carried out subjective and objective evaluations on their static contents. The experiment used a stereoscopic 3D TV for the subjective tests, and the test participants had to interact with the visualized content by changing perspective with the help of a computer mouse (by clicking and dragging). Light field databases were also presented by Rerabek *et al.* [161] and Wang

et al. [162], and the so-called "classic" datasets — such as the Stanford Archive[9] — were reviewed by Wanner *et al.* [163].

The work of Wang *et al.* [164] investigates the QoE of multi-layer light field displays (MLLFD). For such displays, special considerations regarding the perceived quality are required, as quality factors (i.e., perceived resolution) may differ based on the implementation. Other important components of user experience are also measured, such as naturalness.

Agus *et al.* [129] investigated the visualization of 3D medical (radiology) data on a light field display. The authors state that their subjective tests — involving physicians and radiologists — have proven that such visualization method is "clearly superior" to conventional 2D displays. The work of Cserkaszky *et al.* [165] also highlights the potential for nuclear medicine, and points out research synergies. Furthermore, the paper of Lévêque *et al.* [166] considers light field in their analysis on the perceived quality of medical contents.

The recent work of Wijnants *et al.* [167] proposed HTTP adaptive streaming to transmit the data of static light fields. The core idea of the concept is approaching the source views of the model or the scene as the segments of a video sequence.

Interactive features were addressed by the contribution of Adhikarla *et al.* [109], describing a research in which free-hand gestures were tracked to carry out tasks (touching highlighted red sections) on a projected surface. Although perceived quality was not the primary research focus, the subjective test carried out is quite noteworthy, as they used the User Experience Questionnaire (UEQ) [168] to measure attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty. The work of Marton *et al.* [31] used multiple methods to navigate through large-scale static models visualized on a light field display. For instance, hand gestures were tracked using depth sensors, enabling actions such as zooming or panning. The subjective assessment in the study also used a 3D mouse as the controller, with dedicated buttons for zooming in and out. Similarly to the previous work, several quality aspects were investigated, including ease of learning, ease of reaching desired positions, and the perceived 3D image quality. The results indicate that the 3D mouse was preferred by the test participants over hand gestures.

The previously introduced researches on perceived quality mainly focused on the quality of static content visualization. There are works on light field video as well, such as the video service feasibility research of Dricot *et al.* [124], the live capture system

[9]The (New) Stanford Light Field Archive:
http://lightfield.stanford.edu/lfs.html

of Adhikarla *et al.* [169], the video frame quality analysis of Tamboli *et al.* [170], the proposed 3D telemedicine system of Xiang *et al.* [171] or the real-time telepresence system of Cserkaszky *et al.* [132].

The work presented in this thesis addresses both static content and light field video. Static models were used in the research efforts of Phases 1 and 2, and videos are investigated in Phases 3 and 4. Before the introduction of these contributions, let us first review the apparatus and the research environment.

4.4 Displays and Research Environment

In this section, the two light field displays that were used for the subjective studies are briefly introduced, along with the research environment where the tests took place.

4.4.1 HoloVizio 80WLT

The 80WLT is a television-like back-projection light field display. It has a 30-inch screen on which the content can be viewed in a full 180-degree FOV. It has an angular resolution of 1.5 degrees, and the screen brightness is $300\,\mathrm{cd/m^2}$.

Because of its great FOV, this was selected as the light field display of the FOV-related research. However, due to its small screen and the relatively small angular resolution, the C80 was used for all the other subjective tests.

4.4.2 HoloVizio C80 Light Field Cinema

The HoloVizio C80 light field cinema is a high-end front-projection system, possessing some of the greatest properties among all commercially available light field displays at the time of conducting the experiments and also at the time of writing this thesis. First of all, it has a 140-inch screen, which is 3 meters wide and 1.8 meters tall. More importantly, it has an angular resolution of 0.5 degrees, which makes it perfectly suitable for the different subjective tests with variations in content angular resolution. In fact, both these two values are currently unmatched by other systems. The FOV is 45 degrees, and the screen brightness is $1500\,\mathrm{cd/m^2}$.

As the C80 is a front-projection light field display, the location of the optical engine array determines the default viewing distance at 4.6 meters from the plane of the screen, which is the equivalent of a 2.5 H distance. Although the vertex of the FOV is closer to the screen than this distance, but it is common practice that observers should not be positioned between the projectors and the screen due to the possible light ray occlusion.

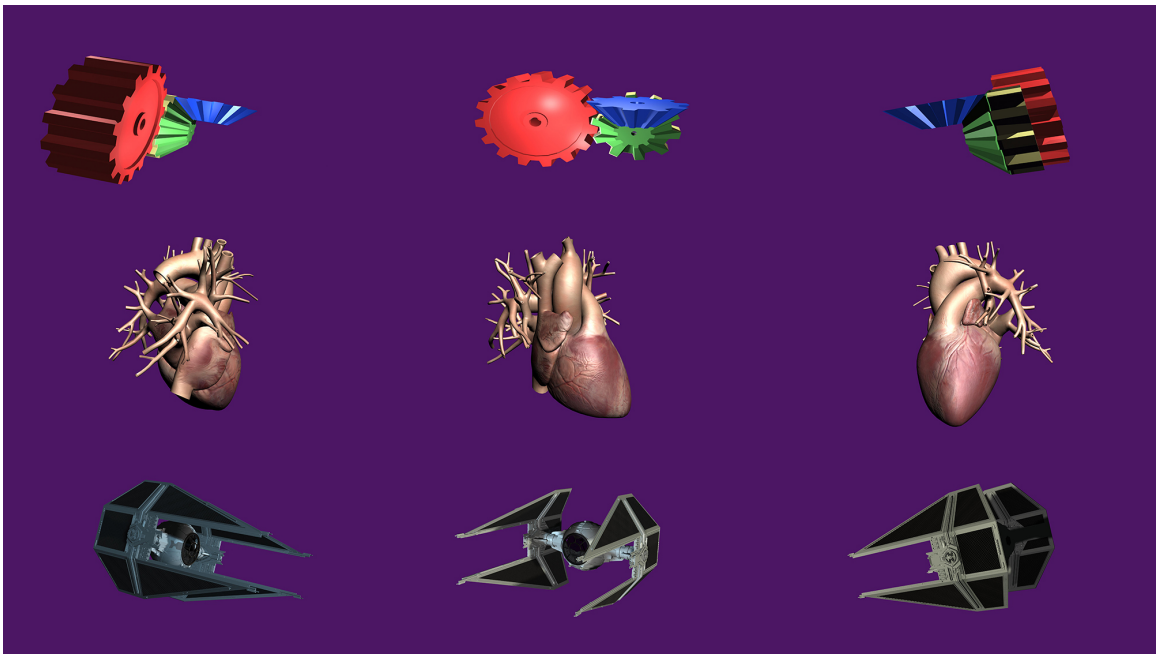


Figure 4.4: LF: Models used in Phase 1.

4.4.3 Research Environment

The tests were carried out in a laboratory, isolated from external audiovisual distractions. The lighting condition of the environment was 20–25 lx for all experiments, except during the Phase 2 research on static observers (section 4.5.2), where approximately 8–10 lx was set for a more cinematic experience.

4.5 Phase 1: Initial Research

Each experiment in Phase 1 was completed by 20 individuals (11 males and 9 females). The age range was from 19 to 50, and the average age was 26. As these were the first experiments from the series of tests on light field visualization, fundamental research efforts on perceived quality were carried out, using static models.

4.5.1 Models

In Phase 1, three models have been used, provided by Holografika. All of them had the exact same, plain-colored background. Source stimulus A was a collection of three shapes (gears) with RGB colors (one of each) and hard model edges, B was a human heart with notable textures and a smooth structure, and C modeled a tie fighter[10]

[10] © Lucasfilm Ltd. LLC, The Walt Disney Company

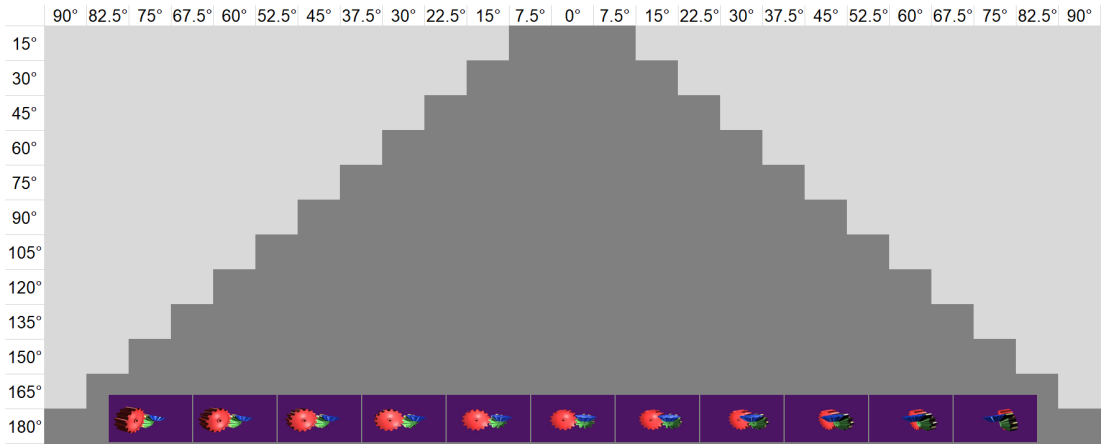


Figure 4.5: LF: Test cases (vertical axis) and the corresponding left and right viewing angles (horizontal axis) of the subjective test on Field of View. Viewing angles covered by the dark area indicate observable content. Views of stimulus A are shown as example.

with a detailed structure. Figure 4.4 shows these models from different angles, all of which were observable by the test participants on the 80WLT.

4.5.2 Research on Field of View

4.5.2.1 Research Aim

The aim of the research was to assess the perceived FOV size, and to address the related issues on the willingness to pay and to use [172].

4.5.2.2 Test Conditions

This was the one and only experiment that used the HoloVizio 80WLT, as its FOV can be utilized up to 180 degrees. In order to accurately simulate different FOV values, the stimuli were first rendered for the entire FOV, and then views on the sides were replaced with the background. This means that in the simulated FOV, the test participant could properly observe the model; outside the FOV, one only saw the background color; and on the edge of the FOV, the quickly fading model was seen.

A total of 12 test cases were defined, each with its own FOV (as it was the only varying test condition), ranging from 15 to 180 degrees. Every FOV increment of 15 degrees was tested, so the chosen test cases were 15, 30, 45 ... 180 degrees. Figure 4.5 demonstrates the FOV sizes of the test cases. For example, when the test condition of 30 degrees was visualized for a given source stimulus, it was possible for the human

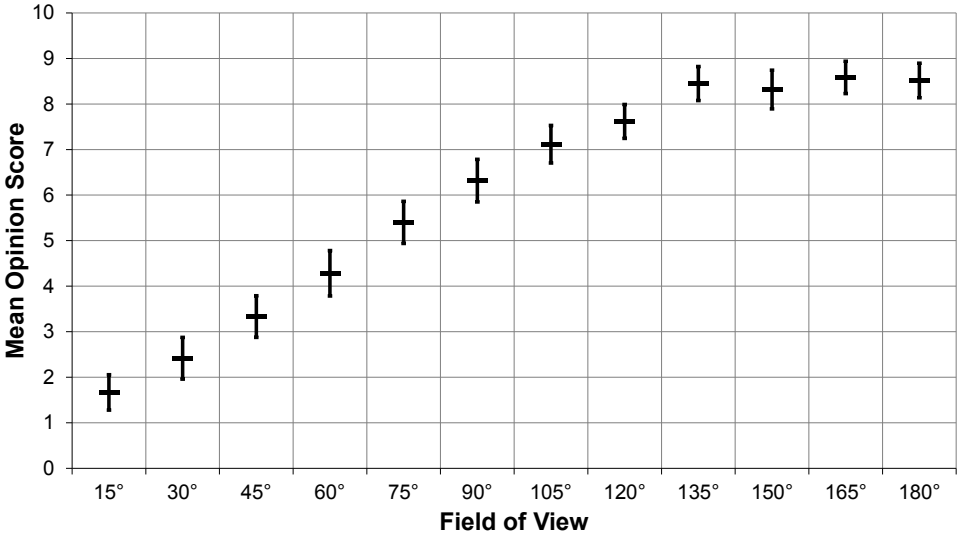


Figure 4.6: LF: Mean Opinion Scores of the subjective test on Field of View.

observer to move roughly 15 degrees left or right inside the valid FOV with respect to the center view of the screen without experiencing model fading.

The 36 test stimuli were shown in a randomized order, and they were to be subjectively evaluated using three scales. One was on general user experience, ranging from 1 to 10, and the others were binary scales, regarding the willingness to pay and to use.

The test participants observed the FOV in a semicircle in front of the display. They could go as close to the display as they desired, and the farthest they could go was 5 meters. Although movement was arbitrary, but the test participants were instructed to view the given stimulus from different distances before assessing.

4.5.2.3 Results

The results for the 10-point scale are shown on Figure 4.6, and the percentages derived from the binary ratings are shown on Figure 4.7. The mean scores indicate a linear relationship between user experience and FOV up until 135 degrees. Between 135 and 180 degrees, the obtained means do not differ significantly, and they fit into a small interval of 0.27 on a scale from 1 to 10.

Similar tendencies can be observed for the binary scores, but they already close to peak at 105 degrees. As this is also the 50% threshold for WTP, these results recommend that future TV-like light field displays provide at least 105 degrees of FOV. Furthermore, the slight drop in both binary scores beyond 135 degrees indicates

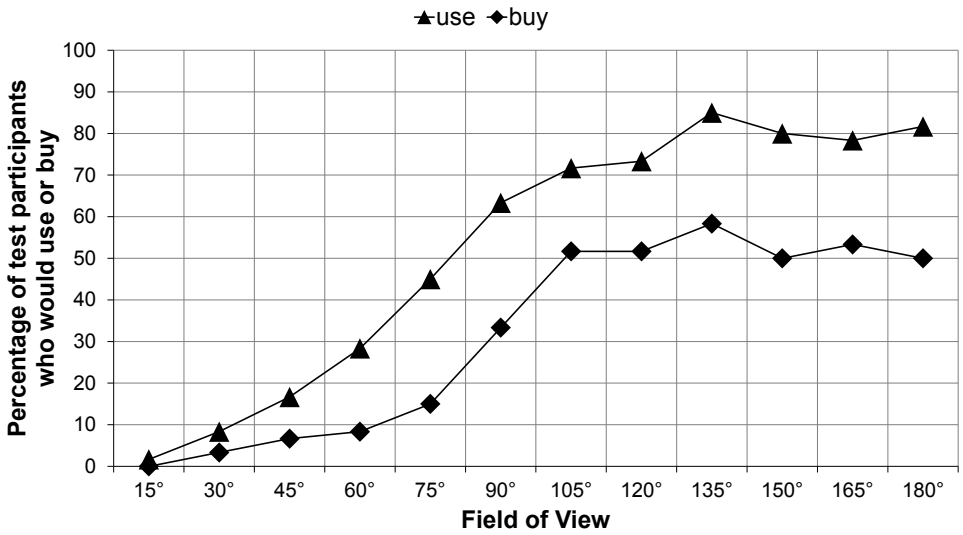


Figure 4.7: LF: Willingness to use and to buy results of the subjective test on Field of View.

that test participants were actually able to perceive the larger FOV, but due to the lack of perceptual added value, these ratings were slightly penalized.

4.5.3 Research on Spatial Resolution

4.5.3.1 Research Aim

The aim of the research was to assess the perceived spatial resolution of the content [173].

4.5.3.2 Test Conditions

The source stimuli were directly rendered in 5 spatial resolutions: 854×480 (WVGA), 1024×576 (PAL), 1280×720 (720p), 1920×1080 (1080p) and 3840×2160 (2160p). The resulting 15 test stimuli were subject to a full DCR comparison, which included self-comparisons as well (i.e., identical stimulus pairs). This is a helpful method to combat cognitive bias that may occur, as comparison results are not only processed on their own, but they are also compared to the corresponding self-comparisons. This method was involved in the final experimental setup due to the small perceivable differences between the test stimuli.

As angular resolution was not investigated, the stimuli were rendered using 180 virtual cameras, which corresponded to 4 views per degree. Regarding viewing conditions, the observation of each test stimulus began from the 2.5 H standard distance,

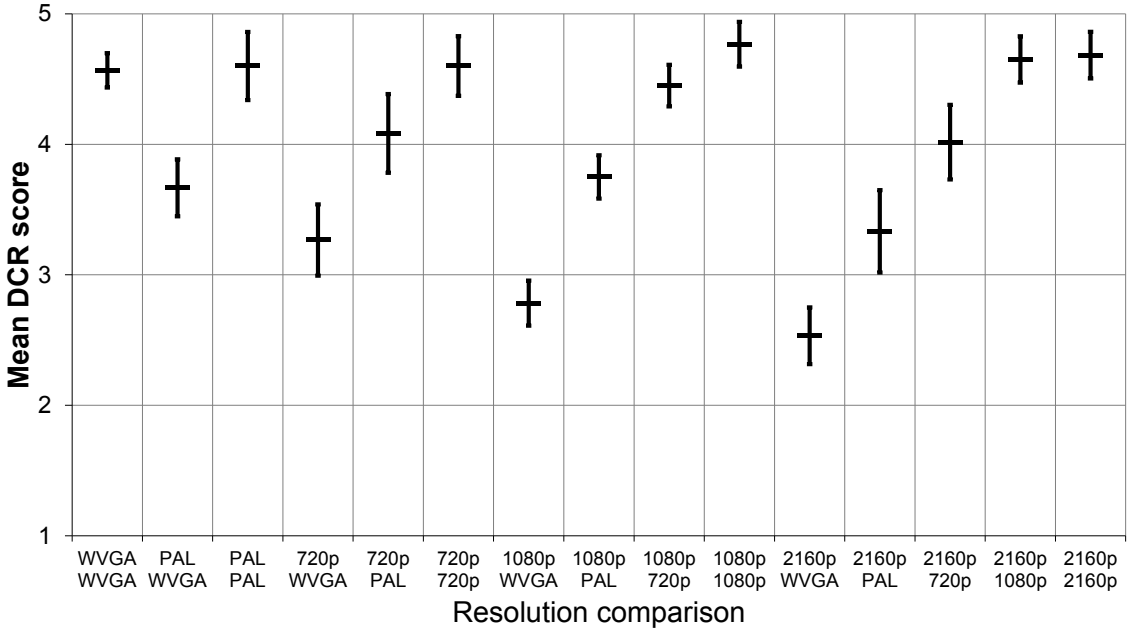


Figure 4.8: LF: Mean DCR scores of the subjective test on spatial resolution.

and it could be increased by two meters at most, along with any horizontal change within the valid FOV. This additional freedom in observer movement was necessitated by the minor differences between adjacent pairs in the test matrix.

4.5.3.3 Results

The mean DCR scores and the scoring distribution are shown on Figures 4.8 and 4.9, respectively. First of all, the obtained mean scores indicate a measurable level of cognitive distortion, as self-comparisons were far from being ideal. This was due to the small perceptual differences. Generally speaking, it can be stated that apart from the two greatest differences between resolutions (2160p/WVGA and 1080p/WVGA), the degradations were rather tolerable. In fact, the visualization of even those two cases were to closer to being *slightly annoying* than *annoying*. Furthermore, the highest neighboring resolutions were statistically indistinguishable, especially in the case of 2160p/1080p. Finally, the high number of ratings indicating imperceptible visual differences between various resolutions should also be noted (see Figure 4.9).

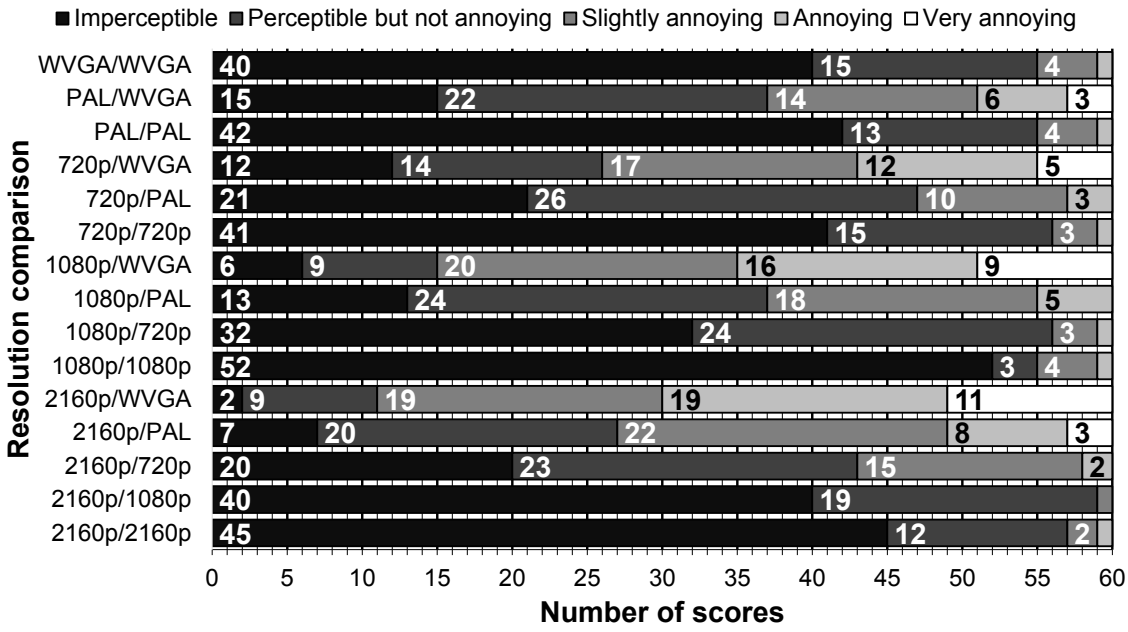


Figure 4.9: LF: Scoring distribution of the subjective test on spatial resolution.

4.5.4 Research on Angular Resolution

4.5.4.1 Research Aim

The aim of the research was to assess the perceived angular resolution of the content, and to address experimental validity [174].

4.5.4.2 Test Conditions

The source stimuli were directly rendered in 10 different angular resolutions, from 15 to 150 views, incremented by 15 views each step: 15, 30, 45 ... 150. The spatial resolution for all 30 test stimuli was a fixed 1024×576. The evaluation of test cases was carried out on an ACR scale from 1 to 10, where 1 was the lowest possible score and 10 represented the reference quality. The test participants observed the stimuli from a fixed 2.5 H distance, and had a sideways movement of a meter in each direction.

In this experiment, the training phase prior to the subjective tests did not include an extensive additional training regarding the phenomenon of parallax disturbance, which is a visual phenomenon that is commonly not experienced by individuals; it is not present in real life and visualization technologies on the current consumer market do not exhibit it. The aim of this decision in the experimental configuration was to address measurement validity through scoring consistency. For all the other experiments on the perceived quality of light field visualization, the extensive additional

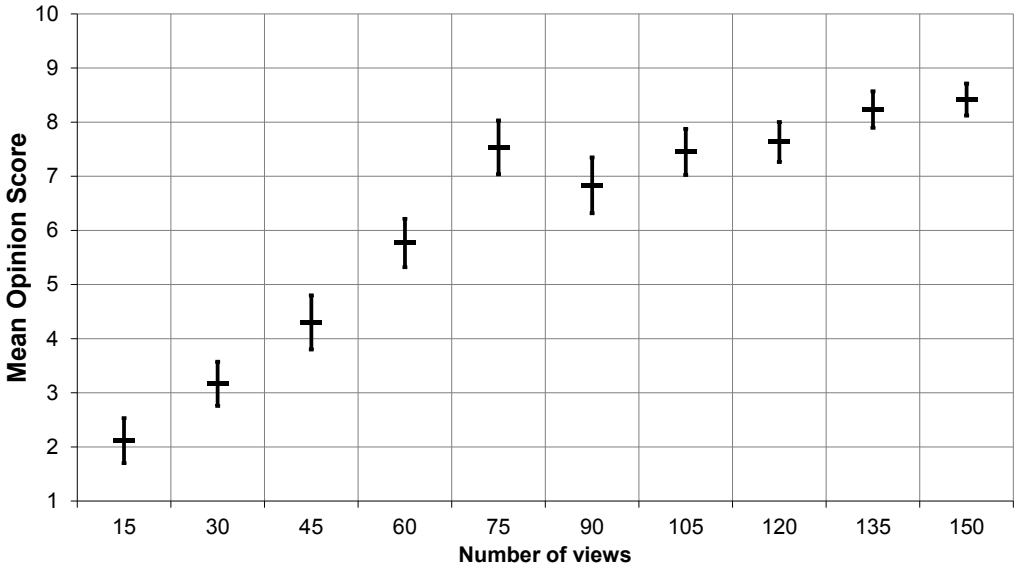


Figure 4.10: LF: Mean Opinion Scores of the subjective test on angular resolution.

training regarding parallax disturbance was involved in order to support assessment accuracy.

4.5.4.3 Results

The mean scores of the experiment are shown on Figure 4.10. Throughout the test, participants made many inconsistent ratings when assessing the quality, but what stands out the most from the obtained results is the particularly high value assigned to 75 views. When the angular resolution of 90 views over the 45-degree FOV (2 views per degree) and its neighboring test cases are investigated (see Figure 4.11), the cause of these results become clearer. There were four instances when at least half of the entire scale was used to provide higher scores to 75 views compared to 90 views, and in more than half of all evaluations, the case of 75 views was rated to be better than 90 views. In fact, in multiple cases, the ratings of 75 views even surpassed 105 views. The figure also highlights the cases of incorrect scoring relations between 90 and 105 views.

The case of 90 views is of scientific interest, since it is the theoretical limit for light field visualization on the given display. However, as the results show, higher angular resolutions managed to achieve higher scores, but of course, with a less steep elevation (see Figure 4.10).

Although for these particular angular resolutions were the mean scores the most effected, great extents of inconsistencies were found throughout the entire set of test

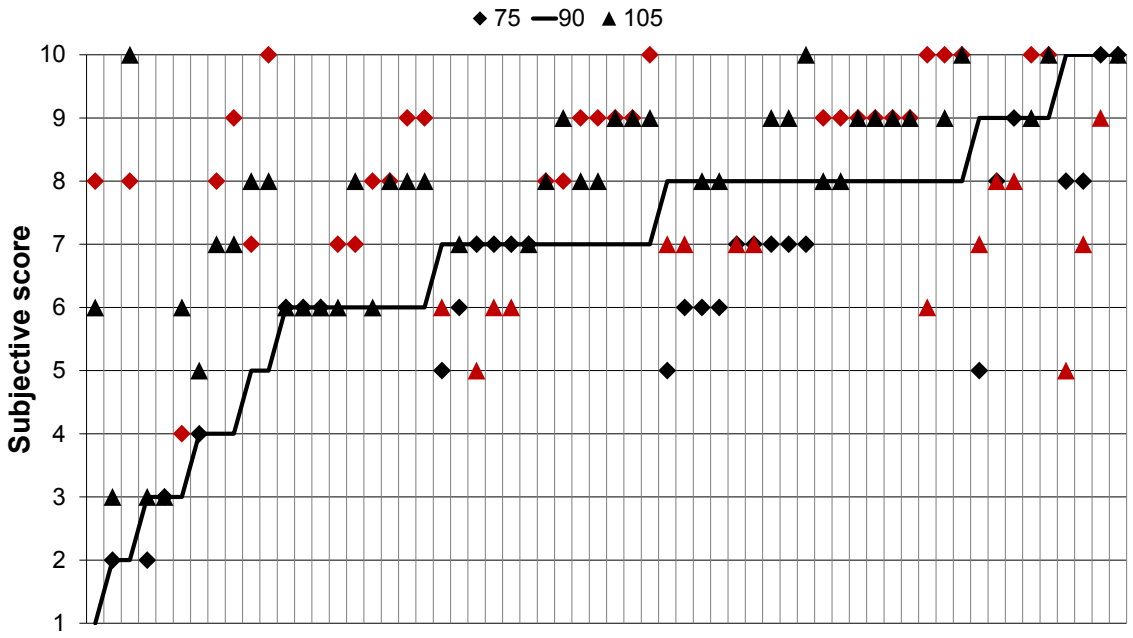


Figure 4.11: LF: Subjective scores for 75, 90, and 105 views. Red markers indicate incorrect relations with respect to 90 views.

conditions. In fact, certain test participants provided higher scores to 45 or 60 views than to 105 or even 135 views. Furthermore, the results suggest that the threshold of tolerance should be around 45–60 views, which points towards an on-going scientific debate in both industry and academia: Can the angular resolution of 1 view per degree be sufficient for future use cases?

4.5.5 Research on View Synthesis

4.5.5.1 Research Aim

The aim of the research was to assess how view synthesis affects the perception of the content compared to reductions in angular resolution [175].

4.5.5.2 Test Conditions

The test stimuli for this experiment were created with two types of settings: either with reduced angular resolution or with light field reconstruction. The selected angular resolutions were 30, 60, and 90 views for the given 45-degree FOV of the display. As for light field reconstruction, Shearlet transform [176] was applied to the source models in 3 different ways, resulting the image set for the measurement. First, it was decimated by a factor of 2 (*D2*), then by 3 (*D3*), which means that every second and

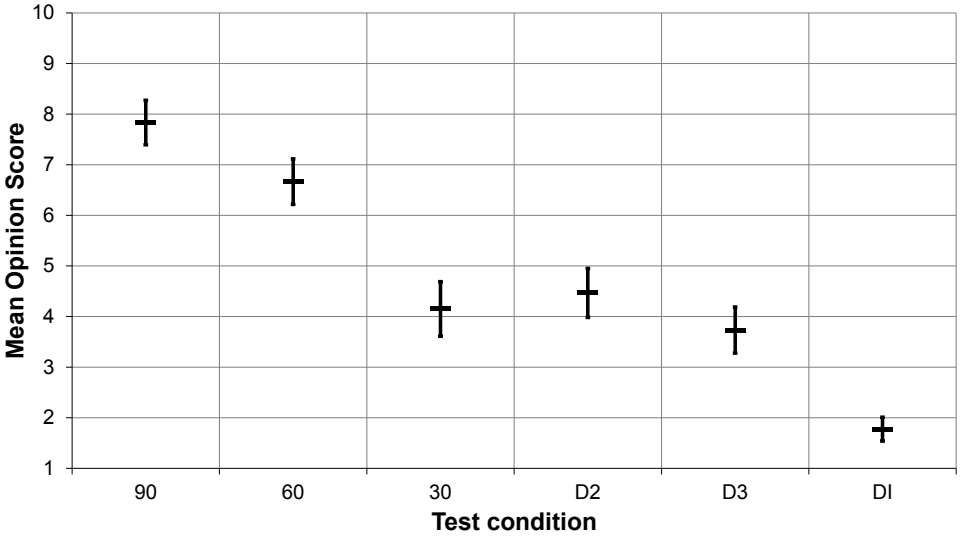


Figure 4.12: LF: Mean Opinion Scores of the subjective test on view synthesis.

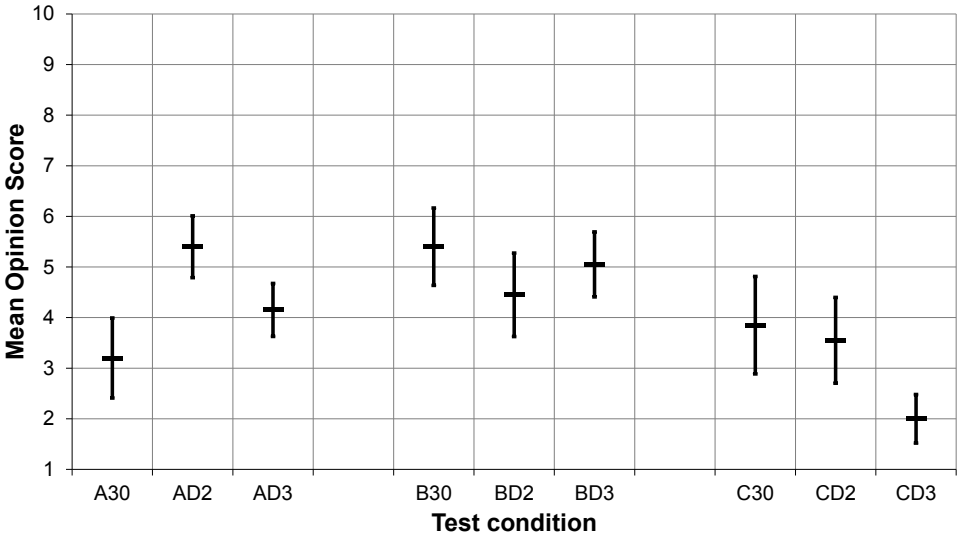


Figure 4.13: LF: Mean Opinion Scores per stimulus for conditions *30*, *D2*, and *D3* in the subjective test on view synthesis.

third row was preserved while the others were discarded and then resynthesized. The Shearlet transform was also used for light field reconstruction with maximum disparity between two adjacent images adjusted too high for the algorithm (DI), making the output of the process highly degraded and blurry. Originally, the solution provided 1024 views per reconstructed stimuli, from which every fourth was kept, so during the measurement all reconstructed images consisted of 256 views. All test stimuli had a fixed spatial resolution of 1024×576.

Similarly to the previous experiment, a 10-point ACR scale was used to assess the

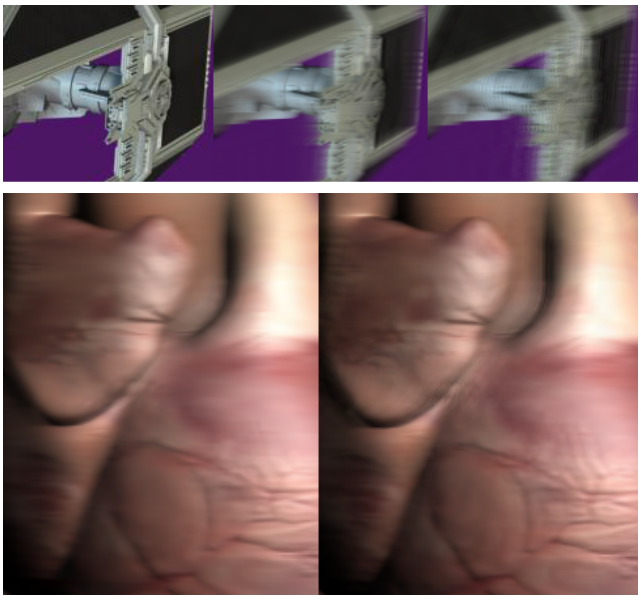


Figure 4.14: LF: The effect of view synthesis on source stimulus *C* (top) and *B* (bottom). From left to right, for stimulus *C*, conditions *30*, *D2* and *D3* are shown, and for stimulus *B*, conditions *D2* and *D3* are shown.

stimuli. Viewing conditions were also identical to the test on angular resolution.

4.5.5.3 Results

The mean subjective scores of the experiment are presented on Figure 4.12. Test conditions *90* and *60* received significantly higher scores compared to the previous experiment, due to the underwhelming experience of the other test conditions. While it is clear that *DI* was simply unacceptable for the test participants — as the contents were barely recognizable due to the excessive blur — the case of the remaining three test conditions is not that evident. From all the studies of Phase 1, this particular one had the most notable, peculiar content dependencies.

Figure 4.13 shows the results for these test conditions separately for each stimulus. Due to the sharp edges of stimulus *A*, condition *30* was highly penalized compared to the other two. With the fine details of stimulus *C*, conditions *30* and *D2* were equally bad, but *D3* was much worse because of the higher amount of blur. A part of stimulus is demonstrated on Figure 4.14. The figure also shows a region of stimulus *B*, which was essential in the understanding of how the objectively worse *D3* received better scores than *D2*. As explained by multiple test participants after the subjective tests, *D3* was very similar to *D2*, however, it had a better, more pleasing perceived contrast. The difference is not statistically significant, but the individual scores report

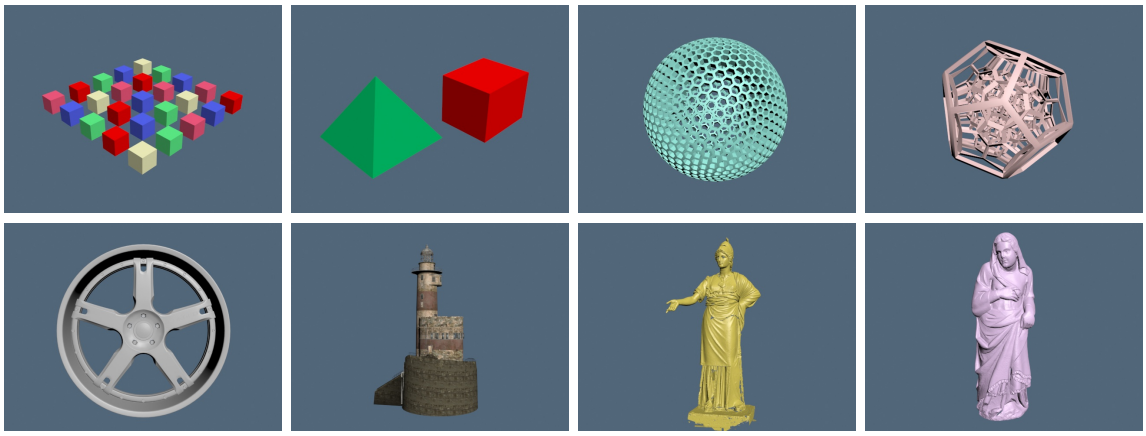


Figure 4.15: LF: Models used in Phase 2.

that while 12 test participants found condition *30* to be better than *D2* for stimulus *B*, 3 did not distinguish them and 5 preferred *D2*; the corresponding numbers for *D3* were 8, 6, and 6.

4.6 Phase 2: Research on Static Content

Similarly to the previous research phase, static models with plain-colored backgrounds were used. The models differed from the previous three, and more complex experimental configurations were used.

4.6.1 Models

The experiments of Phase 2 used 8 static models, 4 of which were frequently selected for stimulus rendering. Figure 4.15 shows these models, which can be clustered into 4 categories, 2 sources each: (top left) simple shapes with plain colors, (top right) mathematical bodies with high structural complexities, (bottom left) spatially diverse and textured objects, and (bottom right) laser-scanned real statues. The models were either designed by Holografika, or accessed from publicly available databases (the mathematical bodies[11] and the laser-scanned statues[12]).

The mathematical bodies received a particularly high level of attention during the times of experimental design and the related results were much anticipated, as they are extremely sensitive to angular resolution reduction due to the detailed structures

[11]George W. Hart's Rapid Prototyping Web Page:
www.georgehart.com/rp/rp.html

[12]Jotero.com 3D-Scan and 3D Measurement:
forum.jotero.com/viewtopic.php?t=3

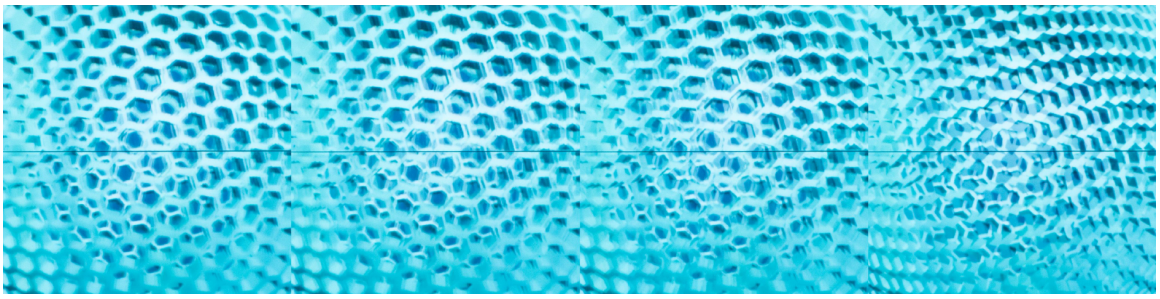


Figure 4.16: LF: Angular resolution reduction of a Phase 2 model.

and the variations in depth. Figure 4.16 shows a series of camera-captured images, continuously reducing the angular resolution of the content (80, 60, 45, and 20 views).

4.6.2 Research on Static Observers

4.6.2.1 Research Aim

The aim of the research was to assess the perceived angular resolution of the content with static observers [177].

4.6.2.2 Test Conditions

The only variable in the test conditions was the angular resolution of the stimuli. The 10 different number of views started from 80, and was reduced by steps of 10 until 45 (80, 70, 60, 50, and 45 views) and then it was reduced by steps of 5 until 20 (40, 35, 30, 25, and 20 views). The aim of this setup was to obtain a finer differentiation between angular resolutions that come with notable disturbances in the smoothness of the parallax effect. As the two mathematical bodies and the two laser-scanned statues were used as models, there were 40 test stimuli to be assessed.

For each and every test condition, participants had to evaluate quality on two scales. One was a 25-point quasi-continuous scale for the perceived quality. Due to the fact that the size of the scale was not apparent to the test participants, visual decisions were made instead of numerical ones, enabling the expressions of minuscule visual differences. The other one was a binary scale for quality acceptance. This scale was found particularly important for the research, as such a scale not only clarifies the data collected by the other scale — e.g., a 15 out of 25 can be interpreted completely differently for two test participants — but also reports on the final judgment of the user regarding the quality of the displayed content. In practice, at the end of the day, most of the active user decisions are simply binary (e.g., buying a display or not).

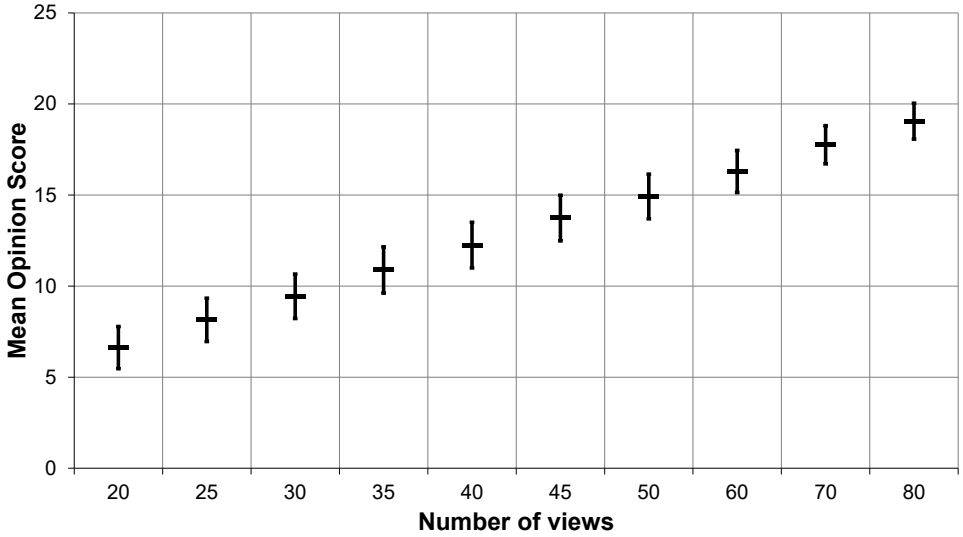


Figure 4.17: LF: Mean Opinion Scores of the subjective test on static observers.

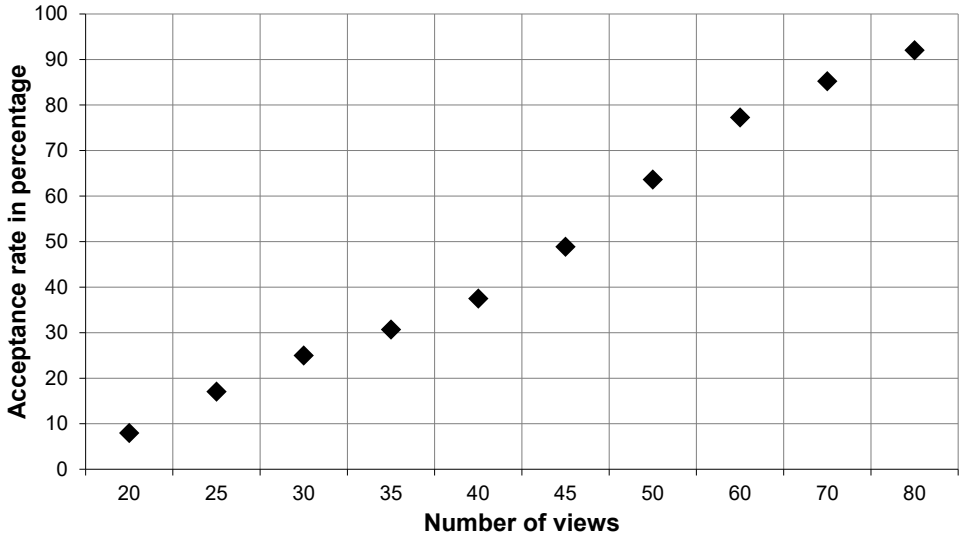


Figure 4.18: LF: Overall quality acceptance in the subjective test on static observers.

In case of this experiment, the test participant decided whether he or she found the perceived quality acceptable or not.

Viewing conditions in this research played an essential role. There was no observer motion in this experiment. In fact, test participants were assigned actual seats during the subjective tests. There was a total of 6 seats, and they were aligned in the following setup: one seat was the usual center view at 4.6 meters, with one seat on the left and the right with a meter separation, and these three seating positions were repeated a meter behind this row, at the viewing distance of 5.6 meters, which

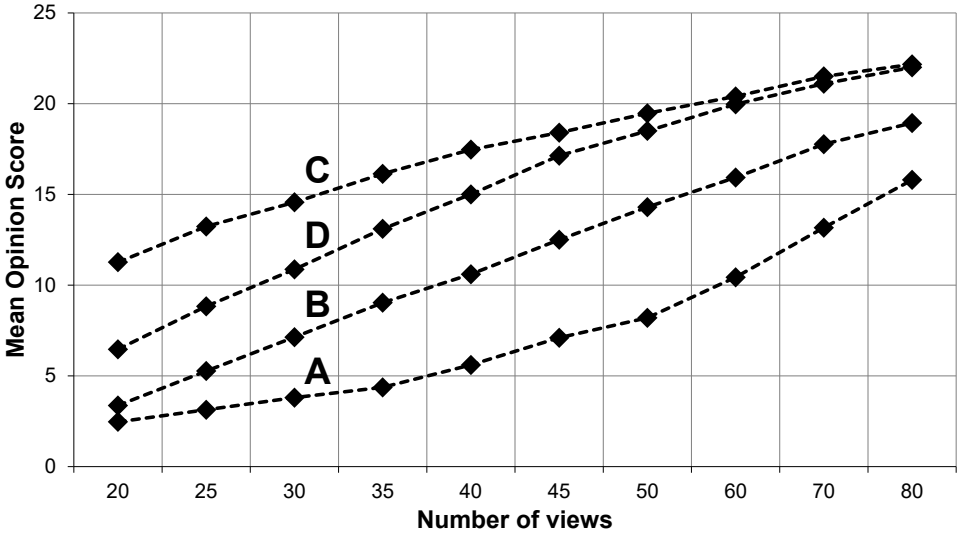


Figure 4.19: LF: Mean Opinion Scores per source stimulus in the subjective test on static observers.

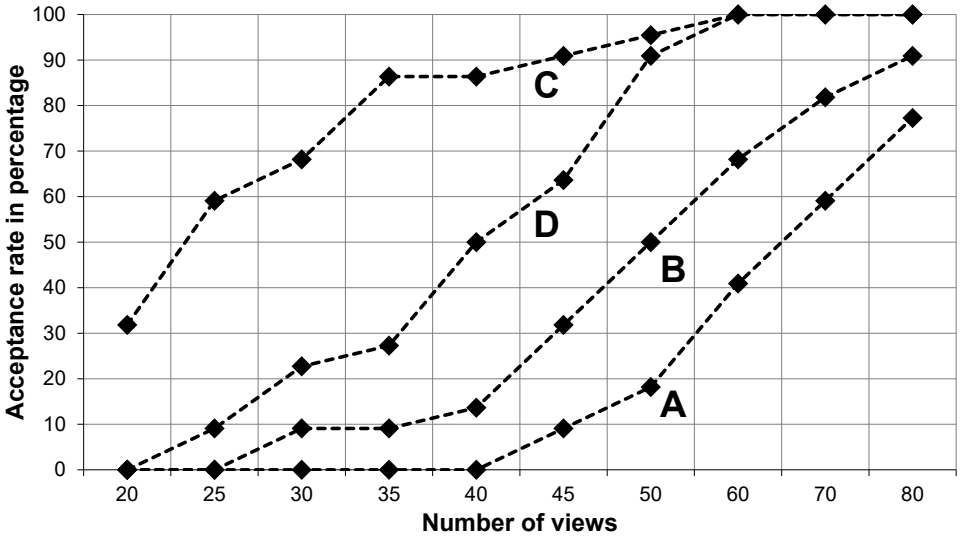


Figure 4.20: LF: Quality acceptance per source stimulus in the subjective test on static observers.

corresponded to approximately 3 H.

4.6.2.3 Results

A total of 22 individuals participated in the experiment (18 males and 4 females). The age range was from 23 to 59, and the average age was 31.

The mean scores are shown on Figure 4.17, and the overall rates of quality acceptance are shown on Figure 4.18. Beyond the near-linear nature of the results, the

mean scores indicate that the test participants did not utilize the top and bottom of the 25-point quasi-continuous scale. Or at least this is something one could assume without investigating the quality assessment in detail.

However, Figures 4.19 and 4.20 show the results per source stimulus, and we can see significant differences originating from the models themselves. Stimulus A (polyhedron with 972 faces) suffered the most from angular resolution reduction; its high level of sensitivity was earlier shown on Figure 4.16. In fact, below 1 view per degree (45 views), not a single test participant found it acceptable. Stimulus B (structure of 120 regular dodecahedra) showed similar tendencies, but since only a lower portion of its body stood out much from the plane of the screen, visual degradations came more from its internal structure — which was not as close to the test participants as the surface of stimulus A — and therefore, it was more resilient to angular resolution reduction. From the two laser-scanned statues, stimulus C with its small spatial dimensions endured the loss of angular resolution very well. Between 60 and 35 views, the only notable degradation in its quality was regarding the arm reaching out. The greatest difference in acceptance can be observed at 45 views: while stimulus C was at 91%, stimulus A was at 9.1%.

In general, the results suggest the non-trivial value correspondence between quality ratings scales and the binary acceptance scale due to subjectivity. Furthermore, with the extended training session prior to the subjective tests, the results achieved a level of rating consistency that was lacking from the research on angular resolution in Phase 1.

4.6.3 Research on Interpolation

4.6.3.1 Research Aim

The aim of the research was to assess the perceived quality of interpolation techniques against low angular resolution [178].

4.6.3.2 Test Conditions

The two mathematical bodies and the two laser-scanned statues were selected for this experiment, and the stimuli were created by either using direct rendering in a given angular resolution or by interpolation techniques. The two interpolation techniques used in this work were the disparity based [179] and the sweeping planes based [180,181] interpolation. Three particularly low angular resolutions were chosen, as light field view interpolation benefits the content only if it is perceptually lacking

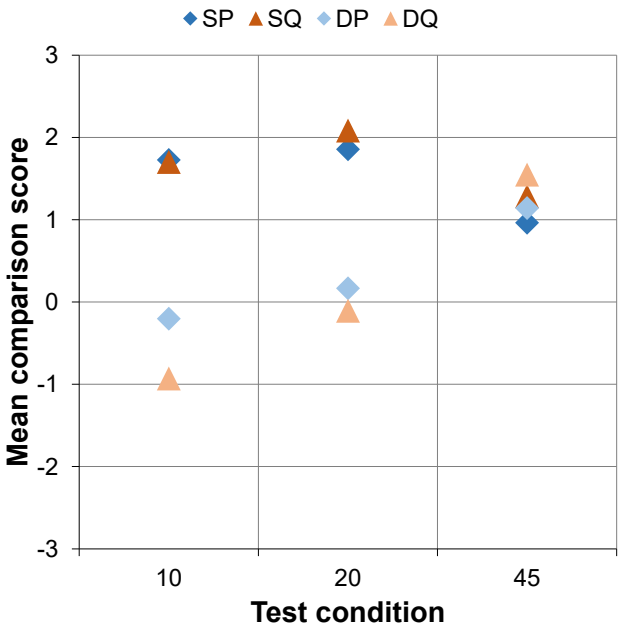


Figure 4.21: LF: Mean comparison scores of the subjective test on interpolation.

in the angular domain. These values were 45, 20, and 10 views. Note that 10 source views for a 45-degree FOV is so insufficient that from all the ten experiments on light field visualization presented in this thesis, this is the only one which involves such a low angular resolution. Of course, in this experimental scenario, using it was a valid choice, as it tested the capabilities of the interpolation techniques.

Applying these techniques to the sources, the respective views were 221, 191 and 181. Reference stimuli were also created, directly rendered in these high angular resolutions. Subjective assessment was carried out using a 7-point scale in a paired comparison, along two criteria. One was the general visual quality of the displayed content, and the other one focused on the smoothness of the horizontal motion parallax. It was important to separately address QoE with these criteria, as interpolation — which is an estimation, after all — affects the visual quality of intermediate views, and low angular resolution affects the parallax effect. Each interpolated stimulus was compared with the other type of interpolation, the corresponding low-resolution and the high-resolution rendering. Thus, there were 15 distinct pairs, and with 4 source models, the total number of paired comparisons was 60. The test participants observed the stimuli from a fixed 2.5 H distance and had a sideways movement of a meter in each direction.

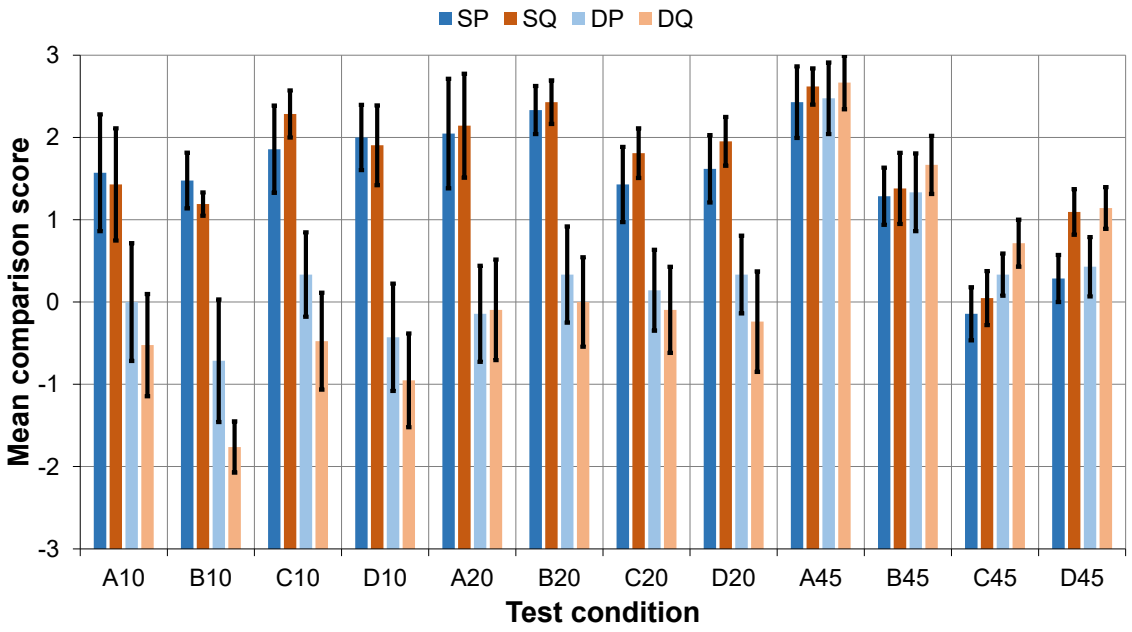


Figure 4.22: LF: Mean comparison scores per stimulus in the subjective test on interpolation.

4.6.3.3 Results

A total of 21 individuals participated in the experiment (16 males and 5 females). The age range was from 23 to 49, and the average age was 31.

In this joint scientific effort with A. Cserkaszky, while his research focus was on the performance comparison of the two interpolation techniques — comparing the interpolated stimuli to the reference ground truth and to each other — I solely addressed the preference between the inputs and the outputs of these techniques. Therefore, the results presented in this subsection are limited to the comparison of stimuli with low angular resolutions and the interpolated ones with the corresponding high angular resolutions.

The mean comparison scores of the experiment are shown on Figure 4.21. The sweeping planes based technique (S) performed exceptionally well for both the aspect of the smoothness of the parallax effect (P) and general visual quality (Q), especially for the lowest angular resolutions. The disparity based method (D) only performed adequately when the number of input views was sufficient, otherwise the output of interpolation was either approximately the same quality or slightly worse. Although no statistically significant difference was found between the aspects, we can see that visual quality was notably degraded due to the insufficient input of the disparity based method.

Let us now view the results in detail, separated by the source stimuli, as shown on Figure 4.22. The naming convention of the four stimuli are exactly the same as in the previous experiment: A and B are the mathematical bodies, C and D are the laser-scanned statues.

The results obtained from the comparison of the 45-view stimuli and the corresponding interpolated stimuli are in perfect alignment with the ratings of the previous experiment on angular resolution. As stimulus A had a very poor visual appearance at this angular resolution, interpolating it, and thus, increasing its angular resolution significantly boosted all of its quality aspects, for both interpolation techniques. On the other hand, since the visual quality of stimulus C was already highly assessed and commonly accepted at this angular resolution, it did not benefit much from interpolation.

For the lower angular resolutions, as seen on the mean scores, while the sweeping planes based method actually managed to improve the quality aspects, the disparity based method struggled to keep the level of the input, since its inaccurate estimations due to the low-density inputs severely damaged the appearance of the stimuli. Although interpolation techniques, in general, do improve the smoothness of the horizontal motion parallax, it is challenging to notice such improvement when the visual quality of the stimulus is greatly degraded.

4.6.4 Research on Resolution Interdependence

4.6.4.1 Research Aim

The aim of the research was to assess the perceived interdependence between spatial and angular resolution [182].

4.6.4.2 Test Conditions

In order to investigate how spatial resolution affects angular resolution, 3 values were chosen for each, and the test stimuli were directly rendered in the 9 combinations. For spatial resolution, 1440×1080 (*S1*), 1024×768 (*S2*) and 640×480 (*S3*) were selected. For angular resolution, 135 (*A1*), 45 (*A2*) and 30 (*A3*) views were used. As this was designed as a more exhaustive experiment, all 8 source models were involved.

A seven-point paired comparison scale was used to assess the smoothness of the horizontal motion parallax. The pairs were configured in a way that the angular resolution was the same, but spatial resolution differed. Therefore, there were 9 comparison types, and with 8 models, there was a total of 72 comparison pairs. The

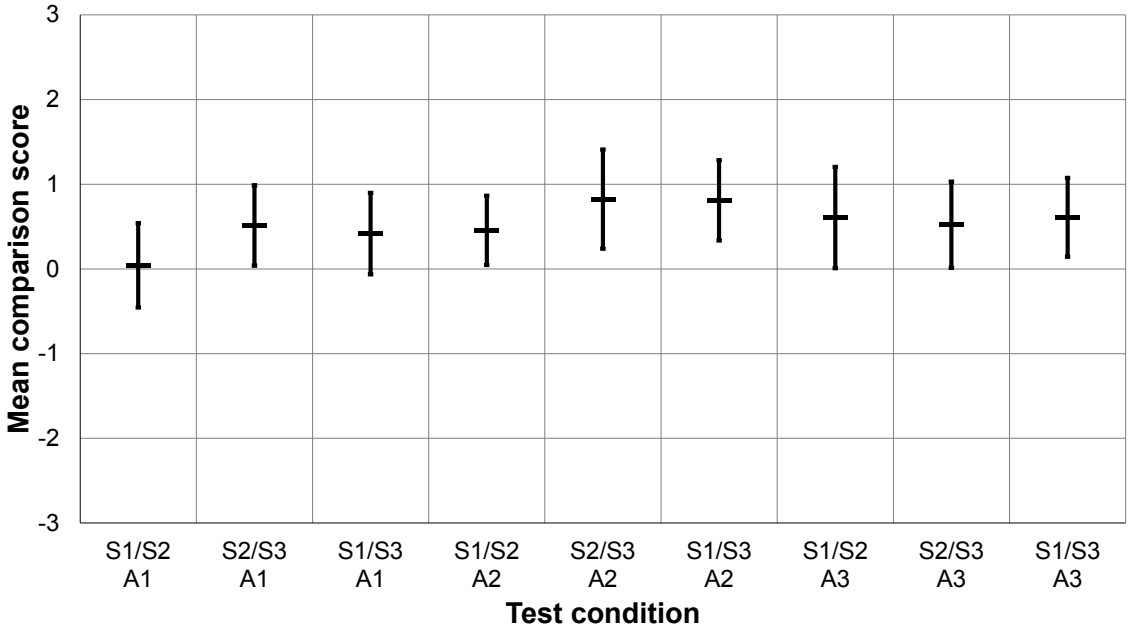


Figure 4.23: LF: Mean comparison scores of the subjective test on interdependence.

test participants observed the stimuli from a fixed 2.5 H distance, and had a sideways movement of a meter in each direction.

4.6.4.3 Results

A total of 22 individuals participated in the experiment (16 males and 6 females). The age range was from 18 to 58, and the average age was 31.

Figures 4.23 and 4.24 show the mean comparison scores and the scoring distribution, respectively. Although no statistically significant difference was found among the test conditions, the results indicate that the reduction of spatial resolution increased the perceived smoothness of the parallax effect. The smallest benefit was measured at *A1*, since the smoothness of the horizontal motion parallax was undisturbed, due to the high angular resolution. It is important to note that it is not necessarily true that the lower the angular resolution is, the more spatial resolution reduction compensates parallax smoothness. In practice, it needs to be taken into account that if the angular resolution is *too* low, then the level of degradation that affects visualization prohibits the visual distinction of this aspect between the stimuli.

In total, 18.7% of the test participants experienced degradations in parallax smoothness after spatial resolution reduction, 31.3% could not distinguish this aspect of the stimuli, and 50% reported improvements. For *A1*, *A2* and *A3*, the average rates of improvement were 40.2%, 58.3% and 51.5%, respectively. The rating tendencies were

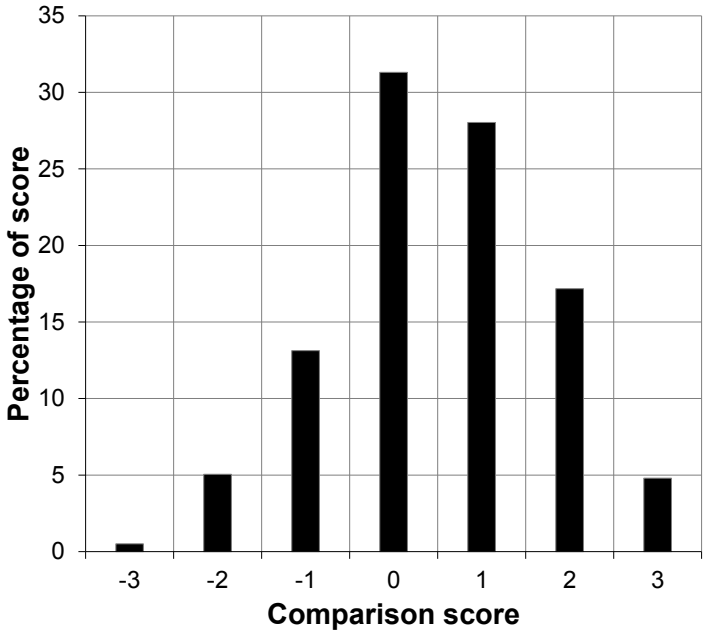


Figure 4.24: LF: Scoring distribution of the subjective test on interdependence.

the same for all sources and no significant difference was found. The average amount of improvement regarding the smoothness of the parallax effect was indeed modest, but measurable, and these finding shall be taken into account for the upcoming phases of the research on light field visualization.

4.7 Phase 3: Research on Light Field Video

In Phase 3 and Phase 4, video stimuli were used for the subjective tests.

4.7.1 Source Videos

A frame of each source stimulus — captured by a pinhole camera — is shown on Figure 4.25. The first two, named *Red* (14.4 sec, 130,240 kbps) and *Yellow* (13.6 sec, 704,160 kbps), were provided by Freelusion[13] with the aim of using the depth budget of the light field display to a great extent. The following two, *Ivy* (10 sec, 38,000 kbps) and *Tesco* (12.5 sec, 76,800 kbps), were created by Post Edison[14], targeting subtle and intense motions, respectively. The final source stimulus, *Gears* (7.2 sec, 286,000 kbps), was a looping animation based on a static model of Phase 1, rendered by Holografika.

[13]Freelusion Video Mapping + Dance Company:
http://freelusion.com/

[14]Post Edison Computer Graphics:
http://www.postedison.hu/

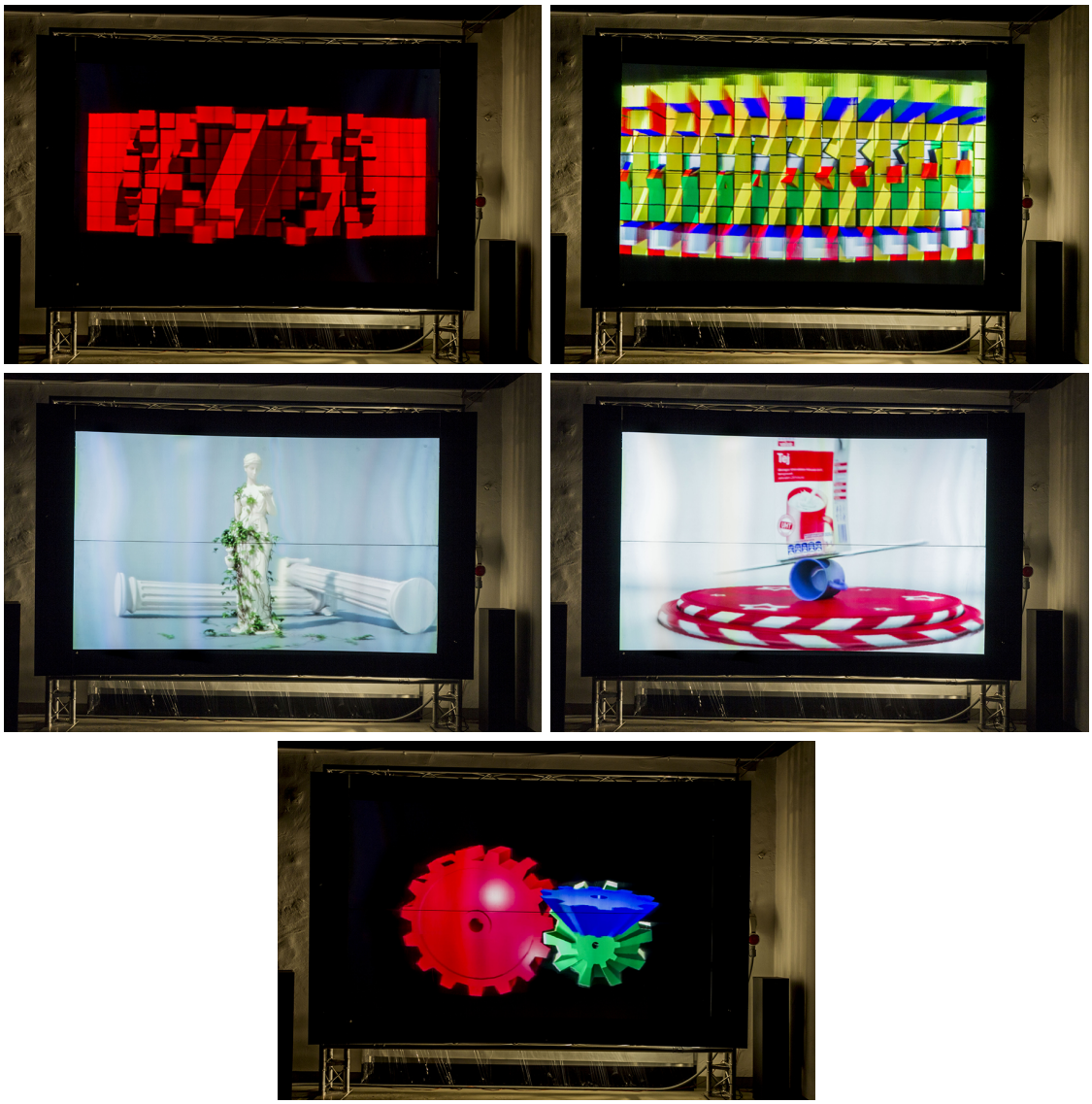


Figure 4.25: LF: Source video contents (*Red*, *Yellow*, *Ivy*, *Tesco*, and *Gears*) used in Phase 3 and Phase 4, visualized on the HoloVizio C80.

All source videos had a frame rate of 25 fps and all were involved in every experiment of both Phase 3 and Phase 4.

In the process of content rendering for the light field display, a total of 20 NVIDIA GeForce GTX 960 2GB GDDR5 128 bit GPUs were used as conversion hardware. The computation with this equipment took roughly 30 ms per video frame. All test stimuli were locally played on the device during the experiments, and therefore, no video transmission was necessary. The spatial resolution of the source varied per content and all the above-mentioned videos were available in an angular resolution of 0.5 degrees.

4.7.2 Research on Viewing Conditions

4.7.2.1 Research Aim

The aim of the research was to address static and moving viewing conditions with regards to the variation in angular resolution [183].

4.7.2.2 Test Conditions

At the time of writing this thesis, the viewing conditions of subjective QoE tests on light field visualization quality are not standardized yet. There are certain issues and phenomena that need to be considered when addressing viewing conditions.

First of all, in case we consider a human observer with two eyes, we can say that the 3D experience requires that the two eyes can be addressed with two separate light rays. If the distance compared to the angular resolution of the display is too great, then this does not occur, and visualization is perceived to be 2D. Let us denote the angular resolution of the display by AR, the average distance between the eyes of the observer as D_E, and the viewing distance at which the 3D experience is still supported as D_V. The rule of thumb in this case is

$$D_V = \frac{D_E}{tan(AR)}. \tag{4.1}$$

If we take D_E as 6.5 cm and AR as 0.5 degrees, then D_V is 745 cm. However, (4.1) is calculated for a perfectly still human observer, which does not exist in practice. The general approach in industry is that still (i.e., not moving) observers can experience 3D visuals at $2 \times D_V$, and moving observers at $3 \times D_V$. Therefore, the investigation of subjective stereoscopic disparity is out of the scope of the experiment due to the spatial limitations of the given laboratory environment. However, it does propose a valid research question and it should be addressed in future works, especially with regards to standardization.

Second, the position of a static observer determines the perceived orientation of the video scene. If the observer is moving, then variations in observer positions will affect the perceived orientation of video frames; basically, if two observers do not have their motion patterns perfectly synchronized, they see certain frames from different angles, so different test participants do not see the content in the exact same way. From an experimental point of view, this is something to be avoided, as such variations may induce bias in the collected results, yet motion needs to be addressed, since the angular-dependent nature of light field visualization is one of its most valuable features.

In case of HPO displays, the parallax effect in the context of this visualization technology is commonly known as the smooth, continuous horizontal motion parallax. An observer does not need two separate eyes to perceive the parallax effect in the natural world, and does not even need to move. A perfectly still human head with one functioning eye is sufficient to collect visual information from which the brain can create the experience of parallax. However, this only works if the eye can change its orientation. If it cannot, then it corresponds to a mere pinhole camera in a fixed position and orientation. Therefore, test participants are always capable of experiencing the parallax effect of the light field display, regardless of viewing conditions, of course, if they are located within the valid FOV and view the display from a reasonable distance that evidently supports stereoscopic disparity.

Again, regardless of viewing conditions, there is always perceived parallax. Even though it is trivial that parallax is perceived, it is not trivial at all *how* it is perceived. This is particularly relevant when the smoothness of the parallax is disturbed. Also, it is called "motion" parallax, as the effect can only be fully perceived with sufficient alteration in the horizontal component of the viewing position; otherwise, the difference in the speed of perceived location change between closer and farther objects are difficult to witness.

Taking all the aforementioned considerations into account, three static positions were chosen: one was the default 2.5 H center view, and two other positions were its 1-meter sideways shifts (left and right). These were directly compared with motion patterns: one going from the left position to the right, and returning to the left, and its corresponding mirrored pattern.

At first, the reference stimulus was shown for each content, with high angular (2 source views per degree) and spatial resolution (see Table 4.1 for *HS* values). The two degraded test conditions were one with reduced angular resolution (1 source view per degree), and one with reduced angular and spatial resolution, denoted as *LA-HS* and *LA-LS*, respectively. Note that according to (4.1), $2 \times D_V$ is 7.45 m if AR is reduced to 1 degree, thus, the 4.6 m corresponding to 2.5 H distance was deemed suitable for 3D experience.

The test participants had to use a 5-point DCR scale to directly compare the perceived smoothness of the motion parallax of the different viewing conditions, with the static position as reference. They also had the option to provide extensive verbal feedback regarding what they perceived, assisting the understanding of the results.

Table 4.1: LF: Spatial resolution of test conditions per source.

Source ID	**HS**	**LS**
Red	1024×768	640×480
Yellow	1024×768	800×600
Ivy	960×540	640×360
Tesco	1280×720	640×360
Gears	1920×1080	640×360

4.7.2.3 Results

A total of 18 individuals participated in the experiment (15 males and 3 females). The age range was from 20 to 42, and the average age was 29.

The first test compared the motion patterns. The research aim here was to check whether both these patterns needed to be compared to the 3 fixed position or not. A binary scale was used to assess the perceived difference in the smoothness of the parallax effect. Results indicate that not a single observer was able to differentiate these patterns with regards to parallax smoothness, therefore, only one was used in the subjective test, whichever the observer personally preferred.

The results of the subjective test are introduced on Figure 4.26, in the forms of obtained scoring distribution and mean scores. First of all, only 36.92% of the scores were 5 (no perceivable difference in the smoothness of the horizontal parallax); 63.08% of the scores indicate a perceivable difference induced by observer motion. In fact, the number of 4 ratings (perceivable difference that does not annoy the observer) alone surpassed the number of 5 ratings. 3 ratings (slight annoyance) and 2 ratings (considerable annoyance) were at 21.8% and 4.1%, respectively, and not a single observer used the bottom of the scale (high level of annoyance).

As for the two test conditions, the obtained results show a statistically significant difference, therefore, the horizontal parallax was deemed to be smoother when spatial resolution was reduced as well (*LA-LS*). This is quite notably reflected in the scoring distribution as well. Roughly twice as many ratings registered annoyance for *LA-HS* than *LA-LS*, and the two test conditions also reached more than a 10% difference for 5 ratings.

Regarding the viewing positions, no statistical difference was found. With 5 contents and 2 test conditions, there were 10 position triplets per observer. Nearly half, 47.7% of these triplets were without variations (all 3 scores were the same), 40% had a single different score and 12.3% had 3 different scores. However, due to the

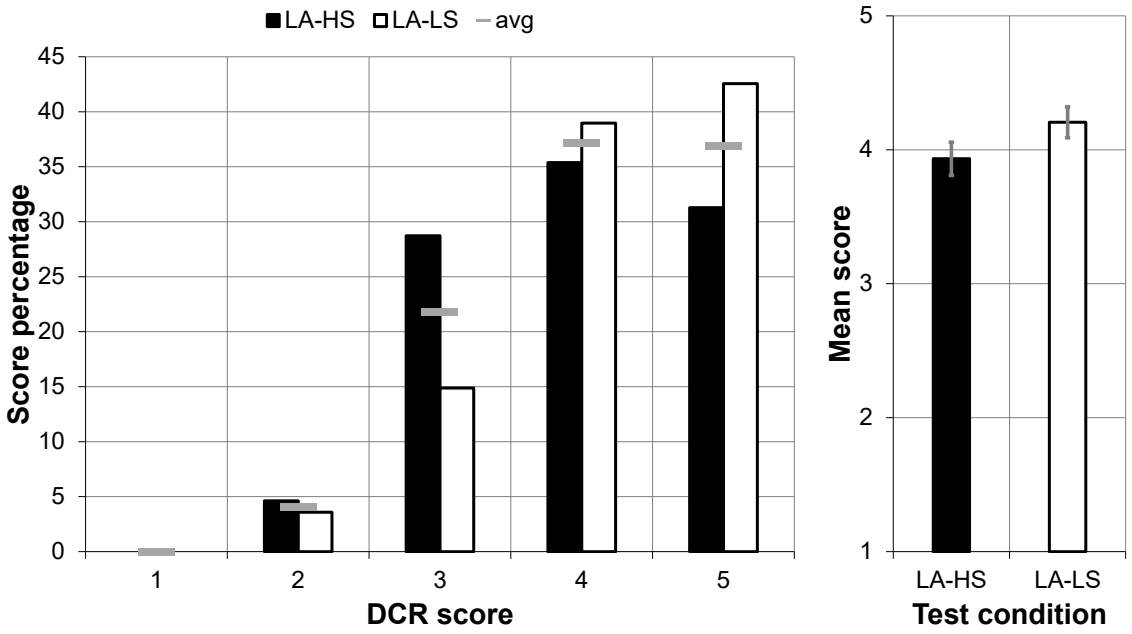


Figure 4.26: LF: Scoring distribution (left) for test conditions with low angular resolution and high spatial resolution (*LA-HS*), with low angular and low spatial resolution (*LA-LS*), and for all scores (avg) in the subjective test on viewing conditions. Mean scores (right) for *LA-HS* and *LA-LS*.

great variations in the scoring dissimilarities, there was no dominant pattern and the averages fit into an interval of 0.05.

Finally, the findings for the selected source contents are similar to viewing positions. Although there were indeed certain differences for content scores per observer, there was no obvious perceptual preference. For example, some did not perceive any degradation for *Red* and *Yellow*, while others not only perceived disturbances in the smoothness, but also rated them to be slightly annoying.

4.7.3 Research on Video Resolution

4.7.3.1 Research Aim

The aim of the research was to assess the perceived spatial and angular resolution of the video content [184].

4.7.3.2 Test Conditions

As the research focused on the spatial and angular resolution of light field video, these two parameters were the only variables. To each source video content, 2 settings per

Table 4.2: The investigated test conditions. *S1* is the first stimulus in the pair and *S2* is the second. *AR* indicates the setting of angular resolution, and *SR* refers to spatial resolution.

Test condition	**Rating type**	**S1AR**	**S1SR**	**S2AR**	**S2SR**
A	DCR	High	High	High	Low
B	DCR	High	High	Low	High
C	DCR	High	High	Low	Low
D	PC	High	Low	Low	High
E	PC	High	Low	Low	Low
F	PC	Low	High	Low	Low

parameter were applied, resulting in 4 initial test conditions. The higher resolution was denoted as *High*, and the lower one as *Low*.

The subjective test itself was carried out by using two methodologies. One was the 5-point DCR, which compared test conditions containing *Low* setting(s) to the reference quality (where both settings were *High*). The other type was a 7-point paired comparison (PC), to directly compare the stimuli containing *Low* setting(s) with each other. The list of the investigated test conditions — summarizing the aforementioned cases — is given in Table 4.2. The *High* and *Low* spatial values were the same as in the previous experiment, defined in Table 4.1. The selected angular resolutions for the *High* and *Low* profiles were 2 and 1 source views per degree, respectively.

The video stimuli were presented to the test participants in a randomized order, clustered by test type. This means that test participants did not have to switch back and forth between the two subjective assessment tasks, making evaluation more straightforward and focused. The experiment also accommodated the option of providing detailed feedback regarding the perceived differences, but it was not mandatory.

4.7.3.3 Results

A total of 18 individuals participated in the experiment (15 males and 3 females). The age range was from 21 to 42, and the average age was 28.

Figure 4.27 introduces the mean results and the distribution of the ratings among the different options. Test condition A (low spatial resolution) mainly received scores of 4 ("*Perceptible but not annoying*"), in 58.8% of the ratings. The mean score of A was also boosted by the fact that in 27.7% of the ratings test participants could not detect degradations. Also, A is the only test condition for which no score of 2

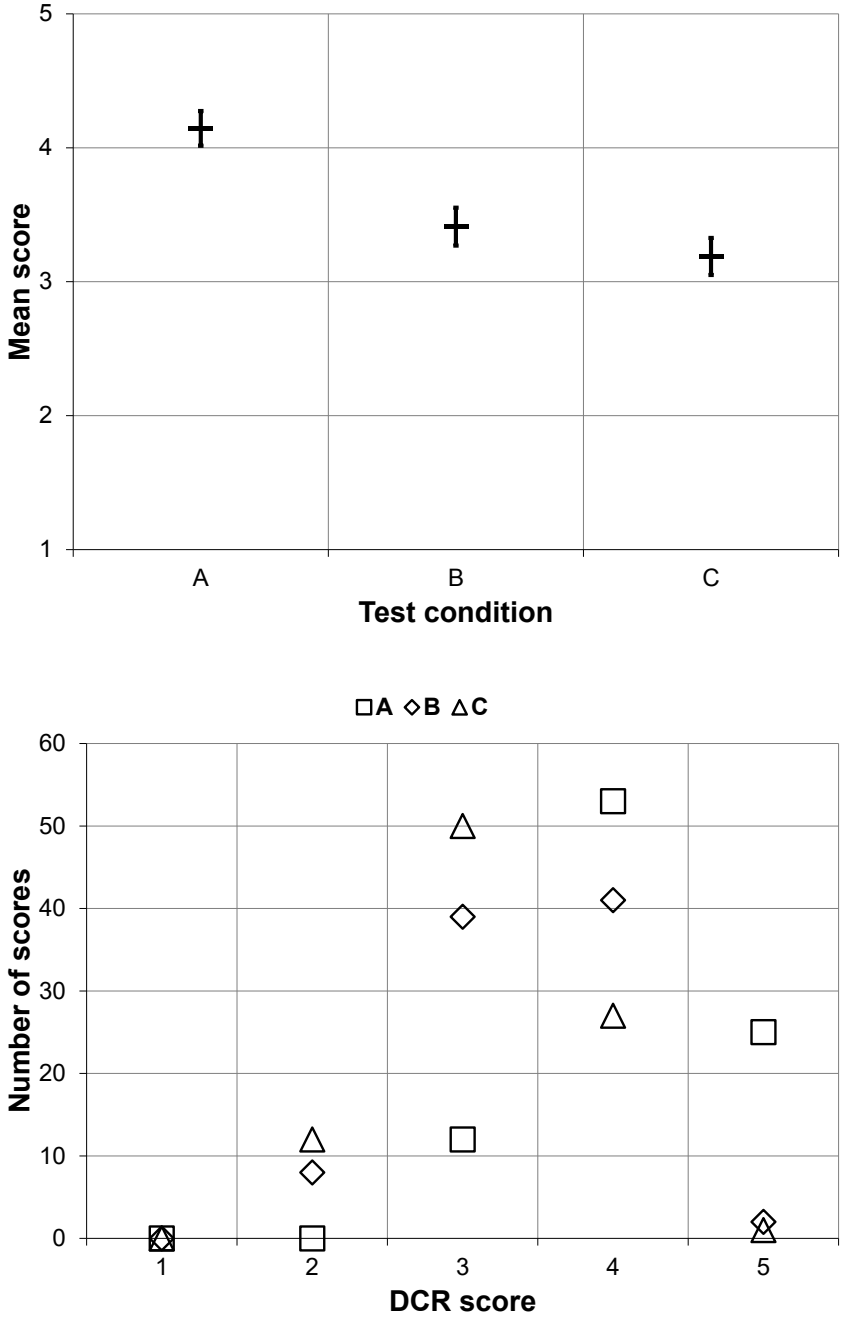


Figure 4.27: LF: Mean scores (top) and rating distribution (bottom) of the DCR assessment of the subjective test on video resolution.

("*Annoying*") or 1 ("*Very annoying*") was given at all. The mean is above 4 and it is significantly better than what the other two test conditions obtained.

Although test condition B (low angular resolution) did receive a higher mean score than C (low spatial and angular resolution), they are not significantly different from each other. Similarly to A, no score of 1 was given; however, there were a few scores

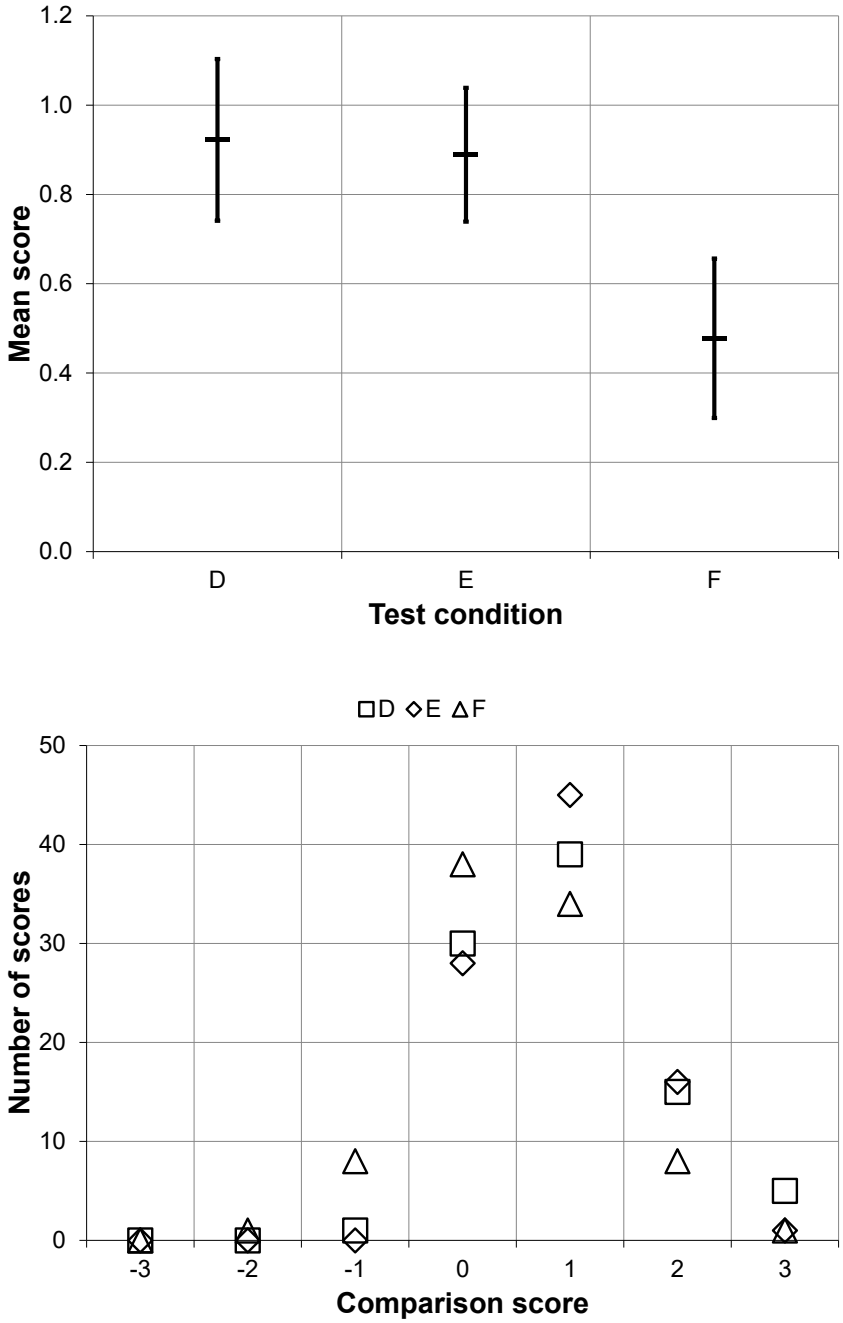


Figure 4.28: LF: Mean scores (top) and rating distribution (bottom) of the PC assessment of the subjective test on video resolution.

of 2 (8 and 12, respectively for B and C). Generally, while some perceived differences for test B were more without annoyance, it came with slight annoyance for test C.

The important findings here are the relations in significant differences. Basically, these results indicate that reducing the spatial resolution of the content will not degrade the perceived quality significantly when content angular resolution is already

low. Combined with the findings on resolution interdependence, the smoothness of the parallax effect can be improved without compromising the overall subjective quality.

Regarding the source video contents, the same scoring patterns generally applied. The highest mean scores were obtained by content *Tesco* and *Gears*. According to the additional feedback provided by the test participants, the multiple simultaneous motions in content *Tesco* made it difficult to focus at specific parts of the scene, as attention was divided. In content *Gears*, this was the opposite, as the main focus was clearly on the edges of the objects, as those were the primary victims of quality degradations. Yet due to this narrowed focus, some perceptual differences went unnoticed. From the 28 "*Imperceptible*" scores in this part of the experiment, 16 (57.14%) were given during the visualization of these two source contents.

In the analysis of the PC scores, positive scores indicate the preference of stimulus 1 (*S1*) over stimulus 2 (*S2*), e.g., -1 means that *S1* was assessed as "*Slightly worse*" / *S2* was assessed as "*Slightly better*".

Figure 4.28 introduces the mean results and the distribution of the ratings among the different options. The first important observation is for test conditions D (low spatial resolution compared to low angular resolution) and E (low spatial resolution compared to low spatial and angular resolution). The obtained scores are very much alike, and it reinforces the DCR assessments of test B and C. This means that the PC study also shows the lack of significant difference between the stimuli with only angular resolution reduced and the stimuli with both parameters reduced.

As the vast majority of the ratings were either positive or zero, it shows that test participants clearly preferred content spatial resolution reduction over angular resolution reduction, regardless of the change in spatial resolution. The highest number of 0 ("*Same*") and -1 ("*Slightly worse*") scores were used to assess test condition F, thus, in those cases, either the stimuli of B and C could not be distinguished from each other, or C was actually perceived to be better. Also, the mean score of F is below 0.5 and it is significantly lower than what D and E received. These further support the findings of the DCR test, portraying the lack of the aforementioned significant difference.

As for the different source video contents, the relations are rather similar to what was seen for the DCR study. Contents *Tesco* and *Gears* provided the smallest observable differences, and therefore, their mean scores are the lowest. Contents *Red* and *Yellow* received the lowest scores in DCR, and quite similarly the highest scores in PC. This is mostly due to the high utilization of the depth budget of the display.

Regarding content *Ivy*, the scene had the least dynamism with its very subtle animation, yet enabled the test participant to assign more focus to the background and its degradations.

In the optional user feedback, it should be noted that most of the test participants distinguished two types of blur: a general blur (spatial resolution reduction) and a blur that applied more to parts outside the plane of the screen (angular resolution reduction). They both affected the perceived quality of light-field video, but as the obtained results show, angular resolution was more critical.

4.8 Phase 4: The Dynamic Adaptive Streaming of Light Field Video

In this final phase, the focus was on the use case of real-time video transmission, particularly light field video streaming. It is an under-investigated topic, as many find it too early to consider such an application on the level of research. The only other work to address the adaptive streaming of light fields is the contribution of Wijants *et al.* [167], which uses static contents. Current efforts have a greater focus on still image visualization than video content transmission. For instance, if we look at the large collaborations, the JPEG Pleno framework [185] is already developing a light field format for standardization purposes. MPEG-I for immersive media has only recently started addressing light field. There are indeed certain milestones that need to be passed before implementing light field video streaming services — i.e., reductions in end-user device cost and in streaming data sizes — but the currently available technology already enables research in the QoE of light field visualization to be carried out, including real-time video transmission.

Dynamic Adaptive Streaming over HTTP (DASH) [186–189] for conventional 2D visualization tackles the choice of the lesser evil in QoE: in case of lower available bandwidth, users tend to tolerate a temporary reduction in spatial resolution more than playback interruptions in the forms of stalling. Still, at the end of the day, both options result in the degradation of user experience, and the less frustrating one needs to be chosen. Today, it is a common practice that the user is actually given the choice of manually selecting the video resolution of a given content, overriding the DASH concept of lower-quality representations if necessary (e.g., if the video is a recording of an online game and the viewer would like to read information that is typically written in small font). In QoE research, similar dilemmas are frequent, such as being caught "between the devil and the deep blue sea" [190], which addresses initial waiting times

and stalling events. In the domain of DASH streaming, quality switching, and stalling are the most critical impairments, and in the domain of light field visualization, video quality and video applications, in general, are currently under-investigated, yet hold great potential.

4.8.1 Concept

The proposed quality switching protocol [191] is based on multiple representations of the video content, with different spatial and angular resolution values. It is important to note that dynamic adaptive light field video streaming only makes sense if display-independent data is considered (e.g., an array of views), that is to be converted at the client's side. It is debatable today whether we can consider conversion to be sufficiently fast and efficient for real-time applications. The process fundamentally depends on the input itself; the higher the resolution, the more the time conversion takes. To demonstrate its feasibility, if we take only 18 source views, such as in the live capture system of Adhikarla *et al.* [169], conversion is in the order of 10 ms, which is perfectly suitable for such time-critical use case scenarios. A higher angular resolution, such as 80 source views, can still fit into the order of 100 ms. Using more sophisticated display-independent light field formats can also maintain a low conversion time [192].

The feasibility for real-time application can be further exhibited in more details via the light field telepresence solution, presented by Holografika at the T.um technological demonstration exhibition of SK Telecom in late 2017 and then published in 2018 [132]. The solution is practically a light field camera system with an integrated light field display. In one room, the light field camera system captures the full body of a standing person with 96 cameras arranged on a 105-degree arc. The camera broadcaster then collects the frames of the cameras and streams them over the network. In the other room, the light field display system, which has a nearly 180-degree FOV, receives the camera frames, and on the computer clusters, it converts the images to the light field of the display. In this specific implementation, the total system delay is approximately 100 ms, which is measured between the arrival of the given camera frame at the display side of the system and the appearance of the actual visualization on the screen.

Conversion on a given light field display results in fixed parameters of spatial and angular resolution. This means that if a client requests and receives representations of a specific content with quality indicator *Q1* (low-quality representation) and another client with quality indicator *Q3* (high-quality representation), the converted

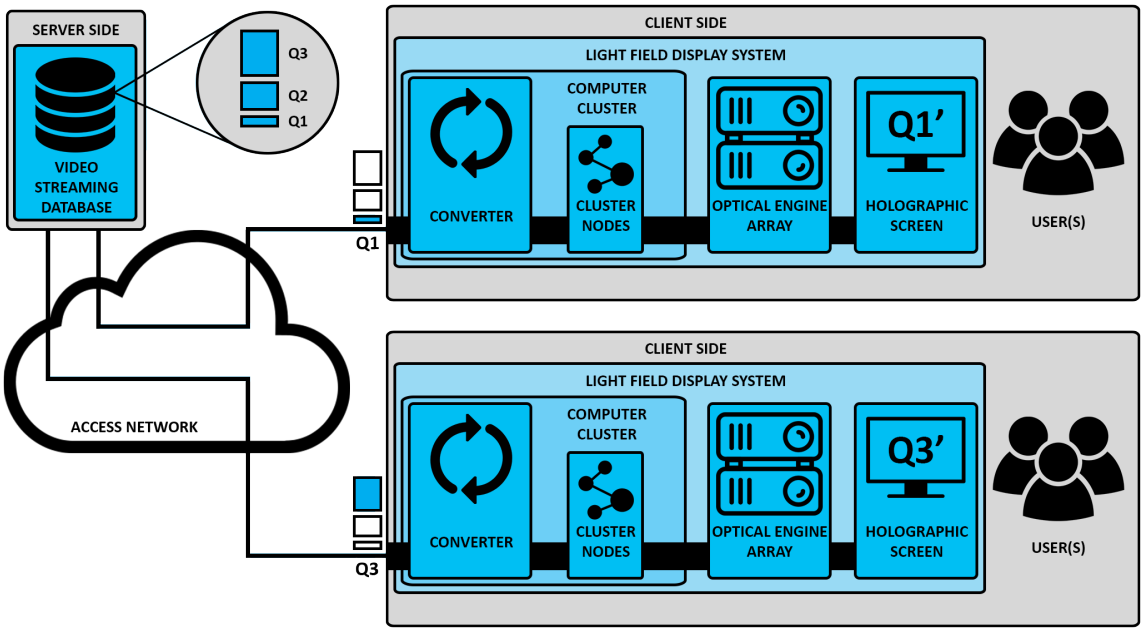


Figure 4.29: LF: Dynamic adaptive streaming of *Q1* and *Q3* quality representations of a light field video. The illustrated architecture of the light field display system employs a holographic diffuser, analogous to the HoloVizio C80 light field cinema.

and visualized *Q1'* and *Q3'* would have the same data volume, but would differ in perceived quality (see Figure 4.29 [191]).

To gain a better understanding of such visualization systems in practice, let us briefly review the functionalities of the primary components and modules present on Figure 4.29. The views of a given content are captured by a camera array, or they are rendered on computers from virtual scenes. These views are stored in the video streaming database with different angular and spatial resolutions, resulting in different quality properties and storage/bandwidth requirements. The client continually requests frames of the selected content with a given quality parameter. Once the light field display system receives the frame through the access network, it distributes them to the computers in the cluster that are responsible for rendering the 3D light field on the optical engines, that project light onto the holographic screen. The conversion process in the computer cluster interpolates the virtually continuous light field from the received discrete camera views, and from this, it renders the ideal optical engine images for the specific type of light field display system. The cluster nodes modify these optical engine images by applying the calibration parameters of the particular display system and send these images to the optical engines. The user(s) then is/are able to experience the 3D light field on the display system.

Sending display-specific data can avoid the phase of conversion, and thus, can

make the process of visualization faster. However, in that case, the server side would need to store the corresponding converted data for every single supported display type. Also, if the capabilities of the display system surpass the original parameters of the content, then the converted video is larger as well. As an example, in a conventional 2D setting, let us imagine a 720p video. If we want to stream this to an end-user system with a UHD/4K display, that means we would have to transmit a version that is upscaled to 2160p, if we follow the concept of display-specific data transmission. On the one hand, sending videos that are already converted to system specifications eliminates the phase of conversion at the user's end, but it disables the option for dynamic adaptive streaming and can also result in inefficient data transmission rates.

As for frame rate, the projector array of such systems can support visualization up to 60 fps. However, in real-time utilization, although 20–25 fps is manageable, 60 fps would be *very* challenging. Also, going below 20–25 fps in playback can easily threaten the user experience. Therefore, in the scope of this research, no optimization of the temporal domain was considered, and playback was limited to 25 fps.

4.8.2 Preliminary Research on Video Stalling

Before configuring the experiment — particularly regarding the stalling events — a series of tests were carried out on different visualization platforms at Kingston University, in order to assist the work at hand. These tests addressed the perceptual thresholds of stalling detection and the user preference regarding stalling distribution. The research also covered miscellaneous topics such as QoE over time, but they are not relevant to the scope of this section and to the presented work in general, and thus, they are not included in the thesis.

4.8.2.1 Research on Stalling Detection

4.8.2.1.1 Research Aim

The aim of the research was to assess the thresholds of 3D stalling detection [99].

4.8.2.1.2 Test Conditions

The subjective tests took place in a research environment similar to what was introduced for the HDR studies in Chapter 3. The display was a 55-inch Philips autostereoscopic 3D display (WOWvx 2D-plus-depth) with Dimenco software[15]. With an HD display visualizing HD content at full screen, the viewing distance was 6 H, which

[15]https://www.dimenco.eu/

Figure 4.30: LF+: Source videos used in the tests on stalling detection.

corresponded to approximately 4 meters. Two 30-fps videos of the manufacturer's demo sequences were used as source contents, namely "*Pinocchio*" and "*Motorcycle*" (see Figure 4.30).

For each test participant, the experiment contained two distinct parts. The first one was a perceptual threshold measurement with the aim of determining detectable stalling durations. The "*Pinocchio*" sequenced was used, with three stalling events (*P1*, *P2*, and *P3*) selected based on their TI values and motion types: *P1* had a low-TI vertical motion, *P2* had a similarly low-TI back-forth motion, and *P3* was a high-TI vertical motion. As the movement to which *P3* was assigned had the highest intensity in the clip by far — since camera orientation was changed as well — it carried the objective of being the critical stalling event for the perceptual threshold measurement (easiest to detect). The role of *P1* and *P2* was to compare two similar motions of a given model, one along the y-axis and one along the z-axis.

The three points of stalling were tested separately in 5-second clips. Each clip was first shown without stalling, and then the duration of stalling — which was implemented as frame freezing without a visual indicator — was increased by one frame in each round. This was repeated until the test participant managed to perceive and identify the stalling event. As this was a 30-fps sequence, each added frame corresponded to 33.3 ms.

The second part involved both videos, and stalling events were to be rated on a 5-point DCR scale. The "*Pinocchio*" sequence used the three points introduced in the first part, and three points for stalling events were also applied to the "*Motorcycle*" video, similarly based on TI and motion type. The duration of the stalling events ranged from 0 to 15 repeated frames, plus 20, 25, 30, and 45, in order to test higher extents of stalling as well, up to 1500 ms.

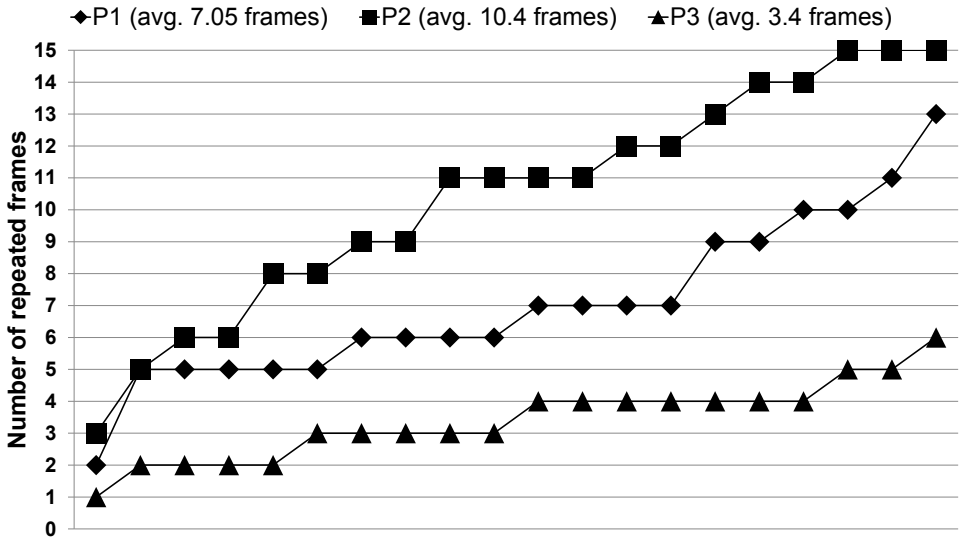


Figure 4.31: LF+: Perceptual thresholds of stalling detection. Each marker indicates the threshold of a test participant for a given stalling event.

4.8.2.1.3 Results

A total of 20 individuals participated in the experiment (16 males and 4 females). The age range was from 18 to 37, and the average age was 24.

Cognitive distortion was not only filtered by the fact that test participants had to clearly identify the stalling event, but also by starting the test with three additional instances of the reference clip before showing the stimuli with the repeated frames. Due to the sheer idea that "there should eventually be a perceptible stalling event", 6 out of 20 test participants stated that they detected stalling in at least one of these three reference videos.

The results on the perceptual thresholds of stalling detection are shown on Figure 4.31. As described in Chapter 1, each marker represents the threshold value of a test participant for a stalling event, but as the results for the different stalling events are visualized in ascending orders independently, three vertically aligned values do not necessarily belong to the same test participant (hence a horizontal axis would not carry meaningful information for the figure). The obtained results indicate a clear difference between the perception of the three stalling events. As expected, *P3* was the easiest to detect, with half of the test participants already noticing it at 3 repeated frames (100 ms) or below. Although *P1* and *P2* were similar with regards to the amount of change between subsequent frames, *P1* (back-forth motion) was easier to detect than *P2* (up-down motion). Furthermore, in order to signify the importance of stalling event position in a given sequence, note that the average of

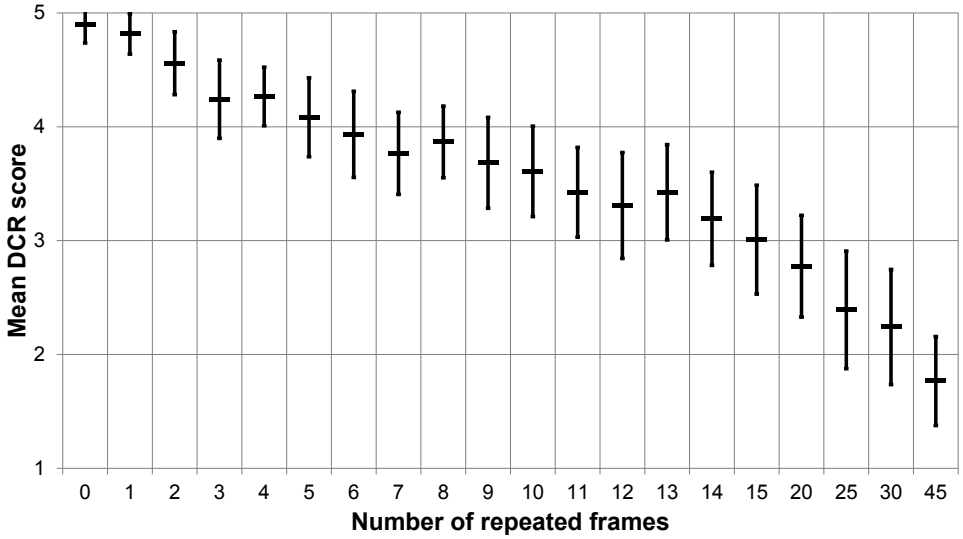


Figure 4.32: LF+: Mean DCR scores of the subjective test on stalling duration.

the number of repeated frames (i.e., stalling duration) necessary for the detection of stalling event *P2* is more than 3 times as high as for *P3*. These results also reinforce that a stalling duration of 500 ms (15 repeated frames at 30 fps) should be clearly detectable regardless of event position, and they emphasize the need for content with movement along the z-axis in QoE research on glasses-free 3D.

The mean DCR scores of the test are shown on Figure 4.32. The most relevant finding is that the results do not indicate a significant difference between the perception of stalling events of 500 ms and 1000 ms, and the same applies to the pair of 1000 ms and 1500 ms. However, the difference between 500 ms and 1500 ms is indeed significant. Therefore, using these two values in the evaluation of the concept investigated in Phase 4 is likely to provide statistically different ratings. Furthermore, while the test subjects generally deemed the stalling event with a duration of 500 ms to be "*Slightly annoying*", the other two were rather assessed with the option "*Annoying*".

4.8.2.2 Research on Stalling Distribution

4.8.2.2.1 Research Aim

The aim of the research was to assess the user preference regarding stalling distribution [193].

4.8.2.2.2 Test Conditions

The test involved three different distributions of a given overall stalling duration, which was 3000 ms. The first one — denoted as *D1* — included five stalling events

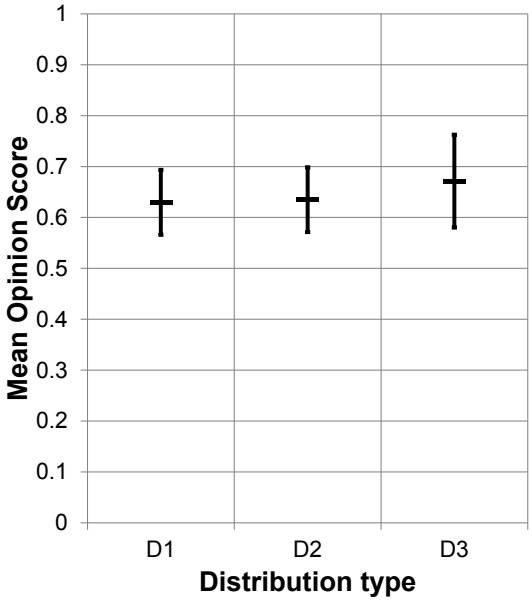


Figure 4.33: LF+: Mean Opinion Scores of the subjective test on stalling distribution.

with the distribution of $2 \times 250\,\text{ms} + 500\,\text{ms} + 2 \times 1000\,\text{ms}$. *D2* also contained five stalling events, but the distribution was $2 \times 250\,\text{ms} + 2 \times 500\,\text{ms} + 1500\,\text{ms}$. Finally, *D3* only had two stalling events, evenly distributed as $2 \times 1500\,\text{ms}$. All three used the same five positions of stalling events — from which D3 only selected two — chosen based on their TI values, covering various extents of subsequent frame change.

These test conditions were applied to different 2D clips of "*Big Buck Bunny*", with varying stalling event orders (4 instances for *D1* and *D2*, and 2 for *D3*). The durations of the stimuli were 19–22 seconds. The subjectively perceived quality of the stimuli were to be rated on a digital continuous scale between 0 (worst quality) and 1 (best quality).

The tests were carried out with similar environmental conditions as the previous experiment. The display was the same Samsung 55-inch JU6400 6 Series Flat UHD/4K Smart LED TV as in the experiments presented in Chapter 2. However, since content resolution was only 1920×1080, the viewing distance was 3 H accordingly.

4.8.2.2.3 Results

A total of 16 individuals participated in the experiment (10 males and 6 females). The age range was from 18 to 37, and the average age was 30.

The results indicate the lack of any significant difference, as shown on Figure 4.33. In fact, 9 out of 16 test participants rated with average differences between the dis-

tribution types that are smaller than 0.1. From the remaining 7, 4 favored *D3* and 3 preferred the other two. Furthermore, the larger confidence interval of *D3* originates from the smaller number of associated tests (as stated earlier), and not from a higher deviation of scores.

Due to the balanced user preference regarding stalling distribution, it was found sufficient to investigate a single-event scenario in the evaluation experiment. Excluding different stalling event distribution patterns from the study enabled a smaller test condition matrix, granting better focus on the research question addressed in the following section.

4.8.3 Evaluation

4.8.3.1 Research Aim

The aim of the research was to evaluate the concept of the dynamic adaptive streaming of light field video [194].

4.8.3.2 Test Conditions

In order to evaluate the concept of dynamic adaptive light field streaming, video stimuli were created with quality switches and stalling events; the variables in the videos were spatial resolution, angular resolution, and stalling duration. There were four types of test conditions for each source content: (a) quality switching with spatial resolution reduction, (b) quality switching with angular resolution reduction, (c) quality switching with spatial and angular resolution reduction, and (d) stalling event. Each variable in these conditions had two parameters in the test. As the goal of this work was to analyze the effects of stalling and reductions in resolutions on the perceived quality in a controlled manner, instead of considering an actual adaptation strategy based on a real bandwidth model, an ad-hoc approach was chosen, and the following fixed parameters were used.

The stalling event was either 500 ms or 1500 ms long; these are typical values in the related research [95,99,195,196]. The duration of 500 ms was chosen as it is above the threshold of being a JND, yet it can be easily tolerated. 1500 ms, on the other hand, is much more difficult to tolerate and can be considered as a significant stalling duration. The stalling events in this study were implemented as frame repetitions (or frame freezing) without graphic indicators, similarly to the research presented on HDR in Chapter 3, and the ones introduced earlier in this section.

Table 4.3: Investigated spatial resolutions in quality switching.

Source ID	High resolution	Low resolution
Red	1024×768	640×480
Yellow	1024×768	800×600
Ivy	960×540	640×360
Tesco	1280×720	640×360
Gears	1920×1080	640×360

Table 4.4: Investigated test condition pairs.

Test ID	Quality switching	Stalling duration
A	Spatial	500 ms
B	Spatial	1500 ms
C	Angular	500 ms
D	Angular	1500 ms
E	Both	500 ms
F	Both	1500 ms

Video stimuli were created with stalling events in this manner as this can be considered to be one of the most common approaches for Video on Demand (VoD) services when quality switching is not implemented; the last visualized frame freezes on the screen until playback continues. The only major deviation from the general practice is the lack of a typical spinning, circular graphical indicator. Its implementation would have been possible, either in 2D or 3D; however, it can be visually distracting for the observers, and it would have made stalling duration very explicit, disregarding certain features of the content.

Content angular resolution was either 1 or 2 source views per degree (1 or 0.5 degrees). The higher value was chosen because, according to previous studies [124, 175], it can provide a smooth horizontal motion parallax, and thus, a good user experience (at least in that regard). The lower value was chosen for multiple reasons. First of all, the work presented on interdependence earlier in this chapter [182] pointed out the potential gain at this content angular resolution, that might compensate the overall QoE during quality switching with both resolutions. Furthermore, prior researches [146, 174] argue whether this can be considered as a sufficiently high value for light field visualization or not, making it more of scientific interest.

Quality switching regarding content spatial resolution varied for each source video

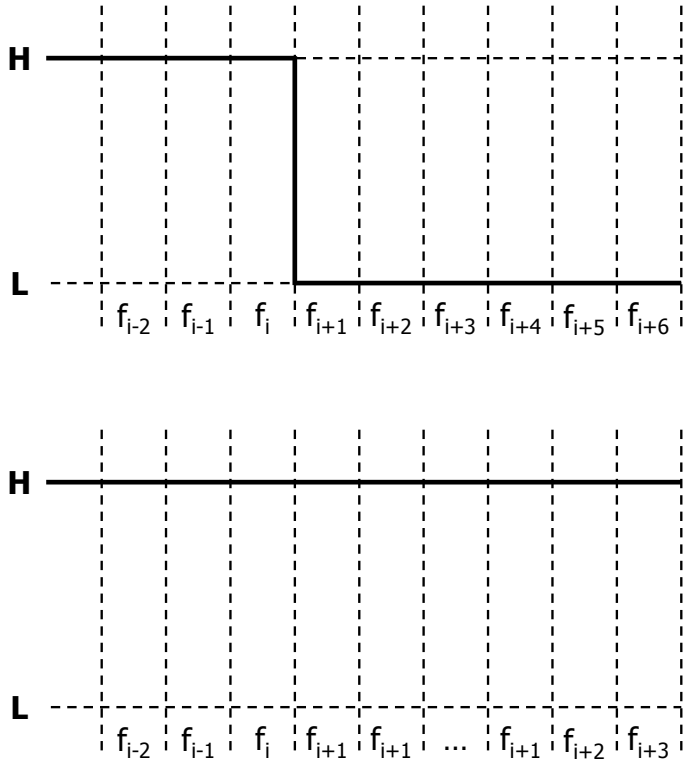


Figure 4.34: LF: The implementation of quality switching (top) and stalling (bottom). H and L are the high- and low-quality representations, respectively, and f represents the frames of the video. For stalling, the length of the event is determined by the number of the repeated f_{i+1} frames.

(see Table 4.3), using two different aspect ratios. The choice of the spatial resolution was based on the prior findings in the area [173], and the configuration was the same as during the research on viewing conditions and video resolution (see Table 4.1).

It is vital to point out that this work was focused on "down-switching", so quality switching was always performed as a sudden change from higher to lower spatial and/or angular resolution. Also, each and every stimulus only contained either one quality switching or a stalling event, and the frame of quality switching and stalling was always at the middle of the content.

Figure 4.34 demonstrates the implementation of quality switching and stalling in the experiment. For each source video, f_i represents the frame at the middle of the video (at the half of the duration). As either quality switching or stalling was applied only from f_{i+1}, the frames of the stimuli between f_1 and f_i were identical, with quality H (high quality) for both angular and spatial resolution. In case of quality switching,

if we denote the final frame of a given stimulus as f_n, then from frame f_{i+1} to f_n the stimulus was shown in quality L (low quality), where L either had a reduced spatial resolution, angular resolution or both. As stalling was implemented as frame repetition, f_{i+1} was repeated according to the different stalling durations, and then was followed by f_{i+2} and the rest of the frames until f_n. This means that no frames were skipped, and stalling increased the total duration of the stimulus.

As there were 5 source videos and 5 test conditions with 3 quality switchings and 2 stalling events, the total number of video stimuli was 25. The subjective test was a paired comparison, in which the different quality switching types were compared to the stalling events. Thus, there were 6 comparisons, which were applied to the 5 sources, so the total number of comparisons was 30. This also means that a test participant viewed 60 video stimuli during the test. Furthermore, this is one of the reasons why the stimulus duration was limited: using longer source videos would have resulted in a prolonged total test duration, as every extra second in average source length would have meant 1 minute more for the total duration. The other reason is that the experiment was centered on the quality switching event itself, and one stimulus only contained a single switching event. Similar video stimulus durations were used in the work of Duanmu *et al.* [197], addressing streaming QoE.

The evaluation was performed on a 5-point comparison scale ("*Much worse* (-2)", "*Worse* (-1)", "*Same* (0)", "*Better* $(+1)$", "*Much better* $(+2)$"). The task of the test participants was to compare the second stimulus in the pair to the first one. The comparison task targeted the overall QoE, taking every aspect of perceived quality into consideration. The notion of QoE was, of course, limited to the visual experience, as the test stimuli contained no audio, similarly to every other work on video quality in this thesis. The test pairs and also the videos inside the pairs were separated by a 5-second blank screen[16].

4.8.3.3 Results

A total of 20 individuals participated in the experiment (16 males and 4 females). The age range was from 20 to 38, and the average age was 26.

As detailed in the previous section, every comparison pair contained one video stimulus with quality switching and one with stalling. In the analysis of this section, positive values indicate the preference of stalling, and negative values indicate the preference of quality switching (e.g., a score of -2 means that a given stimulus with

[16]Rec. BT.2021: Subjective methods for the assessment of stereoscopic 3DTV systems

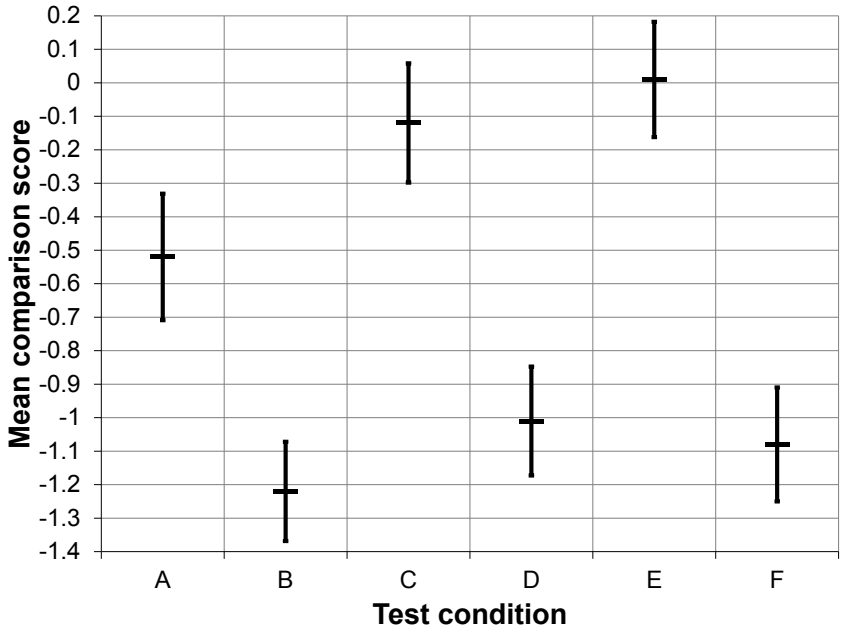


Figure 4.35: LF: Mean comparison scores of the evaluation of dynamic adaptive light field streaming.

quality switching was rated "*Much better*" compared to the stimulus it was paired with, containing a stalling event).

Figure 4.35 shows the mean comparison scores obtained for the test conditions defined in Table 4.4, and Figure 4.36 provides the distribution of the scores. When the quality switches were compared to a high stalling duration of 1500 ms (conditions B, D, and F), typically the quality switches were preferred; the preference of stalling was 6% or less in all three conditions. Yet it needs to be noted that in this research angular resolution was only reduced from 0.5 to 1 degree, which can be considered a borderline of toleration [174, 175, 177]. It is expected that further, larger extents of reductions in angular resolution could easily reverse this ratio, as users would rather wait 1500 ms than face severe visual degradations (e.g., see Figure 4.16).

Generally, it can be stated that the quality switching based purely on spatial resolution (conditions A and B) performed better than the other two types, as it is reflected in both mean scores and scoring distribution. However, from a statistical point of view, there is no significant difference between the scores of conditions B, D, and F.

The same does not apply to the comparisons with a low 500 ms stalling (conditions A, C, and E). While C and E were rather balanced in preference (see Figure 4.36), in condition A, quality switching was preferred by 59% of the test participants and

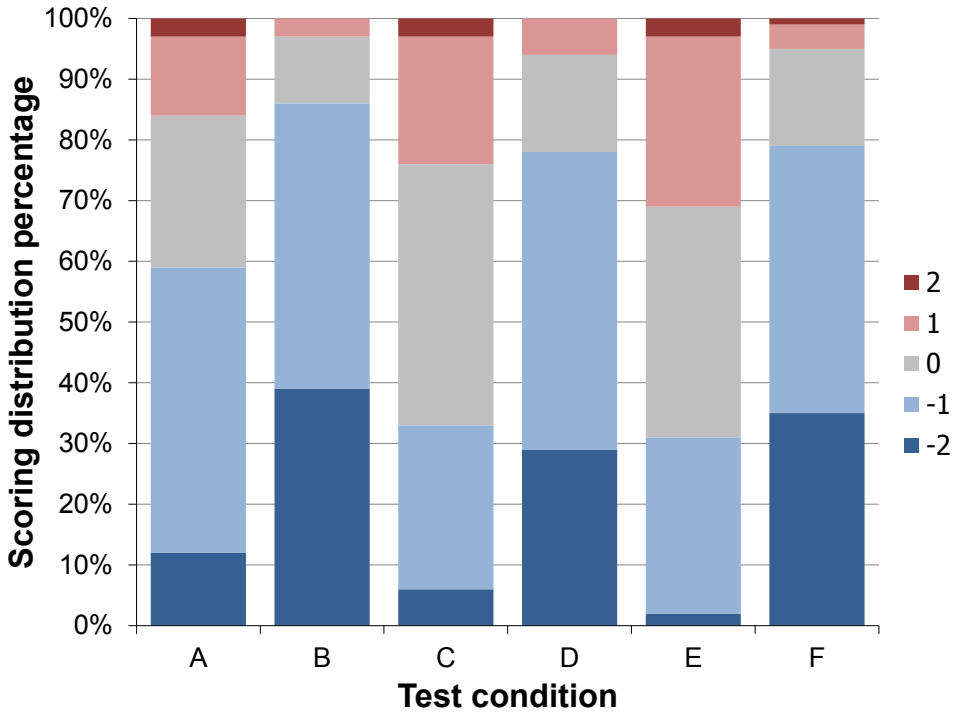


Figure 4.36: LF: Distribution of comparison scores per test condition.

stalling was chosen by 16%.

The most important finding in these results is the negligible difference between quality switching using angular resolution reduction only (C and E) and using both angular and spatial resolution (D and F). If the difference between them regarding perceived quality is such, then the combined switching can be used in a practical application of the protocol, which can come with a major reduction in bandwidth requirement compared to angular switching, without compromising user experience.

As an example for bandwidth requirement reduction, let us take the data sizes of *Red*, *Yellow*, *Ivy*, *Tesco*, and *Gears* at full spatial and angular resolution. Decreasing spatial and/or angular resolution by selecting lower-quality segments evidently reduces the transmission data rate, as the segment sizes (in bytes) are smaller. Table 4.5 shows how this applies to the source video stimuli of the subjective test, e.g., the size of *Gears* at combined low resolutions is 5.5% of the size at full spatial and angular resolution (corresponding to a compression ratio of approximately 18:1). Note that these short video sequences, not even reaching 15 seconds in duration, were the size of 5–6 GB as high-quality source contents (i.e., in their original high spatial and angular resolutions).

As investigated in the work on interdependence, reducing spatial resolution when

Table 4.5: Data rate reduction through lowered spatial and/or angular resolution for each video content. 25% in the table means that only a quarter of the corresponding high-resolution data was required for light field visualization.

Source ID	**Spatial**	**Angular**	**Both**
Red	39%	50%	19.5%
Yellow	61%	50%	30.5%
Ivy	44.4%	50%	22.2%
Tesco	25%	50%	12.5%
Gears	11%	50%	5.5%

Figure 4.37: LF: A part of *Gears* before (left) and after (right) a quality switch, reducing both spatial and angular resolution.

angular resolution does not provide a continuous horizontal motion parallax with undisturbed smoothness can lessen the impact of visual phenomena such as the crosstalk effect. Figure 4.37 compares frame f_i and f_{i+1} (i.e., before and after quality switching) from the sequence *Gears*, where the quality switch included both resolutions. Although certain levels of the crosstalk effect and ghosting were visible, the blur induced by the lowered spatial resolution applied to the entire scene, including the visual degradations that disturbed the parallax effect. This was able to mask the insufficient angular resolution to a given extent. Therefore, the blur reduced the effect of such visual phenomena from a perceptual perspective.

The selection of the source video contents used in the experiment did not have a significant impact on the obtained results; no statistically significant difference was found between any two of the contents. There are, of course, certain extents of differences, visualized in the distribution of the scores per video content (see Figure 4.38). These primarily originate from condition *A* and *B*, which compared quality switching with spatial resolution reduction to stalling. Conditions *C*, *D*, *E*, and *F* do not deviate much per source; in fact, some even have the exact same mean values. The only exception is *Tesco*, which obtained fewer ratings favoring quality switching, due to the high mobility of the scene.

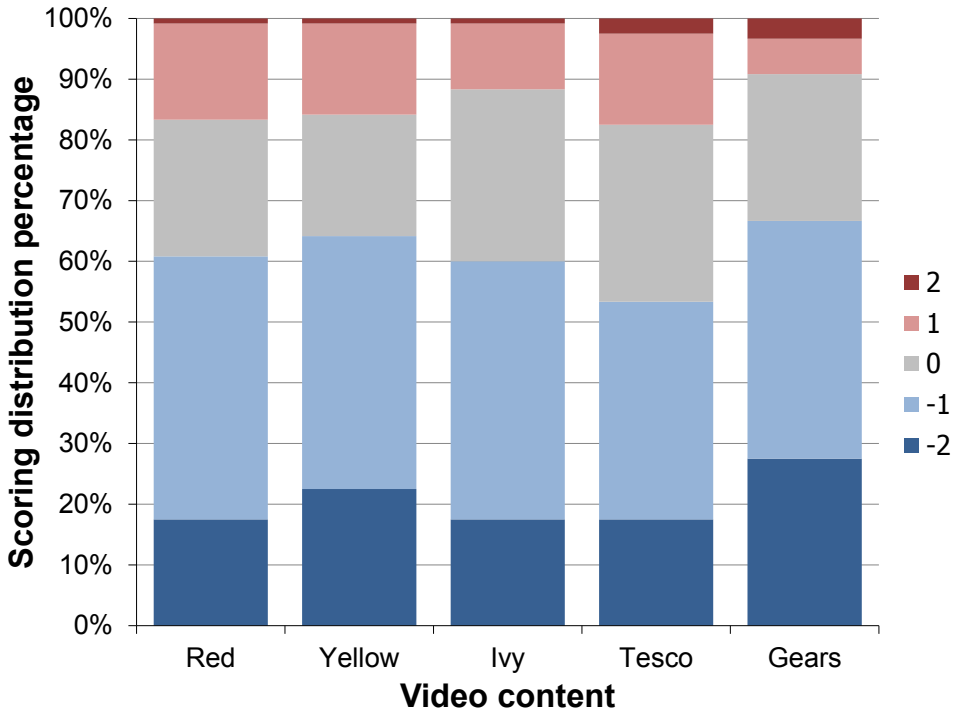


Figure 4.38: LF: Distribution of comparison scores per video content.

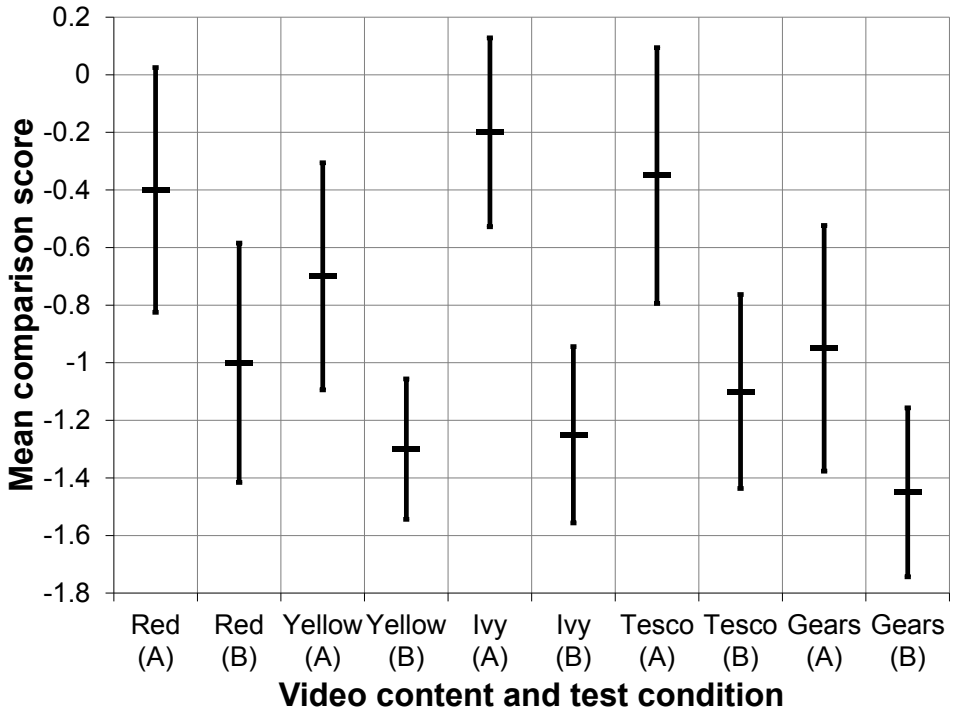


Figure 4.39: LF: Mean scores of test conditions A and B per content.

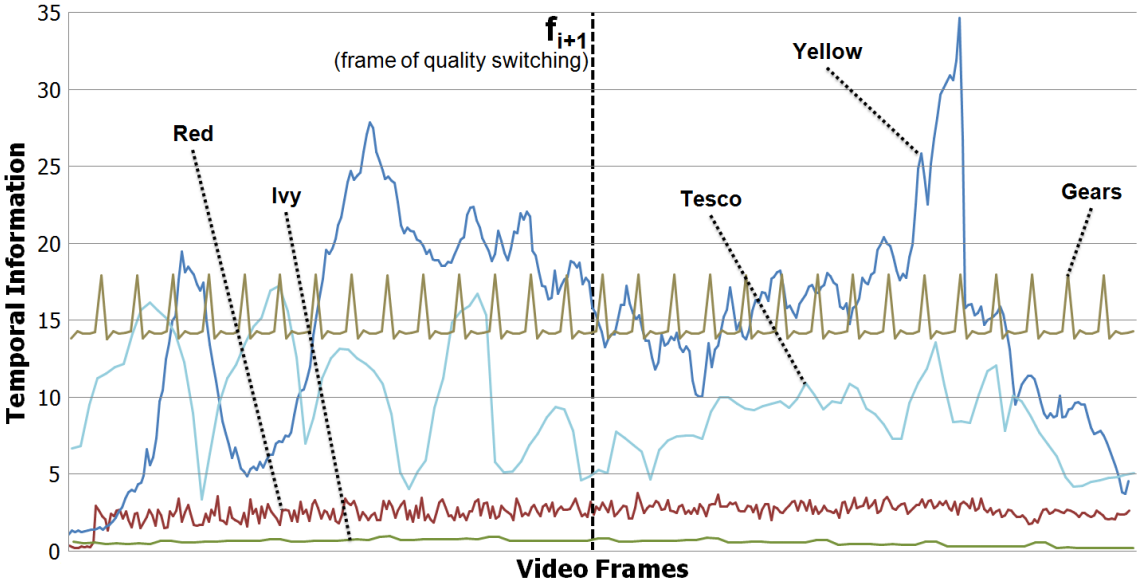


Figure 4.40: LF: Temporal Information of the source video stimuli.

As described in Table 4.3, the five contents switched between different spatial resolutions. Based on the selected values, the descending order of switching magnitude (percentage of difference in source pixel count, see Table 4.5) was *Gears*, *Tesco*, *Red*, *Ivy* and *Yellow*. Theoretically, this would imply that spatial resolution reduction affected *Gears* the most on the level of perception, and contents such as *Ivy* and *Yellow* were less affected. However, the results report the opposite (see Figure 4.39).

Although there is no statistically significant difference between the source contents in the analysis focusing on A and B, the relations between the mean comparison scores are quite noteworthy; for both A and B, *Gears* was the least affected, while *Ivy* received comparably low preference scores for quality switching, particularly for A. The results indicate that the content itself had a greater influence than the change in source spatial resolution. Even though *Ivy* was limited to a subtle animation, the additional blur due to quality switching degraded the visual appearance of the statue and the detailed ivy plant growing around it. On the other hand, in case of *Gears*, this quality transition only softened the edges of the rotating gears, even though the source spatial resolution was reduced from 1920×1080 to the same 640×360 as *Ivy* (see Table 4.3). As for *Red* and *Yellow*, their comparison scores were more in alignment with the difference in source spatial resolutions, for both A and B.

Figure 4.40 shows the application of the conventional 2D TI measurement to the middle source camera view. The x-axis of the figure indicates the passage of time, and as the source videos were different in duration, the TI figures are temporally

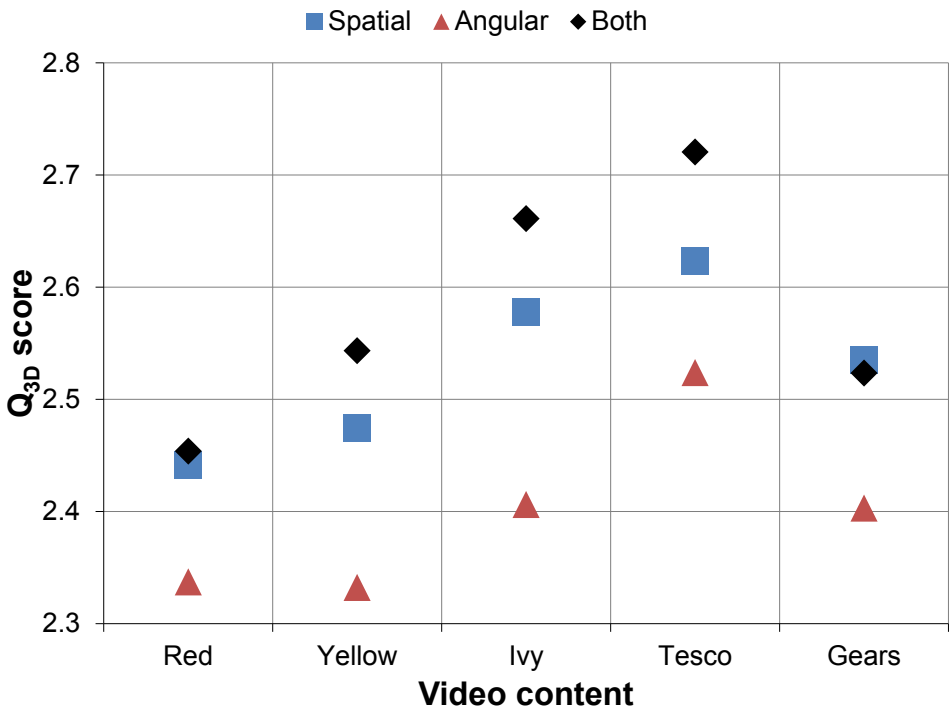


Figure 4.41: LF: Q_{3D} scores at frame f_{i+1} per video content and $\alpha = 0.89$. Higher objective scores suggest higher levels of QoE degradation.

stretched to have the same f_{i-1} middle frame where the quality switching took place. The subtle animation of *Ivy* barely registered in the measurement, while the multiple rotating columns of *Yellow* resulted in high levels of TI. As *Gears* was a short looping animation, this is well reflected in the repeating TI pattern. However, such application of TI cannot measure motions and alterations along the z-axis (depth); TI applied to *Red* mainly measured the shadows cast by the columns that moved closer to the observer, instead of the actual movement. Therefore, in future works, there should be an aim to develop a TI metric for discrete light field contents (camera view arrays), for accurate content classification. Such knowledge regarding the content is particularly important, as the utilization of the depth budget can fundamentally affect perceptual sensitivity towards the parallax effect, and thus, the requirements for angular resolution.

The obtained subjective test results were also compared to the full reference (FR) objective quality metric proposed by Tamboli *et al.* [145], which was selected due to its consideration for the angular quality component. The Q_{3D} values of the metric are calculated as:

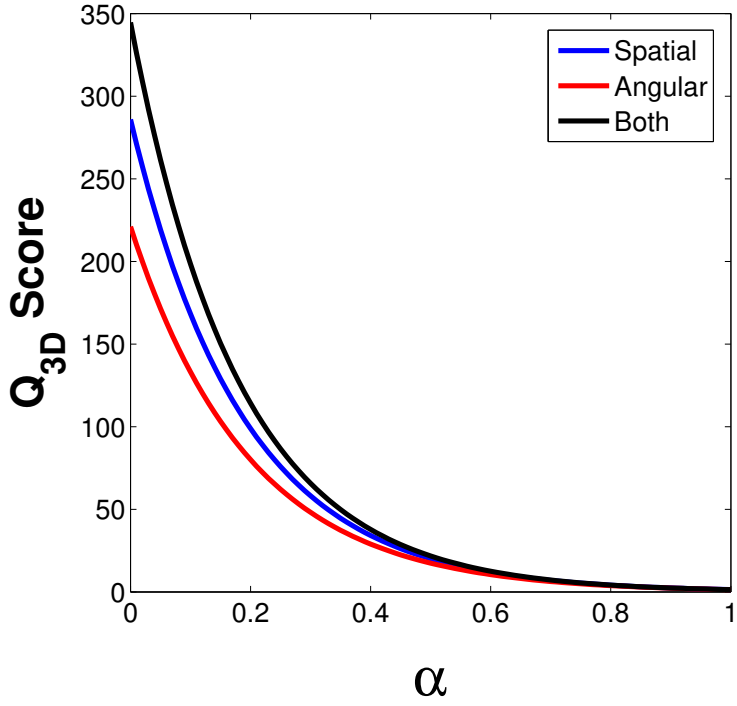


Figure 4.42: LF: Q_{3D} scores of *Gears* at frame f_{i+1}. Higher objective scores suggest higher levels of QoE degradation.

$$Q_{3D} = Q_{2D} \cdot (\frac{Q_\theta}{Q_{2D}})^\alpha \tag{4.2}$$

where Q_{2D} is the spatial component, which is based on the transformation of images into a parameter space and their comparison in that space, using a steerable pyramid decomposition; Q_θ is the angular component, calculating MS-SSIM for optical flow vectors; and α is a parameter whose value is between 0 ($Q_{3D} = Q_{2D}$) and 1 ($Q_{3D} = Q_\theta$).

The metric was applied to frame f_{i+1} (see Figure 4.34) of each stimulus with a quality switching event, and the reference was the corresponding frame with high angular and spatial resolution. For the objective evaluation, the α value of 0.89 was used as set by Tamboli *et al.*, which was determined via a 1000-fold cross-validation, based on their subjective quality assessment scores.

The objective Q_{3D} scores (see Figure 4.41) fit into an interval of 0.4 (the difference between the highest and lowest value was 0.39). This indicates small differences between the degradations of the contents, as this metric in practice can provide Q_{3D} scores between 5 and 10 for distorted content at an α of 0.89. However, in this

experiment, the contents were not degraded visually (e.g., via added noise), but only by reducing spatial and/or angular resolution. The Q_θ value, which runs between 0 (most extreme extent of angular distortion) and$\sqrt{2}$ (no measurable angular distortion) deviated the most in case of *Gears* (1.37), due to the sharp edges in the scene. Its Q_{3D} scores based on the possible α values are shown on Figure 4.42.

According to these objective results, shifting down from the angular resolution of 2 views per degree to 1 view per degree had a lesser impact on the estimated scores than the changes in spatial resolution, for all video contents (see Figures 4.41 and 4.42). However, this is in contradiction with the obtained subjective results (see Figure 4.35). Yet it needs to be noted that the authors of the metric strictly used reference visual data with an angular resolution of 1 view per degree during every process of creation, including calibration and evaluation.

Furthermore, as the objective metric was designed for still content, it is difficult to efficiently apply it to light field video frames, as such frames might not be as clear and sharp as a static scene, due to the motion and changes in the content [170]. In such case, motion blur and other visual degradations affect the reference of the FR metric, also modifying the quality of the image (video frame) to which spatial and/or angular resolution reduction is applied. For accurate QoE estimation, light field video metrics would be necessary. However, at the time of this thesis, no objective quality metric for light field video exists, as video, in general, for this visualization technology is currently under-investigated.

Regarding the subjective test results, one could argue that the additional cognitive load via the "wow effect" evoked by the novel visuals of light field technology could have biased the perceived stalling duration. As it was presented in the previous chapter, the stunning new visuals of such displays may draw away the attention of the test participants from stalling events. These results were taken into consideration during the design of the study, and therefore, during the training phase, the fact of the presence of the stalling event was emphasized for the test participants, along with its location within the video.

4.9 Chapter Summary

The chapter introduced ten studies on light field QoE. Whenever angular resolution was a test variable among the conditions, the angular resolution of 1 degree was always included (except in the study on view synthesis, where 1.5 degrees was a more suitable base of comparison due to the higher level of visual degradation). Across

all the obtained results, for both still contents and video sequences, a considerable variation can be found in the subjective ratings of such visual stimuli. There is an apparent dependency on the content, but observer movement also had a significant impact on the perceived quality. Perhaps the best term to approach this angular resolution is "adequate"; an angular resolution of 1 degree may be sufficient for many contents and use case scenarios, but it rarely achieves visual excellence, especially if the viewers observe the light field display with unconstrained movement within the FOV. In the first subjective test on angular resolution, it received poor ratings, while the study on static observers not only performed better overall, but almost all test participants found one of the laser-scanned statues to be acceptable at 1 degree. Yet an important part of the truth is that almost none of them found one of the complex mathematical bodies acceptable at the exact same resolution, signifying the content-dependent nature of the issue. 1 degree was also the angular resolution where resolution interdependency was the most notable; the reduction of spatial resolution at this angular resolution managed to improve the perceived smoothness of the parallax effect. This applied to the related experiments on both still content and video.

In the end, these findings were integrated into the concept of quality switching, relying also on the separately investigated tolerance towards spatial resolution reduction. The subjective test on concept evaluation was the first ever to address the dynamic adaptive streaming of light field video. Due to technological limitations (e.g., bandwidth and processing power), no such service exists at the time of this thesis; in fact, it may take several more years for the first light field service to emerge in the commercial world, and the industry also needs to wait patiently for the common utilization of real-time applications of light field technology. Yet there are many research directions along which these scientific efforts could be continued. The possible future work related to the experiments presented in the three main chapters is introduced in the following chapter, concluding the thesis.

Chapter 5

Conclusions

The thesis presented the results of a series of subjective tests that were carried out to investigate the QoE of three emerging display technologies. Two of them, namely UHD and HDR visualization, were approached from the angle of cognitive bias via the labeling effect, and the third one, light field visualization, was first addressed on a more fundamental level, and then these results were used to propose something more complex with novelty in the scientific area.

The obtained results from the tests on UHD video indicate that the labeling effect had a significant impact on the subjective scores, regardless of test condition and source content. It needs to be particularly noted how similar the results were for test conditions baring the same labels.

The corresponding study without labels concludes the lack of statistical differences between the two video resolutions, regardless of rating scale. Both subjective studies used a close viewing distance of 110 cm, which is standard 1.6 H for the given UHD display, showing UHD content on its entire screen. The only aspect of the experimental configuration that may be criticized is the choice of content, as more common UHD contents were used instead of typical "eye candy" videos. However, this work argues that studies should aim to have a greater focus on source content types and quality levels that the average user may meet during daily multimedia consumption.

The choice of rating scale greatly affected the test with the labels, as the more fine-grained 7-point comparison scale enabled the expression of the slighter perceived differences, in contrast to the 3-point scale. On the one hand, a 3-point scale may avoid the inclusion of some biased results in the outcome of the research, as the cognition of preconception can only override genuine perception to a certain degree. On the other hand, the use of such scale may also result in the loss of subjective feedback that reports slight differences, and in many cases, these slight differences are of great research interest.

The thesis also introduced four experiments on HDR video QoE. The first one addressed different quality aspects, and investigated their cognitive distortions caused by the labeling effect. The obtained subjective scores indicate that for aspects like luminance, color, and image quality, the positive label "Premium HDR" resulted in a positive bias, but for frame rate — which was more difficult to directly connect to HDR visualization — the rating patterns were not obvious. It was found that several test participants approached frame rate as an aspect which generally suffers degradations due to a trade-off between visuals and frame rate.

The second and the third experiment focused on stalling event detection and tolerance for HDR and LDR visualization, respectively. The comparison of the results showed that even 500 ms stalling events may go unnoticed due to the presence of the so-called "wow effect" and "visual awe" that comes with HDR visualization. The studies indicate statistically significant differences between the evaluation of the LDR and HDR sequences.

The fourth experiment investigated the perceived duration of stalling events, in a scenario similar to the first study; the test participants were influenced by the label "Premium HDR". The findings clearly suggest the presence of the prior idea; the preconception, which — similarly to the first experiment — builds on the trade-off between visuals and other quality aspects that do not contribute to the appearance of HDR visualization. The results indicate that the stalling events in "Premium HDR" videos were perceived to be longer. This applied to stalling events with 500 ms and 1000 ms duration as well, but the latter suffered significantly more cognitive bias.

The third and final display technology addressed by this thesis was light field. The fundamental research on light field visualization found that display and content FOV of 135 degrees is definitely adequate for use cases with unlimited observer movement; that the loss in content spatial resolution can be highly tolerated; that the acceptance of an angular resolution of 1 degree is highly debatable; that subjective tests with disturbed parallax smoothness require additional training in order to support rating consistency; that light field reconstruction may be visually unfavorable but may also enhance the visual experience through increased content contrast; and that view interpolation may significantly improve the QoE, provided that the selected technique has a sufficiently dense view set to work with.

The subjective studies also investigated the interdependence between spatial and angular resolution, coming to a conclusion that the blur created by the reduction of spatial resolution may benefit the smoothness of the horizontal motion parallax, in

case it is already disturbed via insufficient angular resolution. The results of the experiments on viewing conditions indicate that test participants with sideways movement during observation may perceived a given stimulus to be significantly worse regarding angular continuity. One of these tests particularly highlighted the connection between content resolutions with regards to the movement of the test participants.

The last major section of the thesis is the concept and evaluation of the dynamic adaptive streaming of light field video. The results indicate that quality switching in the domains of spatial and angular resolution is clearly favorable to 1500 ms stalling events, and that reduction in spatial resolution is preferred over 500 ms waiting times. The lack of difference between stimuli with low angular and low combined resolution reinforces the findings of the previous studies, and therefore, shows that the reduction of spatial resolution in a scenario of insufficient transmission rates is a beneficial decision, especially if the angular resolution is reduced as well.

5.1 Future Work

There are numerous ways to continue the research presented in this thesis. Regarding the work on UHD visualization, the study using explicit labels could be repeated with implicit labels, and then, of course, a post-experiment questionnaire could be used to record whether the test participants considered the label or not. Longer sequences could be included in the tests, not only to investigate QoE over time but also to address the fading of the labeling effect in studies with labels that are either frequently presented or only once at the beginning of the test. Compression could be involved in order to study its influence on the subjective ratings and to enhance the realism of the experimental context. The role of source content could also be addressed in more detail, using videos ranging from low-quality user-generated content to exceptional demo materials. Furthermore, user decisions could be involved in order to investigate the effect of post-decision dissonance on the perceived quality. Additionally, the experiments presented in Chapter 2 and all their potential continuations could be performed using 4K and 8K instead of HD and 4K.

As for the future work on HDR visualization, first of all, the second and the third study could be repeated with several other stalling event durations and patterns (i.e., varying stalling frequencies within a stimulus with different durations), particularly targeting short events around the level of JND and longer durations beyond 1000 ms. These tests could also utilize equipment to track the eye movement of the test participants, and thus, record where they look at the time of the stalling event. Regarding

quality aspects, the assessment of frame rate and stalling event duration could be simultaneously integrated into an experiment with the presence of the labeling effect. Repeating the four studies with a frame rate of 60 fps and UHD resolution could also provide valuable scientific insight, along with 10-bit and 12-bit videos, in order to address the significance of bit depth in such studies. Furthermore, similarly to the suggested continuation of the work on UHD, the fading of cognitive bias over time could also be investigated, with various test methodologies and significantly longer video sequences. The experiments on both UHD and HDR could involve various physiological measurements as well, in order to quantitatively measure cognitive load during quality assessment tasks.

Regarding the continuation of the work on light field visualization, the research on the QoE of light field video streaming could be extended with longer sequences and with different stalling distributions, including various frequencies and durations. A logical next step in research would be to use actual adaptation strategies based on real bandwidth models. An extended range of angular resolution variation could also be investigated, particularly around 1 degree. Furthermore, as technology progresses, super resolution should be addressed. Possibly the most intriguing research in the topic would be study angular resolution around the threshold of super resolution, as the ability of proper depth focusing is expected to be a real game changer. The industry could benefit from light field QoE studies that target the optimal combinations of spatial and angular resolution, as such results may highly influence future display and service development. Regarding displays, the majority of the presented experiments may be adapted to other front-projection and back-projection displays in a straightforward manner — as most of the test conditions were based on general KPIs of light field technology — and similar findings would be expected. Of course, it cannot be generalized to each and every form of light field visualization, as certain test conditions are not applicable to specific device types (e.g., wide-baseline parallax observation on near-eye gears). A rather intriguing research would be to have several test participants simultaneously observe the video stimuli, without any constraint on viewing conditions, besides being limited to the FOV of the display. Physiological measurements could be used to determine the connection between immersion and content resolution. Finally, a long-term goal could be to develop and test streaming solutions and schemes for cost-efficient real-time transmission of light field video.

Appendix A

Subjective Test Results

This appendix introduces the collected raw subjective assessment data from notable experiments, one from each main chapter. Each column represents the ratings of specific test participant. A maximum of 15 columns are provided on a single page. In experiments where there were more than 15 test participants, the test conditions (marked on the left) are repeated per cluster; e.g., as 36 test individuals participated in the selected HDR test, there are 3 iterations presented in total $(15 + 15 + 6)$.

A.1 UHD: Tests with Labels, 7-point Scale

Condition 1

SRC01	1	3	0	1	2	1	1	3	2	2	0	1	2	1	1
SRC02	3	2	0	1	2	3	1	3	3	2	1	1	1	0	-1
SRC03	0	2	0	1	2	2	1	3	1	1	-1	1	3	2	2
SRC04	2	2	0	0	1	1	1	3	-2	2	1	1	3	3	-1
SRC05	-3	3	0	1	1	2	1	3	0	0	0	1	2	2	-1
SRC06	1	2	0	2	2	-1	1	3	-2	-1	1	1	2	0	1
SRC07	1	2	0	1	2	1	1	3	1	-1	1	1	1	1	1
SRC08	0	2	0	2	2	3	1	3	2	1	-1	1	1	0	1

Condition 2

SRC01	1	-2	1	-1	-1	0	-1	-3	1	-1	-1	-1	-1	-1	1
SRC02	1	-2	0	1	-1	1	-1	-3	2	-1	1	-1	-1	-1	-1
SRC03	-1	-2	-1	1	-1	-1	-1	-3	0	1	-1	-1	-3	-1	-1
SRC04	0	-2	0	0	-1	0	-1	-3	3	1	-1	0	-3	-3	-2
SRC05	1	-2	0	-1	0	-1	-1	-3	1	-2	-1	0	-1	0	-1
SRC06	0	-3	0	1	0	-2	-1	-3	1	-1	0	-1	-2	0	-2
SRC07	3	-2	0	-1	0	-2	-1	-3	0	-2	-1	-1	-2	0	-1
SRC08	2	-2	-1	-1	-1	-3	-1	-3	1	0	-1	0	-1	0	1

Condition 3

SRC01	1	3	0	2	2	1	1	3	3	-2	-1	0	2	0	-2
SRC02	0	2	0	1	2	3	1	3	-1	1	1	0	2	1	2
SRC03	-1	2	0	-1	2	-1	1	3	1	-2	1	1	3	2	1
SRC04	2	3	1	1	-1	0	1	3	-1	1	0	1	3	3	1
SRC05	-1	2	0	1	2	1	1	3	1	2	1	1	3	1	1
SRC06	-2	2	0	2	1	2	1	3	2	-2	-1	0	2	1	-1
SRC07	2	2	0	0	1	0	1	3	2	1	0	1	2	0	-1
SRC08	0	2	0	3	3	0	1	3	0	-1	1	1	1	1	-1

Condition 4

SRC01	-1	3	0	1	2	0	1	3	-1	-1	0	1	2	1	1
SRC02	1	2	0	1	2	2	1	3	0	2	0	1	1	1	2
SRC03	1	3	0	0	1	2	1	3	0	2	-1	1	2	2	2
SRC04	0	2	0	0	2	0	1	3	-1	2	0	1	2	3	-2
SRC05	-1	2	-1	0	2	1	1	3	-1	0	1	0	2	1	-1
SRC06	0	2	0	-2	1	0	1	3	1	1	1	1	1	1	1
SRC07	1	2	0	0	1	-2	1	3	-1	2	0	1	1	0	-2
SRC08	3	2	0	2	2	2	1	3	-1	1	-1	1	1	1	-2

Condition 5

SRC01	0	-2	0	0	-1	1	-1	-3	1	0	-1	-1	-1	-1	-1
SRC02	1	-1	0	0	-1	3	-1	-3	2	2	0	0	-1	0	-2
SRC03	2	-1	0	-1	-1	-1	-1	-3	1	-1	1	-1	-2	-2	1
SRC04	1	-2	2	0	-1	0	-1	-3	1	-3	0	0	-1	-3	1
SRC05	-1	-1	0	-1	-1	-1	-1	-3	0	1	0	-1	-1	0	2
SRC06	1	-3	0	0	1	0	-2	-3	0	1	0	-2	-3	0	1
SRC07	0	-2	0	-2	-1	-1	-1	-3	0	0	-1	-1	-2	-1	1
SRC08	3	-2	0	0	-1	-2	-1	-3	0	1	-1	0	-1	-1	1

Condition 6

SRC01	1	-2	0	0	-1	2	-1	-3	-2	0	-1	0	-1	-1	-1
SRC02	2	-3	-1	-1	2	0	-1	-3	2	0	0	0	-1	-1	1
SRC03	1	-2	0	-1	-1	-1	-1	-3	2	1	1	0	-2	-2	1
SRC04	0	-1	0	0	-1	0	-1	-3	0	1	-1	0	-2	-3	-1
SRC05	1	-1	0	0	-1	-1	-1	-3	0	-1	1	-1	-2	0	-2
SRC06	0	-3	0	1	-1	2	-1	-3	0	2	-1	0	-1	-2	1
SRC07	0	-3	0	-1	-1	-1	-1	-3	-1	-2	0	-1	-2	0	2
SRC08	-1	-2	0	-1	0	1	-1	-3	1	3	-1	0	-1	-1	2

Condition 7

SRC01	2	0	-1	-1	2	3	0	2	0	1	-1	0	0	0	-2
SRC02	2	1	1	0	2	-1	0	2	-1	0	0	0	0	0	-2
SRC03	-2	1	0	1	2	2	0	2	0	-2	1	0	0	0	-2
SRC04	-1	1	0	0	1	0	0	1	-2	-2	-1	0	0	0	-2
SRC05	-1	0	0	-1	1	1	0	2	0	0	0	0	0	0	-1
SRC06	2	1	0	0	0	-1	0	1	0	-1	1	0	0	0	-1
SRC07	1	0	0	1	2	0	0	2	2	2	0	0	0	0	2
SRC08	2	1	0	0	0	3	0	2	0	2	0	0	0	0	2

Condition 8

SRC01	1	0	1	0	2	2	0	1	-1	3	-1	0	0	0	-1
SRC02	1	0	0	0	2	1	0	-2	-2	1	0	0	0	0	-2
SRC03	-1	0	0	1	-1	0	0	-1	1	-1	0	0	0	0	1
SRC04	1	1	0	0	-1	3	0	-1	-1	2	-1	0	0	0	2
SRC05	-1	1	0	-1	0	2	0	-1	0	2	-1	0	0	0	1
SRC06	0	-1	0	-1	-1	1	0	-2	0	-2	-2	0	0	0	1
SRC07	1	0	1	-1	2	-1	0	2	-2	-2	0	0	0	0	-1
SRC08	3	0	-2	1	-1	0	0	-1	2	1	1	0	0	0	1

A.2 HDR: Tests on Stalling Duration

1AS	1	1	1	-1	0	-1	1	0	1	-1	2	1	0	1	0
1CS	-1	1	0	0	-1	0	0	1	0	-1	2	1	1	1	-1
1AL	-2	-2	-1	0	2	-1	0	1	2	0	3	1	2	2	1
1CL	1	2	-1	1	-2	0	2	1	1	0	3	1	2	2	1
2AS	0	1	1	1	-2	1	-1	0	0	1	3	1	2	1	1
2BS	1	1	2	1	-1	0	0	0	0	0	1	1	0	0	-1
2AL	-2	2	1	2	-2	2	1	2	1	-1	3	1	3	2	1
2BL	1	1	1	1	-1	0	0	1	1	1	3	1	1	1	0
3AS	0	1	0	0	0	1	1	1	0	0	3	1	2	1	1
3CS	-1	1	0	0	1	2	0	-1	0	0	3	1	1	-1	0
3AL	2	1	1	0	-1	-1	-2	1	0	1	3	1	2	1	-1
3CL	0	1	0	0	-2	0	2	1	1	1	3	1	3	2	1
4BS	0	1	1	1	0	0	0	0	0	1	3	1	1	1	2
4CS	1	1	1	0	-2	1	-1	-1	-1	0	2	1	-1	1	-1
4BL	2	2	2	1	-1	-2	2	2	1	1	3	1	2	2	1
4CL	-1	1	1	0	-3	0	2	1	0	1	2	1	1	1	1
5AS	1	1	0	1	-2	0	0	-1	0	1	2	1	1	0	1
5CS	-2	1	1	0	-1	-1	1	0	1	0	2	1	2	1	1
5AL	-1	1	1	1	0	2	-1	0	2	1	2	1	1	2	0
5CL	1	2	1	1	1	-1	-2	2	0	1	3	1	1	2	1

6BS	0	-1	0	1	-1	1	0	0	0	1	2	1	0	1	0
6CS	-1	-1	0	0	-2	0	1	0	0	0	3	1	-1	0	0
6BL	0	1	1	1	-2	-1	2	1	2	1	2	1	1	1	1
6CL	1	-1	0	1	-3	0	-1	0	1	0	2	1	1	0	0
7BS	2	1	0	0	-1	1	1	0	0	1	2	1	-1	1	1
7CS	0	1	1	0	0	0	-1	0	0	-1	2	1	0	1	-1
7BL	-1	1	2	0	-1	2	1	0	1	-1	3	1	2	2	1
7CL	-2	1	1	0	0	0	-2	1	1	1	3	1	1	1	0
8AS	1	1	0	1	0	1	1	0	0	-1	3	1	1	2	1
8CS	0	1	0	0	1	-1	-1	0	0	1	3	1	1	2	1
8AL	1	1	1	1	-2	-2	2	-1	0	-1	3	1	1	2	0
8CL	0	1	1	0	0	1	0	0	0	1	3	1	2	2	2
9AS	1	1	0	0	-1	1	0	-1	0	0	2	1	1	1	-1
9CS	0	1	1	0	-2	0	-1	0	0	-1	2	1	0	1	0
9AL	1	1	1	0	0	2	1	2	1	1	2	1	2	2	1
9CL	-1	1	0	0	-1	1	2	2	-1	1	2	1	1	1	1
10AS	1	1	1	0	0	0	0	0	0	0	1	1	0	-1	1
10BS	2	1	1	0	-1	-1	-1	0	0	0	3	1	1	1	1
10AL	-2	1	0	1	1	1	2	1	0	1	3	1	2	2	1
10BL	0	1	1	1	1	0	1	-1	1	1	3	1	3	2	2

1AS	1	1	2	1	1	1	1	1	-1	1	1	0	-1	1	-1
1CS	-1	1	-1	1	2	0	1	1	0	1	-1	0	2	0	1
1AL	2	2	2	1	2	2	2	2	3	-1	0	2	-2	-1	2
1CL	2	2	2	1	2	0	2	1	0	0	1	1	1	-2	1
2AS	-1	2	2	1	2	0	1	1	1	0	1	1	-2	-1	1
2BS	-2	1	-2	1	1	0	1	1	-1	0	0	0	0	-1	-1
2AL	1	3	2	1	3	0	2	1	1	-1	2	1	3	-3	1
2BL	1	2	1	1	2	0	2	2	0	0	1	-1	2	2	-1
3AS	1	2	2	1	2	0	1	1	0	0	0	1	2	1	-1
3CS	-2	1	2	1	1	0	0	1	0	0	-1	0	-2	-2	2
3AL	1	2	1	1	2	1	2	2	-1	1	2	1	2	-1	1
3CL	-1	2	3	1	2	0	2	2	1	1	-2	1	2	2	2
4BS	1	2	1	1	2	0	1	1	0	1	1	1	-1	2	0
4CS	2	1	-2	1	1	0	1	1	0	0	-1	0	1	0	0
4BL	2	3	3	1	3	1	2	2	2	1	0	1	2	-3	2
4CL	1	3	2	1	3	1	2	1	-1	0	-1	1	1	1	1
5AS	0	1	1	1	1	0	0	0	-1	1	0	1	1	0	0
5CS	-1	1	2	1	1	1	0	0	0	0	0	1	2	2	2
5AL	1	2	1	1	2	1	2	2	1	1	1	1	1	1	1
5CL	1	2	3	1	2	0	2	2	0	1	-1	1	3	-2	1

6BS	-1	2	2	1	1	0	1	0	0	0	0	1	1	2	-1
6CS	1	1	-1	1	1	-1	1	0	0	0	-2	1	0	-1	0
6BL	2	2	3	1	3	0	1	2	-1	0	-1	2	2	3	2
6CL	1	2	2	1	2	0	2	1	1	1	0	1	1	-2	-1
7BS	-1	2	1	1	1	-1	1	1	1	-1	0	1	1	0	0
7CS	1	2	0	1	1	0	0	0	0	0	1	0	-1	2	1
7BL	1	3	2	1	2	1	0	0	2	1	1	1	1	1	1
7CL	1	3	1	1	1	0	1	1	-1	-1	0	0	1	-1	2
8AS	1	2	1	1	1	0	1	1	0	0	-1	-1	1	2	2
8CS	1	1	1	1	1	0	0	1	-1	1	1	-1	-2	2	3
8AL	1	1	-1	1	1	0	1	2	0	-1	0	1	1	-1	-1
8CL	1	2	2	1	2	0	2	1	2	0	1	2	2	-3	2
9AS	0	1	1	1	1	0	1	1	1	0	0	1	-1	2	-1
9CS	0	1	0	1	1	1	0	0	0	0	1	-2	0	-1	-1
9AL	1	2	2	1	2	1	2	2	0	0	0	1	2	-3	2
9CL	1	2	1	1	2	0	2	2	1	0	-1	-1	1	1	1
10AS	1	2	1	1	1	0	1	1	0	0	0	-1	0	0	1
10BS	1	2	3	1	1	0	2	2	1	1	-1	2	2	-2	2
10AL	1	3	2	1	3	0	1	1	2	2	0	1	-1	1	-1
10BL	1	3	3	1	2	-1	2	2	2	0	-1	3	2	-2	2

1AS	0	-2	0	0	2	-2
1CS	0	-2	0	1	1	1
1AL	0	-3	1	1	1	2
1CL	1	-2	0	1	1	2
2AS	-1	-2	1	0	-2	-1
2BS	0	0	0	0	0	1
2AL	1	-3	1	2	3	-2
2BL	-1	-1	0	-1	1	-1
3AS	-1	-1	0	1	-1	0
3CS	0	-2	0	2	1	2
3AL	1	-3	-1	-1	2	-1
3CL	-1	-3	1	-2	2	3
4BS	1	-2	1	0	-1	-1
4CS	0	0	0	0	0	-1
4BL	-1	-3	1	0	-2	-2
4CL	0	-1	0	1	-1	1
5AS	0	-1	0	1	1	0
5CS	1	-2	0	-1	1	-1
5AL	-1	-1	1	1	1	2
5CL	1	-3	1	2	3	2

6BS	1	-2	0	-1	1	-1
6CS	1	-1	0	0	1	-1
6BL	-1	-2	1	2	2	-2
6CL	1	-2	0	1	1	1
7BS	1	-1	1	0	1	0
7CS	0	-2	-1	-1	1	-1
7BL	-1	-1	1	1	-1	1
7CL	1	-3	0	2	2	2
8AS	1	-1	0	1	-2	2
8CS	2	-1	0	2	-1	2
8AL	1	-2	0	1	-1	2
8CL	2	-2	1	2	-2	2
9AS	-1	-2	0	1	1	-1
9CS	0	0	0	1	0	-1
9AL	1	-2	1	-1	2	-2
9CL	1	-1	-1	1	-2	-1
10AS	0	-1	0	1	2	1
10BS	2	-3	-1	2	-1	2
10AL	1	-2	1	2	2	1
10BL	2	-3	1	2	-2	3

A.3 LF: Tests on Dynamic Adaptive Streaming

Red

A	-1	-1	1	-1	1	0	-1	1	1	-1	1	0	-1	0	-1
B	-2	-1	-1	-1	0	-1	0	-1	0	-1	-1	1	-2	1	-2
C	-1	-2	0	-1	1	0	0	-1	0	1	0	1	-1	-1	0
D	-2	-1	-1	-1	0	-1	-1	0	-1	1	-1	-1	-2	-2	-1
E	-1	0	1	-1	0	1	-1	2	1	1	-1	1	-1	1	0
F	-2	-1	-1	-1	-1	-1	-2	0	0	1	-1	-1	-1	-2	0

Yellow

A	-1	-2	-2	-1	1	0	-1	-1	0	1	-1	1	-1	-1	0
B	-2	-1	-1	-1	0	-1	-2	-2	-1	-1	-1	-1	-2	-2	-1
C	-2	1	0	0	0	-1	-1	1	0	1	1	2	0	-1	0
D	-2	-2	1	-1	1	-1	-2	0	-2	-1	-2	1	-1	-2	-1
E	-1	-1	0	-1	1	1	-1	-1	1	0	1	1	1	-1	0
F	-2	0	0	-1	1	-1	-2	-2	0	-1	-2	0	-1	-2	-2

Ivy

A	0	-1	0	0	1	-1	1	-1	-1	1	1	0	-1	0	-1
B	-1	-1	-2	-1	-1	-1	0	-1	0	0	-1	-2	-2	-2	-2
C	0	-2	1	0	-1	0	-1	1	0	-1	-1	-1	1	1	0
D	-1	-2	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	0	-1
E	0	-2	-1	0	1	0	-1	0	0	1	0	-1	-1	2	1
F	-1	-1	-1	-1	1	-1	-2	-2	-1	0	-2	-2	0	-1	-1

Tesco

A	2	-1	0	0	0	-1	0	-2	0	0	-1	2	-1	0	-1
B	-1	-2	-1	0	-1	-1	-2	-1	-1	-1	-1	0	-2	1	-1
C	-2	1	0	0	1	-1	-1	0	0	1	-1	1	-1	2	0
D	-1	0	-1	0	0	0	-1	-1	0	0	-2	1	-2	-1	-2
E	-1	-1	-1	-1	1	1	-1	1	0	1	1	0	0	1	0
F	-1	-2	-1	-1	-1	1	-2	0	-1	-1	0	0	-1	-2	-2

Gears

A	-2	-1	1	-1	-1	-1	-1	2	-2	-1	-1	0	-2	-1	-2
B	-2	-1	0	-1	-1	-1	-2	0	-1	-2	-2	-1	-2	-1	-2
C	-1	-2	0	0	-1	0	-1	2	-1	1	1	0	-1	-2	0
D	-1	-1	-1	-1	-1	-1	-2	1	0	0	-1	-1	-2	-2	-1
E	0	1	1	0	0	-1	-1	1	-1	0	0	2	0	0	0
F	0	0	-1	0	-1	-1	-2	-2	-2	-1	-2	2	-1	-2	-1

Red

A	-1	-2	0	-1	-2
B	-2	-2	-1	-2	-2
C	0	-1	1	0	-1
D	-1	-2	0	-1	-2
E	0	-1	0	0	-2
F	-1	-2	-1	-1	-2

Yellow

A	-1	-2	-1	-1	-1
B	-1	-2	-1	-2	-1
C	0	0	1	0	-1
D	-1	-1	-1	-1	-2
E	-1	-1	0	0	0
F	-2	-1	-1	-2	-2

Ivy

A	0	0	0	-1	-1
B	-1	-2	-2	-2	-1
C	1	-1	0	0	0
D	0	-2	-1	-1	-2
E	0	-1	0	0	0
F	-1	-2	-1	-1	-2

Tesco

A	-1	0	0	-2	-1
B	-2	-1	-1	-2	-2
C	0	0	0	0	1
D	-2	-2	0	-1	-2
E	1	1	1	1	-1
F	-1	-2	-1	0	-2

Gears

A	-1	-2	-1	-1	-1
B	-2	-2	-2	-2	-2
C	0	-1	0	0	0
D	-1	-2	-2	-1	-2
E	0	0	0	0	-1
F	-1	-2	-2	-2	-2

Appendix B

Duration of the Experiments

This appendix provides a summary of the per-subject duration of the subjective tests.

Experiment	Duration in minutes
UHD: Test with Labels	40
UHD: Test without Labels	35
HDR: Quality Aspects	10
HDR: Stalling Detection	30
LDR: Stalling Detection	30
HDR: Stalling Duration	40
LF: FOV	20
LF: Spatial Resolution	25
LF: Angular Resolution	12
LF: View Synthesis	10
LF: Static Observers	12
LF: Interpolation	35
LF: Resolution Interdependence	45
LF: Viewing Conditions	25
LF: Video Resolution	25
LF+: Stalling Detection	20
LF+: Stalling Distribution	30
LF: Dynamic Adaptive Streaming	30

Appendix C

Data Sheets of the Displays

This appendix provides full technical specifications for all the displays involved in the experiments. The shown data tables are based on the information provided by the manufacturers.

C.1 Samsung UN55JU6400

Screen size	55”
Type	Direct LED
Aspect ratio	16:9
Resolution	3840×2160
Dimensions	1242.1 mm $\times$ 718.8 mm $\times$ 63.5 mm
Brightness	300 cd/m^2
Contrast ratio	1000:1

C.2 Panasonic TX-P42S10E

Screen size	42”
Type	Plasma
Aspect ratio	16:9
Resolution	1920 × 1080
Dimensions	1029 mm × 661 mm × 105 mm
Contrast ratio	30000:1

C.3 Philips Dimenco

Screen size	55”
Type	multiview lenticular display with WOWvx 2D-plus-depth input
Aspect ratio	16:9
Resolution	1920×1080
Autostereoscopic views	28
Brightness	700 cd/m^2
Contrast ratio	1300:1
FOV	150 degrees

C.4 SIM2 HDR47ES6MB

Screen size	47”
Type	LCD TFT display with LED back light unit
Aspect ratio	16:9
Resolution	1920×1080
Dimensions	1106 mm × 650 mm × 160 mm
Brightness with full white screen	2300 cd/m^2
Contrast	from 16 to 17.5 f/stops
White point	native 6000K adjustable (5000K – 9000K)

C.5 HoloVizio 80WLT

Screen size	30”
Type	back-projection light field display
Aspect ratio	16:10
Dimensions	920 mm × 568 mm × 479 mm
2D equivalent resolution from one view point	1280 × 768
Brightness	300 cd/m^2
FOV	180 degrees full horizontal
Angular resolution	1.5 degrees

C.6 HoloVizio C80

Screen size	140”
Type	front-projection light field display
Aspect ratio	16:9
Dimensions	4000 mm × 3500 mm × 5000 mm viewing area (auditorium) not included
3D resolution	63 Mpixel
Brightness	1000 cd/m^2
FOV	40 degrees
Angular resolution	0.5 degrees
Mass	900 kg

Bibliography

[1] S. Bouchard, S. Dumoulin, J. Talbot, A.-A. Ledoux, J. Phillips, J. Monthuy-Blanc, G. Labonté-Chartrand, G. Robillard, M. Cantamesse, and P. Renaud, "Manipulating subjective realism and its impact on presence: Preliminary results on feasibility and neuroanatomical correlates," *Interacting with Computers*, vol. 24, no. 4, pp. 227–236, 2012.

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

[3] J. Casal, M. Camara, C. Diaz, J. Ruano, and N. Garcia, "Device characterization for conditional encoding," in *Meeting of the Video Quality Experts Group (VQEG)*, 2018.

[4] L. Sax, "Sex differences in hearing: Implications for best practice in the classroom," *Advances in Gender and Education*, vol. 2, pp. 13–21, 2010.

[5] M. Hyder, K. u. R. Laghari, N. Crespi, M. Haun, and C. Hoene, "Are QoE requirements for multimedia services different for men and women? Analysis of gender differences in forming QoE in Virtual Acoustic Environments," in *International Multi Topic Conference*. Springer, 2012, pp. 200–209.

[6] I. C. Zündorf, H.-O. Karnath, and J. Lewald, "Male advantage in sound localization at cocktail parties," *Cortex*, vol. 47, no. 6, pp. 741–749, 2011.

[7] A. Large, J. Beheshti, and T. Rahman, "Gender differences in collaborative web searching behavior: an elementary school study," *Information Processing & Management*, vol. 38, no. 3, pp. 427–443, 2002.

[8] K. Lamm, T. Mandl, C. Womser-Hacker, and W. Greve, "The influence of expectation and system performance on user satisfaction with retrieval systems." in *EVIA @ NTCIR*, 2010, pp. 60–68.

[9] G. J. Szybillo and J. Jacoby, "Intrinsic versus extrinsic cues as determinants of perceived product quality," *Journal of Applied Psychology*, vol. 59, no. 1, pp. 74–78, 1974.

[10] S. Burton, A. Biswas, and R. Netemeyer, "Effects of alternative nutrition label formats and nutrition reference information on consumer perceptions, comprehension, and product evaluations," *Journal of Public Policy & Marketing*, pp. 36–47, 1994.

[11] K. Berger, Y. Koudota, M. Barkowsky, and P. Le Callet, "Subjective quality assessment comparing UHD and HD resolution in HEVC transmission chains," in *Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, 2015, pp. 1–6.

[12] A. Cserkaszky, P. A. Kara, R. R. Tamboli, A. Barsi, M. G. Martini, L. Bokor, and T. Balogh, "Angularly continuous light-field format: Concept, implementation, and evaluation," *Journal of the Society for Information Display*, 2019.

[13] W. Verbeke and J. Viaene, "Consumer attitude to beef quality labeling and associations with beef quality labels," *Journal of International Food & Agribusiness Marketing*, vol. 10, no. 3, pp. 45–65, 1999.

[14] J. Li, Y. Koudota, M. Barkowsky, H. Primon, and P. Le Callet, "Comparing upscaling algorithms from HD to Ultra HD by evaluating preference of experience," in *Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2014, pp. 208–213.

[15] B. Szajna and R. W. Scamell, "The effects of information system user expectations on their performance and perceptions," *MIS Quarterly*, vol. 17, no. 4, pp. 493–516, 1993.

[16] A. Sackl, K. Masuch, S. Egger, and R. Schatz, "Wireless vs. wireline shootout: How user expectations influence Quality of Experience," in *Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2012, pp. 148–149.

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

[18] Y. Tanaka and D. Ochi, "QoE Assessment Methodologies for 4K Video Services," *NTT Technical Review*, vol. 12, no. 5, 2014.

[19] L. Hamzaoui and D. Merunka, "The impact of country of design and country of manufacture on consumer perceptions of bi-national products' quality: an empirical model based on the concept of fit," *Journal of Consumer Marketing*, vol. 23, no. 3, pp. 145–155, 2006.

[20] E. Lick, B. König, M. R. Kpossa, and V. Buller, "Sensory expectations generated by colours of red wine labels," *Journal of Retailing and Consumer Services*, vol. 37, pp. 146–158, 2017.

[21] S. Al-Juboori, I.-H. Mkwawa, L. Sun, and E. Ifeachor, "Investigation of relationships between changes in EEG features and subjective quality of HDR images," in *International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 91–96.

[22] M. Narwaria, M. P. Da Silva, P. Le Callet, and R. Pepion, "Tone mapping based HDR compression: Does it affect visual experience?" *Signal Processing: Image Communication*, vol. 29, no. 2, pp. 257–273, 2014.

[23] S.-E. Moon and J.-S. 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, pp. 236–247, 2015.

[24] L. Shi, S. Zhao, W. Zhou, and Z. Chen, "Perceptual evaluation of light field image," in *25th International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 41–45.

[25] C. Bist, R. Cozot, G. Madec, and X. Ducloux, "QoE-based brightness control for HDR displays," in *Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2017, pp. 1–6.

[26] S. Y. Rieh and N. Belkin, "Interaction on the web: Scholars' judgement of information quality and cognitive authority," in *Proceedings of the 63rd Annual Meeting of the American Society for Information Science*, vol. 37, 2000, pp. 25–38.

[27] S. Gächter, H. Orzen, E. Renner, and C. Starmer, "Are experimental economists prone to framing effects? A natural field experiment," *Journal of Economic Behavior & Organization*, vol. 70, no. 3, pp. 443–446, 2009.

[28] I. Viola, M. Rerabek, T. Bruylants, P. Schelkens, F. Pereira, and T. Ebrahimi, "Objective and subjective evaluation of light field image compression algorithms," in *Picture Coding Symposium (PCS)*. IEEE, 2016, pp. 1–5.

[29] S. Darukumalli, P. A. Kara, A. Barsi, M. G. Martini, T. Balogh, and A. Chehaibi, "Performance comparison of subjective assessment methodologies for light field displays," in *International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2016, pp. 28–33.

[30] L. L. Garber Jr, E. M. Hyatt, and L. Nafees, "The effects of food color on perceived flavor: a factorial investigation in India," *Journal of Food Products Marketing*, vol. 22, no. 8, pp. 930–948, 2016.

[31] F. Marton, M. Agus, E. Gobbetti, G. Pintore, and M. B. Rodriguez, "Natural exploration of 3D massive models on large-scale light field displays using the fox proximal navigation technique," *Computers & Graphics*, vol. 36, no. 8, pp. 893–903, 2012.

[32] V. Hulusic, G. Valenzise, E. Provenzi, K. Debattista, and F. Dufaux, "Perceived dynamic range of HDR images," in *Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2016, pp. 1–6.

[33] P. Paudyal, F. Battisti, A. Neri, and M. Carli, "A study of the impact of light fields watermarking on the perceived quality of the refocused data," in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2015*. IEEE, 2015, pp. 1–4.

[34] R. R. Tamboli, P. A. Kara, B. Appina, M. G. Martini, S. S. Channappayya, and S. Jana, "Effect of Primitive Features of Content on Perceived Quality of Light Field Visualization," in *10th International Conference on Quality of Multimedia Experience (QoMEX)*, 2018.

[35] P. T. Kovács, R. Bregović, A. Boev, A. Barsi, and A. Gotchev, "Quantifying spatial and angular resolution of light-field 3-D displays," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1213–1222, 2017.

[36] G. Van Wallendael, P. Coppens, T. Paridaens, N. Van Kets, W. Van den Broeck, and P. Lambert, "Perceptual quality of 4K-resolution video content compared to HD," in *Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2016, pp. 1–6.

[37] S. H. Bae, J. Kim, M. Kim, S. Cho, and J. S. Choi, "Assessments of subjective video quality on HEVC-encoded 4K-UHD video for beyond-HDTV broadcasting services," *IEEE Transactions on Broadcasting*, vol. 59, no. 2, pp. 209–222, 2013.

[38] R. Weerakkody, M. Mrak, V. Baroncini, J.-R. Ohm, T. K. Tan, and G. J. Sullivan, "Verification testing of HEVC compression performance for UHD video," in *Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2014, pp. 1083–1087.

[39] F. Xie, M. T. Pourazad, P. Nasiopoulos, and J. Slevinsky, "Determining bitrate requirement for UHD video content delivery," in *International Conference on Consumer Electronics (ICCE)*. IEEE, 2016, pp. 241–242.

[40] Y. Zhu, L. Song, R. Xie, and W. Zhang, "SJTU 4K video subjective quality dataset for content adaptive bit rate estimation without encoding," in *International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 2016, pp. 1–4.

[41] R. Sotelo, J. Joskowicz, M. Anedda, M. Murroni, and D. D. Giusto, "Subjective video quality assessments for 4K UHDTV," in *International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 2017, pp. 1–6.

[42] A. Mackin, M. Afonso, F. Zhang, and D. Bull, "A study of subjective video quality at various spatial resolutions," in *25th International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2830–2834.

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

[44] Netflix. (2017) Toward a practical perceptual video quality metric. [Online]. Available: https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[45] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Benchmarking of quality metrics on ultra-high definition video sequences," in *18th International Conference on Digital Signal Processing (DSP)*. IEEE, 2013, pp. 1–8.

[46] G. H. Mead, *Mind, self and society*. University of Chicago Press, 1934, vol. 111.

[47] F. Tannenbaum, *Crime and the Community*. JSTOR, 1938.

[48] P. C. Wason, "On the failure to eliminate hypotheses in a conceptual task," *Quarterly journal of experimental psychology*, vol. 12, no. 3, pp. 129–140, 1960.

[49] R. S. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises," *Review of general psychology*, vol. 2, no. 2, pp. 175–220, 1998.

[50] E. F. Loftus and H. G. Hoffman, "Misinformation and memory: The creation of new memories." *Journal of Experimental Psychology: General*, vol. 118, no. 1, pp. 100–104, 1989.

[51] A. Tversky and D. Kahneman, "The framing of decisions and the psychology of choice," in *Environmental Impact assessment, technology assessment, and risk analysis*. Springer, 1985, pp. 107–129.

[52] ——, "Prospect theory: An analysis of decision under risk," *Econometrica*, vol. 47, no. 2, pp. 263–291, 1979.

[53] M. Sherif, D. Taub, and C. I. Hovland, "Assimilation and contrast effects of anchoring stimuli on judgments," *Journal of experimental psychology*, vol. 55, no. 2, p. 150, 1958.

[54] P. R. Wilson, "Perceptual distortion of height as a function of ascribed academic status," *The Journal of social psychology*, vol. 74, no. 1, pp. 97–102, 1968.

[55] I. Sinha and R. Batra, "The effect of consumer price consciousness on private label purchase," *International journal of research in marketing*, vol. 16, no. 3, pp. 237–251, 1999.

[56] S. Y. Rieh and N. J. Belkin, "Understanding judgment of information quality and cognitive authority in the WWW," in *Proceedings of the 61st Annual Meeting of the American Society for Information Science*, vol. 35, 1998, pp. 279–289.

[57] L. Festinger, *A Theory of Cognitive Dissonance*. Stanford University Press, 1962, vol. 2.

[58] K. Brunnström, S. A. Beker, K. De Moor, A. Dooms, S. Egger, M.-N. Garcia, T. Hossfeld, S. Jumisko-Pyykkö, C. Keimel, M.-C. Larabi *et al.*, "Qualinet White Paper on Definitions of Quality of Experience," 2013.

[59] J. Jacoby, J. C. Olson, and R. A. Haddock, "Price, brand name, and product composition characteristics as determinants of perceived quality," *Journal of Applied Psychology*, vol. 55, no. 6, pp. 570–579, 1971.

[60] P. S. Richardson, A. S. Dick, and A. K. Jain, "Extrinsic and intrinsic cue effects on perceptions of store brand quality," *The Journal of Marketing*, pp. 28–36, 1994.

[61] D. DelVecchio, "Consumer perceptions of private label quality: the role of product category characteristics and consumer use of heuristics," *Journal of retailing and Consumer Services*, vol. 8, no. 5, pp. 239–249, 2001.

[62] F. L. Heisey, "Perceived quality and predicted price: Use of the minimum information environment in evaluating apparel," *Clothing and Textiles Research Journal*, vol. 8, no. 4, pp. 22–28, 1990.

[63] J. K. Johansson, "Determinants and effects of the use of made in labels," *International Marketing Review*, vol. 6, no. 1, 1989.

[64] S. A. Ahmed and A. d Astous, "Comparison of country of origin effects on household and organizational buyers product perceptions," *European Journal of Marketing*, vol. 29, no. 3, pp. 35–51, 1995.

[65] R. Batra, V. Ramaswamy, D. L. Alden, J.-B. E. Steenkamp, and S. Ramachander, "Effects of brand local and nonlocal origin on consumer attitudes in developing countries," *Journal of consumer psychology*, vol. 9, no. 2, pp. 83–95, 2000.

[66] W. Zhao and X. Zhou, "Status inconsistency and product valuation in the California wine market," *Organization Science*, vol. 22, no. 6, pp. 1435–1448, 2011.

[67] J. Masson, P. Aurier, and F. d'hauteville, "Effects of non-sensory cues on perceived quality: the case of low-alcohol wine," *International Journal of Wine Business Research*, vol. 20, no. 3, pp. 215–229, 2008.

[68] F. d'Hauteville, M. Fornerino, and J. Philippe Perrouty, "Disconfirmation of taste as a measure of region of origin equity: An experimental study on five French wine regions," *International Journal of Wine Business Research*, vol. 19, no. 1, pp. 33–48, 2007.

[69] L. L. Garber Jr, E. M. Hyatt, and R. G. Starr Jr, "The effects of food color on perceived flavor," *Journal of Marketing Theory and Practice*, vol. 8, no. 4, pp. 59–72, 2000.

[70] C. Wagner, *The Wagner Color Response Report.* Wagner Institute for Color Research, 1985.

[71] A. Simonson and B. H. Schmitt, *Marketing aesthetics: The strategic management of brands, identity, and image.* Simon and Schuster, 1997.

[72] M. Schöffler, "Overall Listening Experience — a new Approach to Subjective Evaluation of Audio." Doctoral thesis at University of Erlangen Nürnberg, 2017.

[73] K. Lamm, T. Mandl, C. Womser-Hacker, and W. Greve, "User experiments with search services: Methodological challenges for measuring the perceived quality," in *3rd Workshop on Perceptual Quality of Systems (PQS)*, 2010.

[74] A. Sackl, P. Zwickl, S. Egger, and P. Reichl, "The role of cognitive dissonance for QoE evaluation of multimedia services," in *Globecom Workshops.* IEEE, 2012, pp. 1352–1356.

[75] P. A. Kara, L. Bokor, and S. Imre, "Distortions in QoE assessment of 3D multimedia services on multi-access mobile devices," in *9th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob).* IEEE, 2013, pp. 311–318.

[76] ——, "Seeing is believing and vice versa: Investigation of the altered perception during subjective assessment of streaming multimedia," in *Tenth International Conference on Signal-Image Technology and Internet-Based Systems (SITIS).* IEEE, 2014, pp. 539–545.

[77] A. Sackl, S. Egger, P. Zwickl, and P. Reichl, "The QoE alchemy: Turning quality into money. Experiences with a refined methodology for the evaluation of willingness-to-pay for service quality," in *Fourth International Workshop on Quality of Multimedia Experience (QoMEX).* IEEE, 2012, pp. 170–175.

[78] A. Sackl, P. Zwickl, and P. Reichl, "From Quality of Experience to Willingness to Pay for Interconnection Service Quality." in *Networking Workshops*. Springer, 2012, pp. 89–96.

[79] A. Sackl, R. Schatz, and A. Raake, "More than I ever wanted or just good enough? User expectations and subjective quality perception in the context of networked multimedia services," *Quality and User Experience*, vol. 2, no. 1, 2017.

[80] P. A. Kara, L. Bokor, A. Sackl, and M. Mourão, "What your phone makes you see: Investigation of the effect of end-user devices on the assessment of perceived multimedia quality," in *Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2015, pp. 1–6.

[81] J. Ebbing, "The halo effect of smartphone brands: smartphone's brand equity influence on the user experience of third-party smartphone applications." M.Sc. thesis at University of Twente, 2017.

[82] L. Festinger, *Conflict, decision, and dissonance.* Stanford University Press, 1964.

[83] R. E. Knox and J. A. Inkster, "Postdecision dissonance at post time." *Journal of personality and social psychology*, vol. 8, no. 4, pp. 319–323, 1968.

[84] M. Narwaria, M. P. Da Silva, and P. Le Callet, "High dynamic range visual quality of experience measurement: Challenges and perspectives," in *Visual Signal Quality Assessment*. Springer, 2015, pp. 129–155.

[85] M. Narwaria, M. P. Da Silva, P. Le Callet, G. Valenzise, F. De Simone, and F. Dufaux, "Quality of experience and HDR: concepts and how to measure it," in *High Dynamic Range Video*. Elsevier, 2016, pp. 431–454.

[86] M. Narwaria, M. P. Da Silva, and P. Le Callet, "HDR-VQM: An objective quality measure for high dynamic range video," *Signal Processing: Image Communication*, vol. 35, pp. 46–60, 2015.

[87] M. D. Fairchild, "The HDR photographic survey," in *Color and Imaging Conference*, vol. 2007, no. 1. Society for Imaging Science and Technology, 2007, pp. 233–238.

[88] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling, and H. Brendel, "Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays," in *Digital Photography*, vol. 9023. SPIE, 2014.

[89] J.-L. Blin, "SAMVIQ – Subjective assessment methodology for video quality," *Rapport technique BPN*, vol. 56, p. 24, 2003.

[90] U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J.-N. Antons, K. Y. Chan, N. Ramzan, and K. Brunnström, "Psychophysiology-based QoE assessment: a survey," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 6–21, 2017.

[91] S.-E. Moon and J.-S. Lee, "EEG connectivity analysis in perception of tone-mapped high dynamic range videos," in *23rd International Conference on Multimedia*. ACM, 2015, pp. 987–990.

[92] D. Darcy, E. Gitterman, A. Brandmeyer, S. Daly, and P. Crum, "Physiological capture of augmented viewing states: objective measures of high-dynamic-range and wide-color-gamut viewing experiences," *Electronic Imaging*, vol. 2016, no. 16, pp. 1–9, 2016.

[93] S. Daly, E. Gitterman, and G. Mulliken, "Pupillometry of HDR Video Viewing," *Electronic Imaging*, vol. 2018, no. 14, pp. 1–8, 2018.

[94] P. A. Kara, A. Cserkaszky, and M. G. Martini, "Premium HDR: The Impact of a Single Word on the Quality of Experience of HDR Video," in *International Conference on Multimedia and Expo (ICME), Emerging Multimedia Systems and Applications (EMSA)*. IEEE, 2018.

[95] S. van Kester, T. Xiao, R. E. Kooij, K. Brunnström, and O. K. Ahmed, "Estimating the impact of single and multiple freezes on video quality," in *Human Vision and Electronic Imaging XVI*, vol. 7865. SPIE, 2011.

[96] M. A. Usman, M. R. Usman, and S. Y. Shin, "The impact of temporal impairment on quality of experience (QoE) in video streaming: A no reference (NR) subjective and objective study," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 9, no. 8, pp. 1570–1577, 2015.

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

[98] P. Yu, F. Liu, Y. Geng, W. Li, and X. Qiu, "An objective multi-layer qoe evaluation for tcp video streaming," in *IFIP/IEEE International Symposium on Integrated Network Management (IM)*, 2015, pp. 1255–1260.

[99] P. A. Kara, W. Robitza, M. G. Martini, C. T. Hewage, and F. M. Felisberti, "Getting used to or growing annoyed: How perception thresholds and acceptance of frame freezing vary over time in 3D video streaming," in *International Conference on Multimedia & Expo (ICME), 22nd International Packet Video Workshop (PV)*. IEEE, 2016, pp. 1–6.

[100] P. A. Kara, M. G. Martini, C. T. Hewage, and F. M. Felisberti, "Times change, stalling stays: Subjective quality assessment over time of stalling in autostereoscopic 3D video services," in *12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2016, pp. 787–792.

[101] G. Lippmann, "Epreuves reversibles. Photographies integrals," *Comptes-Rendus Academie des Sciences*, vol. 146, pp. 446–451, 1908.

[102] A. Gershun, "The light field," *Studies in Applied Mathematics*, vol. 18, no. 1-4, pp. 51–151, 1939.

[103] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing*. MIT Press, 1991, pp. 3–20.

[104] M. W. Halle, "Holographic stereograms as discrete imaging systems," in *IS&T/SPIE International Symposium on Electronic Imaging: Science and Technology*, 1994, pp. 73–84.

[105] M. Levoy and P. Hanrahan, "Light field rendering," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 1996, pp. 31–42.

[106] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 1996, pp. 43–54.

[107] D. Lanman and D. Luebke, "Near-eye light field displays," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, p. 220, 2013.

[108] A. J. Hansen, M. Kraus, and J. Klein, "Light field rendering for head mounted displays using pixel reprojection," in *International Conference on Computer Graphics Theory and Applications*, 2017.

[109] V. K. Adhikarla, J. Sodnik, P. Szolgay, and G. Jakus, "Exploring direct 3D interaction for full horizontal parallax light field displays using leap motion controller," *Sensors*, vol. 15, no. 4, pp. 8642–8663, 2015.

[110] T. Balogh, Z. Nagy, P. T. Kovács, and V. K. Adhikarla, "Natural 3D content on glasses-free light-field 3D cinema," in *IS&T/SPIE Electronic Imaging*, 2013.

[111] P. A. Kara, Z. Nagy, M. G. Martini, and A. Barsi, "Cinema as large as life: Large-scale light field cinema system," in *2017 International Conference on 3D Immersion (IC3D)*. IEEE, 2017.

[112] J.-H. Lee, J. Park, D. Nam, S. Y. Choi, D.-S. Park, and C. Y. Kim, "Optimal projector configuration design for 300-Mpixel multi-projection 3D display," *Optics express*, vol. 21, no. 22, pp. 26 820–26 835, 2013.

[113] N. Inoue, S. Iwasawa, and M. Okui, "Public viewing of 200-inch glasses-free 3D display system," *New Breeze*, vol. 26, no. 4, pp. 10–11, 2014.

[114] R. L.-G. Masahiro Kawakita, Shoichro Iwasawa and N. Inoue, "Glasses-free large-screen three-dimensional display and super multiview camera for highly realistic communication," *Optical Engineering*, vol. 57, no. 6, 2018.

[115] P. T. Kovács, K. Lackner, A. Barsi, Á. Balázs, A. Boev, R. Bregović, and A. Gotchev, "Measurement of perceived spatial resolution in 3D light-field displays," in *International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 768–772.

[116] P. T. Kovács, A. Boev, R. Bregović, and A. Gotchev, "Quality measurements of 3D light-field displays," in *Eighth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2014.

[117] T. Balogh, P. T. Kovacs, and A. Barsi, "Holovizio 3D display system," in *3DTV Conference*. IEEE, 2007, pp. 1–4.

[118] P. A. Kara, R. R. Tamboli, O. Doronin, A. Cserkaszky, A. Barsi, Z. Nagy, M. G. Martini, and A. Simon, "The key performance indicators of projection-based light field visualization," *Journal of Information Display*, vol. 20, no. 2, pp. 81–93, 2019.

[119] L. Ni, Z. Li, H. Li, and X. Liu, "360-degree large-scale multiprojection light-field 3D display system," *Applied Optics*, vol. 57, no. 8, pp. 1817–1823, 2018.

[120] R. N. Strickland, *Image-processing techniques for tumor detection.* CRC Press, 2002.

[121] C. Conti, L. D. Soares, P. Nunes, C. Perra, P. A. Assunção, M. Sjöström, Y. Li, R. Olsson, and U. Jennehag, "Light field image compression," in *3D Visual Content Creation, Coding and Delivery.* Springer, 2019, pp. 143–176.

[122] R. R. Tamboli, A. Cserkaszky, P. A. Kara, A. Barsi, and M. G. Martini, "Objective quality evaluation of an angularly-continuous light-field format," in *International Conference on 3D Immersion (IC3D).* IEEE, 2018.

[123] I. Viola, M. Rerabek, and T. Ebrahimi, "Comparison and evaluation of light field image coding approaches," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1092–1106, 2017.

[124] A. Dricot, J. Jung, M. Cagnazzo, B. Pesquet, F. Dufaux, P. T. Kovács, and V. K. Adhikarla, "Subjective evaluation of super multi-view compressed contents on high-end light-field 3D displays," *Signal Processing: Image Communication*, vol. 39, pp. 369–385, 2015.

[125] W. Ahmad, M. Sjöström, and R. Olsson, "Compression scheme for sparsely sampled light field data based on pseudo multi-view sequences," in *Optics, Photonics, and Digital Technologies for Imaging Applications V*, 2018.

[126] B. Guo, J. Wen, and Y. Han, "Two-pass Light Field Image Compression for Spatial Quality and Angular Consistency," *Cornell University's Computing Research Repository (CoRR)*, 2018.

[127] X. Zhang, P. A. Chou, M. Sun, M. Tang, S. Wang, S. Ma, and W. Gao, "Surface Light Field Compression using a Point Cloud Codec," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)*, 2018.

[128] O. Doronin, A. Barsi, P. A. Kara, and M. G. Martini, "Ray tracing for HoloVizio light field displays," in *International Conference on 3D Immersion (IC3D)*. IEEE, 2017, pp. 1–8.

[129] M. Agus, F. Bettio, A. Giachetti, E. Gobbetti, J. A. I. Guitián, F. Marton, J. Nilsson, and G. Pintore, "An interactive 3D medical visualization system based on a light field display," *The Visual Computer*, vol. 25, no. 9, pp. 883–893, 2009.

[130] T. E. Bishop, S. Zanetti, and P. Favaro, "Light field superresolution," in *International Conference on Computational Photography (ICCP)*. IEEE, 2009, pp. 1–9.

[131] Y. Takaki, "Super multi-view display with 128 viewpoints and viewpoint formation," in *Stereoscopic Displays and Applications XX*, vol. 7237. SPIE, 2009.

[132] A. Cserkaszky, A. Barsi, Z. Nagy, G. Puhr, T. Balogh, and P. A. Kara, "Real-time light-field 3D telepresence," in *7th European Workshop on Visual Information Processing (EUVIP)*, 2018.

[133] G. Eilertsen, R. K. Mantiuk, and J. Unger, "A comparative review of tone-mapping algorithms for high dynamic range video," in *Computer Graphics Forum*, vol. 36, no. 2. Wiley Online Library, 2017, pp. 565–592.

[134] H. Najaf-Zadeh, M. Budagavi, and E. Faramarzi, "VR+HDR: A system for view-dependent rendering of HDR video in virtual reality," in *International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 1032–1036.

[135] M. Melo, H. Coelho, K. Bouatouch, M. Bessa, R. Cozot, and A. Chalmers, "Tone Mapping HDR Panoramas for Viewing in Head Mounted Displays," in *13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018)*, vol. 1, 2018, pp. 232–239.

[136] X. Yang, L. Zhang, T.-T. Wong, and P.-A. Heng, "Binocular tone mapping," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 4, 2012.

[137] P. Wang, X. Sang, Y. Zhu, S. Xie, D. Chen, N. Guo, and C. Yu, "Image quality improvement of multi-projection 3D display through tone mapping based optimization," *Optics express*, vol. 25, no. 17, pp. 20 894–20 910, 2017.

[138] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Transactions on Graphics (TOG)*, vol. 21, no. 3, pp. 267–276, 2002.

[139] O. Doronin and A. Barsi, "Estimation of global luminance for HoloVizio 3D display," in *International Conference on 3D Immersion (IC3D)*, 2018.

[140] P. T. Kovács and T. Balogh, "3D visual experience," *High-Quality Visual Experience*, pp. 391–410, 2010.

[141] ——, "3D display technologies and effects on the human vision system," in *2nd International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 2011, pp. 1–33.

[142] ——, "3D light-field display technologies," *Emerging Technologies for 3D Video: Creation, Coding, Transmission and Rendering*, pp. 336–345, 2013.

[143] M. S. K. Gul and B. K. Gunturk, "Spatial and angular resolution enhancement of light fields using convolutional neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2146–2159, 2018.

[144] R. Tamboli, K. K. Vupparaboina, J. Ready, S. Jana, and S. Channappayya, "A subjective evaluation of true 3D images," in *International Conference on 3D Imaging (IC3D)*. IEEE, 2014, pp. 1–8.

[145] R. Tamboli, B. Appina, S. Channappayya, and S. Jana, "Super-multiview content with high angular resolution: 3D quality assessment on horizontal-parallax lightfield display," *Signal Processing: Image Communication*, vol. 47, pp. 42–55, 2016.

[146] R. Tamboli, A. B., S. Channappayya, and S. Jana, "Achieving high angular resolution via view synthesis: quality assessment of 3D content on super multiview lightfield display," in *International Conference on 3D Immersion (IC3D)*. IEEE, 2017.

[147] R. R. Tamboli, M. S. Reddy, P. A. Kara, M. G. Martini, S. S. Channappayya, and S. Jana, "A High-angular-resolution Turntable Data-set for Experiments on Light Field Visualization Quality," in *10th International Conference on Quality of Multimedia Experience (QoMEX)*, 2018.

[148] C. Perra, "Assessing the quality of experience in viewing rendered decompressed light fields," *Multimedia Tools and Applications*, pp. 1–20, 2018.

[149] C. Perra, W. Song, and A. Liotta, "Effects of light field subsampling on the quality of experience in refocusing applications," in *Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018, pp. 1–3.

[150] C. Perra, F. Murgia, and D. D. Giusto, "An analysis of 3D point cloud reconstruction from light field images," in *IPTA*, 2016, pp. 1–6.

[151] C. Perra and P. Assuncao, "High efficiency coding of light field images based on tiling and pseudo-temporal data arrangement," in *International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2016, pp. 1–4.

[152] I. Viola and T. Ebrahimi, "A new framework for interactive quality assessment with application to light field coding," in *SPIE*, 2017.

[153] P. Paudyal, F. Battisti, M. Sjöström, R. Olsson, and M. Carli, "Towards the perceptual quality evaluation of compressed light field images," *IEEE Transactions on Broadcasting*, vol. 63, no. 3, pp. 507–522, 2017.

[154] I. Viola, M. Rerabek, and T. Ebrahimi, "A new approach to subjectively assess quality of plenoptic content," in *Applications of Digital Image Processing XXXIX*, vol. 9971. SPIE, 2016.

[155] P. Paudyal, J. Gutierrez, P. Le Callet, M. Carli, and F. Battisti, "Characterization and selection of light field content for perceptual assessment," in *9th International Conference on Quality of Multimedia Experience (QoMEX)*, 2017.

[156] J. Gutiérrez, P. Paudyal, M. Carli, F. Battisti, and P. Le, "Perceptual analysis and characterization of light field content," *VQEG eLetter*, p. 49, 2017.

[157] P. Paudyal, F. Battisti, and M. Carli, "Effect of visualization techniques on subjective quality of light field images," in *International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 196–200.

[158] ——, "Reduced reference quality assessment of light field images," *IEEE Transactions on Broadcasting*, 2019.

[159] P. Paudyal, R. Olsson, M. Sjöström, F. Battisti, and M. Carli, "Smart: a light field image quality dataset," in *7th International Conference on Multimedia Systems*. ACM, 2016, pp. 374–379.

[160] F. Murgia and D. Giusto, "A database for evaluating the quality of experience in light field applications," in *24th Telecommunications Forum (TELFOR)*. IEEE, 2016, pp. 1–4.

[161] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

[162] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi, "A 4D light-field dataset and CNN architectures for material recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 121–138.

[163] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields." in *Vision, Modeling & Visualization*, 2013, pp. 225–226.

[164] S. Wang, K. S. Ong, P. Surman, J. Yuan, Y. Zheng, and X. W. Sun, "Quality of experience measurement for light field 3D displays on multilayer LCDs," *Journal of the Society for Information Display*, vol. 24, no. 12, pp. 726–740, 2016.

[165] A. Cserkaszky, P. A. Kara, A. Barsi, and M. G. Martini, "The potential synergies of visual scene reconstruction and medical image reconstruction," in *SPIE Novel Optical Systems Design and Optimization XXI*, 2018.

[166] L. Lévêque, H. Liu, S. Baraković, J. B. Husić, M. Martini, M. Outtas, L. Zhang, A. Kumcu, L. Platisa, R. Rodrigues *et al.*, "On the subjective assessment of the perceived quality of medical images and videos," in *Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018, pp. 1–6.

[167] M. Wijnants, H. Lievens, N. Michiels, J. Put, P. Quax, and W. Lamotte, "Standards-compliant http adaptive streaming of static light fields," in *24th Symposium on Virtual Reality Software and Technology*. ACM, 2018.

[168] B. Laugwitz, T. Held, and M. Schrepp, "Construction and evaluation of a user experience questionnaire," in *Symposium of the Austrian HCI and Usability Engineering Group*. Springer, 2008, pp. 63–76.

[169] V. K. Adhikarla, F. Marton, T. Balogh, and E. Gobbetti, "Real-time adaptive content retargeting for live multi-view capture and light field display," *The Visual Computer*, vol. 31, no. 6–8, pp. 1023–1032, 2015.

[170] R. R. Tamboli, P. A. Kara, A. Cserkaszky, A. Barsi, M. G. Martini, B. Appina, S. S. Channappayya, and S. Jana, "3D Objective Quality Assessment of Light Field Video Frames," in *3DTV Conference*, 2018.

[171] W. Xiang, G. Wang, M. Pickering, and Y. Zhang, "Big video data for light-field-based 3D telemedicine," *IEEE Network*, vol. 30, no. 3, pp. 30–38, 2016.

[172] P. A. Kara, P. T. Kovács, M. G. Martini, A. Barsi, K. Lackner, and T. Balogh, "From a Different Point of View: How the Field of View of Light Field Displays affects the Willingness to Pay and to Use," in *Eighth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2016.

[173] ——, "Viva la Resolution: The Perceivable Differences between Image Resolutions for Light Field Displays," in *5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS)*, 2016, pp. 107–111.

[174] P. A. Kara, M. G. Martini, P. T. Kovács, S. Imre, A. Barsi, K. Lackner, and T. Balogh, "Perceived quality of angular resolution for light field displays and the validity of subjective assessment," in *International Conference on 3D Imaging (IC3D)*, 2016.

[175] P. A. Kara, P. T. Kovács, S. Vagharshakyan, M. G. Martini, A. Barsi, T. Balogh, A. Chuchvara, and A. Chehaibi, "The Effect of Light Field Reconstruction and Angular Resolution Reduction on the Quality of Experience," in *12th International Conference on Signal Image Technology & Internet Based Systems (SITIS) 3rd International Workshop on Quality of Multimedia Services (QUAMUS)*, 2016.

[176] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Image based rendering technique via sparse representation in shearlet domain," in *International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 1379–1383.

[177] P. A. Kara, A. Cserkaszky, S. Darukumalli, A. Barsi, and M. G. Martini, "On the Edge of the Seat: Reduced Angular Resolution of a Light Field Cinema with Fixed Observer Positions," in *9th International Conference on Quality of Multimedia Experience (QoMEX)*, 2017.

[178] A. Cserkaszky, P. A. Kara, A. Barsi, and M. G. Martini, "To Interpolate or not to Interpolate: Subjective Assessment of Interpolation Performance on a Light Field Display," in *International Conference on Multimedia and Expo (ICME) 8th Workshop on Hot Topics in 3D Multimedia (Hot3D)*. IEEE, 2017, pp. 55–60.

[179] S. T. Barnard and W. B. Thompson, "Disparity analysis of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 4, pp. 333–340, 1980.

[180] R. Yang, G. Welch, and G. Bishop, "Real-time consensus-based scene reconstruction using commodity graphics hardware," in *Computer Graphics Forum*, vol. 22, no. 2. Wiley Online Library, 2003, pp. 207–216.

[181] S. Smirnov, M. Georgiev, and A. Gotchev, "Comparison of cost aggregation techniques for free-viewpoint image interpolation based on plane sweeping," *Ninth International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2015.

[182] P. A. Kara, A. Cserkaszky, A. Barsi, T. Papp, M. G. Martini, and L. Bokor, "The Interdependence of Spatial and Angular Resolution in the Quality of Experience of Light Field Visualization," in *International Conference on 3D Immersion (IC3D)*. IEEE, 2017.

[183] P. A. Kara, R. R. Tamboli, A. Cserkaszky, M. G. Martini, A. Barsi, and L. Bokor, "The Viewing Conditions of Light-Field Video for Subjective Quality Assessment," in *International Conference on 3D Immersion (IC3D)*. IEEE, 2018.

[184] ——, "The perceived quality of light-field video services," in *SPIE Applications of Digital Image Processing XLI*, 2018.

[185] T. Ebrahimi, S. Foessel, F. Pereira, and P. Schelkens, "JPEG Pleno: Toward an efficient representation of visual reality," *IEEE Multimedia*, vol. 23, no. 4, pp. 14–20, 2016.

[186] I. Sodagar, "The MPEG-DASH standard for multimedia streaming over the Internet," *IEEE MultiMedia*, vol. 18, no. 4, pp. 62–67, 2011.

[187] T. Stockhammer, "Dynamic adaptive streaming over HTTP–: standards and design principles," in *Proceedings of the second annual ACM conference on Multimedia systems*. ACM, 2011, pp. 133–144.

[188] O. Ognenoski, M. M. Nasralla, M. Razaak, M. G. Martini, and P. Amon, "DASH-based video transmission over LTE networks," in *International Conference on Communications (ICC)*. IEEE, 2015, pp. 1783–1787.

[189] O. Oyman and S. Singh, "Quality of experience for HTTP adaptive streaming services," *IEEE Communications Magazine*, vol. 50, no. 4, 2012.

[190] T. Hossfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen, "Initial delay vs. interruptions: Between the devil and the deep blue sea," in *Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2012, pp. 1–6.

[191] P. A. Kara, A. Cserkaszky, A. Barsi, M. G. Martini, and T. Balogh, "Towards Adaptive Light Field Video Streaming," *IEEE COMSOC MMTC Communications – Frontiers*, vol. 12, no. 4, pp. 50–55, 2017.

[192] A. Cserkaszky, P. A. Kara, A. Barsi, and M. G. Martini, "Towards display-independent light-field formats," in *International Conference on 3D Immersion (IC3D)*. IEEE, 2017.

[193] P. A. Kara, M. G. Martini, and S. Rossi, "One Spoonful or Multiple Drops: Investigation of Stalling Distribution and Temporal Information for Quality of Experience over Time," in *International Conference on Telecommunications and Multimedia (TEMU)*. IEEE, 2016, pp. 157–162.

[194] P. A. Kara, A. Cserkaszky, M. G. Martini, A. Barsi, L. Bokor, and T. Balogh, "Evaluation of the Concept of Dynamic Adaptive Streaming of Light Field Video," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 407–421, 2018.

[195] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens, "Quantifying the influence of rebuffering interruptions on the user's quality of experience during mobile video watching," *IEEE Transactions on Broadcasting*, vol. 59, no. 1, pp. 47–61, 2013.

[196] R. K. Mok, E. W. Chan, and R. K. Chang, "Measuring the quality of experience of HTTP video streaming," in *IFIP/IEEE International Symposium on Integrated Network Management (IM)*, 2011, pp. 485–492.

[197] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, and Z. Wang, "A quality-of-experience index for streaming video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 154–166, 2017.