

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

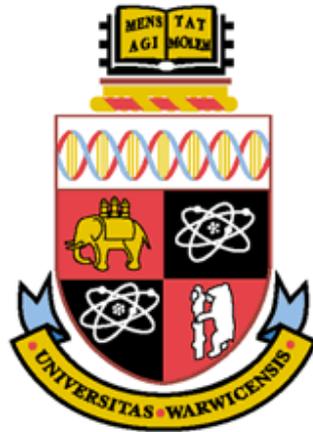
A Thesis Submitted for the Degree of PhD at the University of Warwick

<http://go.warwick.ac.uk/wrap/47727>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.



Auditory-Visual Interaction in Computer Graphics

Vedad Hulusić

BSc

*A thesis submitted in partial fulfilment of the requirements for
the degree of
Doctor of Philosophy in Engineering*

School of Engineering

University of Warwick

May 2011

Contents

Acknowledgements	xiii
Declaration	xv
List of Publications	xvi
Abstract	xix
1 Introduction	1
1.1 High-fidelity rendering	2
1.1.1 Applications	3
1.2 Cross-modal interaction	5
1.3 Research aim and objectives	6
1.4 Thesis outline	7
2 Human Sensory System	9
2.1 Human visual system	10
2.1.1 The anatomy of the HVS	10
2.1.2 Visual perception	15
2.2 Human auditory system	18
2.2.1 Sound properties	18
2.2.2 Peripheral auditory anatomy	24
2.2.3 Auditory perception	28
2.3 Attention and Perception	32
2.3.1 Attentional capture	33
2.3.2 Attentional resources and limitations	35
2.4 Summary	37
3 High-Fidelity Rendering	38
3.1 Radiometry	40
3.2 Light reflectance models	41
3.3 Light transport	42
3.4 Rasterisation	43
3.5 Ray tracing	45

3.6	Path tracing	47
3.7	Irradiance caching	48
3.8	Photon mapping	49
3.9	Summary	49
4	Cross-Modal Interaction	51
4.1	Auditory-visual cross-modal interaction in psychology	51
4.1.1	Auditory influence on visual perception	53
4.1.2	Visual influence on auditory perception	56
4.2	Auditory-visual cross-modal integration	58
4.3	Auditory-visual cross-modal interaction in computer graphics	61
4.3.1	Auditory rendering	61
4.3.2	Visual rendering	62
4.4	Summary	66
5	The Influence of Cross-Modal Interaction on Perceived Rendering Quality Thresholds	68
5.1	Introduction	68
5.2	Experiment	69
5.2.1	Stimuli	70
5.2.2	Visual difference predictor	72
5.2.3	Hardware and rendering time	75
5.2.4	Procedure	75
5.3	Results	77
5.3.1	Statistical analysis of psychophysical experiment	77
5.3.2	Comparison using VDP	83
5.4	Discussion	84
6	Beat Rate Effect on Frame Rate Perception	86
6.1	Introduction	86
6.2	Experiments	87
6.2.1	Design	88
6.2.2	Participants	89
6.2.3	Apparatus	91
6.2.4	Stimuli	91
6.2.5	Procedure	92
6.2.6	Analysis methods	96
6.3	Results	98
6.3.1	Static scenes	98
6.3.2	Dynamic scenes	102
6.4	Discussion	105

7	Exploiting Audio-Visual Cross-Modal Interaction	109
7.1	Introduction	109
7.2	Experiments	110
7.2.1	Design	111
7.2.2	Participants	111
7.2.3	Apparatus	111
7.2.4	Stimuli	113
7.2.5	Procedure	115
7.3	Results	116
7.3.1	Test 1: Camera movement speed influence on animation smoothness perception	116
7.3.2	Test 2: Sound effect's influence on perceived smoothness threshold	117
7.3.3	Test 3: Sound effect's influence on animation smoothness perception	119
7.4	Discussion	121
8	Conclusions and Future Work	123
8.1	Contributions	130
8.2	Impact	131
8.3	Directions for future work	132
8.4	Final remarks	134
	References	135
9	Appendix A: Additional materials from the study presented in Chapter 5	156
10	Appendix B: Additional materials from the study presented in Chapter 6	164
11	Appendix C: Additional materials from the study presented in Chapter 7	169

List of Figures

1.1	Rasterisation: 3D object projection onto a 2D image plane.	2
1.2	An example of different rendering techniques: plain rasterisation (left) and ray tracing technique using Photon mapping. Additional features are included, such as global illumination and caustics (right).	3
1.3	Ray tracing: ray traversal through pixels on the image plane.	3
2.1	The human eye. From [Bri11b]	11
2.2	The electromagnetic spectrum with visible wavelengths. After [Kai11]	12
2.3	Rods and cones normalised spectral sensitivity. From [Sch06].	13
2.4	Rods and cones distribution across the retina. After [BS06].	13
2.5	The top projection of the optic nerves stretching from the eyeballs to the visual cortex. From [Bri11c]	14
2.6	An illustration of wave creation: The number of molecules displaced by a vibration determines the amplitude of a sine wave. After [Alt04].	19
2.7	Sinusoidal representation of a wave: After the string is released, molecules bump into each other, creating a compression. Subsequently, the string moves inwards pulling the molecules away from each other - rarefaction. Inspired by Figure 2.1 from [Alt04].	20

2.8	Two sound waves: The first wave is in phase; the second wave is 90° out of phase. After [Alt04].	21
2.9	Wave interference: w1 and w2 interfering waves; w1+w2 resultant wave.	22
2.10	Response of the human ear to different frequencies. Inspired by Figure 2.3 from [Alt04].	23
2.11	Sound envelope components: Attack, Initial decay, Sustain and Release (Decay).	24
2.12	The anatomy of the human ear. From [Bri11a].	25
2.13	Illustration of the human cochlea.	27
2.14	Bianural cues: Interaural Intensity Difference (IID) and Interaural Time Difference (ITD).	30
3.1	Physically-based rendering examples: images rendered using Path tracing method (<i>courtesy of Piotr Dubla</i>) (top left and bottom right); an image rendered using Radiance rendering package [War94] (top right); an image rendered using Mental Ray (<i>courtesy of Jassim Happa</i>) (bottom left).	38
3.2	Non-physically-based rendering: design concepts (left and middle) and cultural heritage virtual reconstruction example (right). . . .	39
3.3	Light reflectance: diffuse (left), specular (middle) and glossy (right). 41	
3.4	The bidirectional reflectance distribution function (BRDF). After [PH10].	42
3.5	Traditional rasterisation pipeline.	43

3.6	Backward ray tracing. A ray is shot through a pixel. At the intersection point, the ray is spawned: shadow rays S1 and S2 are shot towards light sources; reflectance ray R1 is shot off the surface at the calculated angle; transmission ray T1 is shot through the surface. The process is repeated recursively for R1 and T1.	46
3.7	An example of path traced image (<i>courtesy of Piotr Dubla</i>).	48
5.1	Example of the slide sequence from the experiment.	70
5.2	Scenes used for the experiment: Checkerboard (top left), Corridor (top right), Kalabsha (bottom left), Library (bottom right). See Appendix A for larger images.	71
5.3	White noise frequency spectrum.	73
5.4	An example of the VDP comparison. Top: mask image; Middle: target image; Bottom: difference map with probability of detection - green:0-50%; yellow: 50-75%; red:75-95%; pink:95-100%. See Appendix A for large images.	74
5.5	The frequencies of the participants' scores across sound conditions.	79
5.6	The frequencies of the participants' scores across quality conditions (RaysPerPixel).	80
6.1	A sample frame from each of the animations used in the study. See Appendix B for larger images.	90
6.2	Correlation between beat rates (bps) and frame rates (fps). The diagram shows the number of frames that fit within a beat.	93
6.3	An illustration of the frame rate concept used in the instructions.	93
6.4	Two frames from the sample animation.	94
6.5	Preview of the slider bar used in the experiment.	95
6.6	Mean values of subjective scores across static scenes with standard error. All frame and beat rates are pooled.	99

6.7	Mean values of subjective scores across frame rates with standard error. All static scenes and beat rates are pooled.	100
6.8	Mean values of subjective scores across static scenes with standard error. All static scenes and frame rates are pooled.	100
6.9	Mean values of subjective scores across static scenes and frame rates with standard error.	101
6.10	Mean values of subjective scores across static scenes and beat rates with standard error.	101
6.11	Mean values of subjective scores across frame rates and beat rates with standard error.	101
6.12	Mean values of subjective scores across dynamic scenes with standard error. All frame and beat rates are pooled.	102
6.13	Mean values of subjective scores across frame rates with standard error. All dynamic scenes and beat rates are pooled.	103
6.14	Mean values of subjective scores across beat rates with standard error. All dynamic scenes and frame rates are pooled.	103
6.15	Mean values of subjective scores across dynamic scenes and frame rates with standard error.	104
6.16	Mean values of subjective scores across dynamic scenes and beat rates with standard error.	104
6.17	Mean values of subjective scores across frame rates and beat rates with standard error.	105
7.1	Four frames taken from the walk-through animations. The top two frames are from the animations with camera moving from the corridor to the conference hall, and the bottom two from the animations where the camera is moving from the conference hall to the corridor. See Appendix C for larger images.	112

7.2	Camera path used for the animations (red). Four running animation sequences (yellow) and six walking animation sequences (blue) were used.	114
7.3	Oscillating camera motion along the vertical (z) axis.	114
7.4	The experimental procedure. From left to right: grey box, first animation, grey box, second animation and A/B evaluation screen.	115
7.5	Mean values of the compared running animation test pairs.	120
7.6	Mean values of the compared walking animation test pairs.	121

List of Tables

2.1	The basic properties of the human visual system.	16
2.2	The basic properties of the human auditory system.	21
5.1	Related sounds used for the experiment.	72
5.2	Rendering times for all scenes presented in seconds.	75
5.3	Variables used in the experiment.	77
5.4	Summary of the contingency table for the sound condition. Test-Better shows the count of responses preferring the test image over gold standard one. GoldBetter is the count of responses preferring the gold image over the test one.	78
5.5	Summary of the contingency table for the rpp condition. Test-Better shows the count of responses preferring the test image over gold standard one. GoldBetter is the count of responses preferring the gold image over the test one.	80
5.6	“No sound” group: Chi-Square Analysis (df=1; critical value 3.841 at 0.05 level of significance). Significant results are written in bold.	81
5.7	“Noise” group: Chi-Square Analysis (df=1; critical value 3.841 at 0.05 level of significance). Significant results are written in bold. .	82

5.8	“Related sound” group: Chi-Square Analysis (df=1; critical value 3.841 at 0.05 level of significance). Significant results are written in bold. Results show that the subjects were looking more closely, and were able to find more differences.	82
5.9	VDP comparison of perceivable differences for all scenes showing the percentage of pixels on which the probability of perceiving that difference is the highest.	84
7.1	The details of the experimental design for each test. Numbers represent frame rate, while “r” and “w” stand for <i>running</i> and <i>walking</i> animations respectively.	110
7.2	Test 1: Observed and expected frequencies for the Run - Walk animation smoothness perception comparison.	117
7.3	Test 2: Mean and p values for <i>Audio</i> condition. <i>p - value</i> is given for difference in preference between the test pairs and 60r-60w condition. Lower and upper bounds were 1 (first animation preferred) and 2 (second animation preferred) respectively. *Not inline with a 1-tailed test.	118
7.4	Test 3: Mean and p values for <i>Audio</i> condition. <i>p - value</i> is given for difference in preference between the test pairs and 60vs60 condition. Lower and upper bounds were 1 (first animation preferred) and 2 (second animation preferred) respectively.	119
8.1	The cross-modal phenomena found in psychology and the main studies within the computer graphics that were inspired by this work.	124

Acknowledgements

First and foremost I would like to thank God, the Almighty, for having made all this possible by giving me health, strength and courage to do this work. My deepest gratitude to my supervisor Alan Chalmers, for inviting me to Warwick and his continuous encouragement and support he demonstrated during my study. Many thanks to Kurt Debattista who was there whenever I needed him, providing me with invaluable comments and advice, which greatly contributed to my thesis completion. A special thanks to Selma Rizvic, my undergraduate and masters supervisor, who introduced me to Alan and has been a great friend and support during this journey.

Kurt, Anna and Alex you were like a family! Spending time with you over dinners, coffees and football matches made me feel like at home. I know that I have friends for life. Please don't ever forget the pact we made.

I was so lucky to have such a great group here in Warwick. Vibhor, you are a great guy, always being there to help. Piotr, Matt and Gabriela thanks for being so patient with all my questions and for help with my PhD. I think we have done some significant work together. Tom, Jass, Carlo and Alessandro you are wonderful friends and enhanced my life in UK. Silvester, Francesco, Alena, Elena, Remi, Mike it was terrific having you here. Thank you all.

It was great having the other Bosnians at Warwick. Belma, you were the best flatmate, being so kind and supportive all those years. Elmedin, you are an awesome person, "jaran" who could understand me when no one else did. Thank you both for everything!

Thanks to all the volunteers who participated in the experiments. Without their help and patience this study would have never reached any conclusions. I would also like to express my gratitude to the University of Warwick for funding my research through the Vice Chancellor's scholarship and all the people from the University who assisted me in various ways during my study. Many thanks to all the people from ISOC, who have made my stay in UK special.

There was a person whom I knew for only three months, but who has left an indelible mark in my life. Always smiling, willing to help, enjoying every thing he did and loving every person he knew. A very unique person, whose peculiarity can not be described in only a few sentences... Usama, you will always be remembered.

My family has always supported me through all my endeavours. Without your endless support and love this would have not been the same. I will never be able to thank you for everything you have done for me.

In memory of Usama Monsour.

Declaration

The work in this thesis is original and no portion of work referred to here has been submitted in support of an application for another degree or qualification of this or any other university or institute of learning.

Signed:

Date: 27 May 2011

Vedad Hulusić

List of Publications

The following publications are the result of the work summarised within the thesis:

Journal papers

- **Hulusić, V.**, Debattista, K., Aggarwal, V., Chalmers, A. (2011): Maintaining frame rate perception in interactive environments by exploiting audio-visual cross-modal interaction. *The Visual Computer*.
- **Hulusić, V.**, Debattista, K., Dubla, P., Bessa, M.E.C., Chalmers, A. (2011): Smoothness perception: Investigation of beat rate effect on frame rate perception, submitted to the Computer Graphics Forum, second round of review.

Peer-reviewed Conference Papers

- **Hulusić, V.**, Aranha, M., Chalmers, A. (2008): The influence of cross-modal interaction on perceived rendering quality thresholds, in WSCG 2008 Full Papers Proceedings, The 16th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, Plzen -

Bory, Czech Republic

- **Hulusić, V.**, Czanner, G., Debattista, K., Sikudova, E., Dubla, P., Chalmers, A. (2009): Investigation of the beat rate effect on frame rate for animated content, in Spring Conference on Computer Graphics 2009, Bratislava, Slovak Republic
- **Hulusić, V.**, Debattista, K., Aggarwal, V., Chalmers, A. (2010): Exploiting audio-visual cross-modal interaction to reduce computational requirements in interactive environments, Proceedings of the IEEE conference on Games and Virtual Worlds for Serious Applications, IEEE Computer Society, VS-Games 2011, Braga, Portugal.
- **Hulusić, V.**, Harvey, C., Tsingos, N., Debattista, K., Walker, S., Howard D., Chalmers A. (2011): Acoustic Rendering and Auditory-Visual Cross-Modal Perception and Interaction, State of the Art Reports (STAR), Eurographics 2011.

Additionally, my research has been incorporated with the work of other researchers from the group. As a result of such work following has been published:

- Happa, J., Znyi, E., Hulusić, V., Chrysanthou, Y., Chalmers, A. (2008): The High-Fidelity Computer Reconstruction of Byzantine Art in Cyprus, IV International Cyprological Congress, Lefkosia, Cyprus.
- Happa, J., Znyi, E., Hulusić, V., Chrysanthou, Y., Chalmers, A., (2008): Capturing and Visualising the Past: High-Fidelity Computer Reconstruction of Byzantine Artefacts, Presented at the World Archaeological Congress, Dublin, Ireland.
- Happa, Jassim., Artusi, A., Dubla, P., Bashford-Rogers, T., Debattista, K., Hulusić, V., Chalmers, A. (2009): The Virtual Reconstruction and

daylight illumination of the Panagia Angeloktisti, 10th VAST International Symposium on Virtual Reality, Archaeology and Cultural Heritage, VAST 2009, St. Julians, Malta.

Finally, I have been working on related topics, such as virtual reconstruction of cultural heritage, which resulted in following peer-reviewed conference papers:

- Rizvić, S., Sadžak, A., Ramić-Brkić, B., Hulusić, V. (2011): Virtual Museum Applications and Their Public Perception in Bosnia and Herzegovina, in 4th International Workshop 3D-ARCH 2011: 3D Virtual Reconstruction and Visualization of Complex Architectures 2011, Trento, Italy.
- Hulusić, V., Rizvić, S. (2011): The use of live virtual guides in educational applications, Proceedings of the IEEE conference on Games and Virtual Worlds for Serious Applications, IEEE Computer Society, VS-Games 2011, Athens, Greece

Abstract

Generating high-fidelity images in real-time at reasonable frame rates, still remains one of the main challenges in computer graphics. Furthermore, visuals remain only one of the multiple sensory cues that are required to be delivered simultaneously in a multi-sensory virtual environment. The most frequently used sense, besides vision, in virtual environments and entertainment, is audio. While the rendering community focuses on solving the rendering equation more quickly using various algorithmic and hardware improvements, the exploitation of human limitations to assist in this process remain largely unexplored.

Many findings in the research literature prove the existence of physical and psychological limitations of humans, including attentional, perceptual and limitations of the Human Sensory System (HSS). Knowledge of the Human Visual System (HVS) may be exploited in computer graphics to significantly reduce rendering times without the viewer being aware of any resultant image quality difference. Furthermore, cross-modal effects, that is the influence of one sensory input on another, for example sound and visuals, have also recently been shown to have a substantial impact on viewer perception of virtual environment.

In this thesis, auditory-visual cross-modal interaction research findings have been investigated and adapted to graphics rendering purposes. The results from five psychophysical experiments, involving 233 participants, showed that, even in the realm of computer graphics, there is a strong relationship between vision and audition in both spatial and temporal domains. The first experiment, investigating the auditory-visual cross-modal interaction within spatial domain, showed that unrelated sound effects reduce perceived rendering quality threshold. In the following experiments, the effect of audio on temporal visual perception was investigated. The results obtained indicate that audio with certain beat rates can be used in order to reduce the amount of rendering required to achieve a perceptual high quality. Furthermore, introducing the sound effect of footsteps to walking animations increased the visual smoothness perception. These results suggest that for certain conditions the number of frames that need to be rendered each second can be reduced, saving valuable computation time, without the viewer being aware of this reduction. This is another step towards a comprehensive understanding of auditory-visual cross-modal interaction and its use in high-fidelity interactive multi-sensory virtual environments.

CHAPTER 1

Introduction

High-fidelity rendering in real-time still remains one of the greatest challenges in computer graphics. Despite substantial improvement in the performance of general and dedicated graphics hardware, it is still not possible to generate high-fidelity images of complex scenes on a single machine in real-time. Furthermore, in virtual environments, such as video games, stimulation of auditory, and possibly some other senses, is often also desirable to improve the user experience, since we are used to environments where multiple senses are stimulated simultaneously. This, however, does not necessarily need to be considered as an additional work load, but instead, can be exploited, so that the overall work load is balanced or even reduced, without any perceivable loss in quality. This is possible due to various limitations of the Human Sensory System (HSS). One such limitation is the influence of one sensory input on another, commonly termed cross-modal interaction. In this thesis, auditory-visual interaction is investigated, adapting it for enhancing graphics rendering, while maintaining perceptual quality.

1.1 High-fidelity rendering

Rendering is the process of digital image synthesis of virtual scenes for consumption by current displays - typically in the form of a 2D image. This can be done using different methods with various levels of complexity, and therefore achieving higher or lower image fidelity. A complete calculation of light propagation contributed by each photon individually would be impractical. Therefore, different techniques have been developed for approximating this process, making it achievable on available hardware resources. A coarse image representation can be achieved by straightforward projection of a 3D object on a 2D image plane, see Figure 1.1. This technique is known as rasterisation, see Figure 1.2.

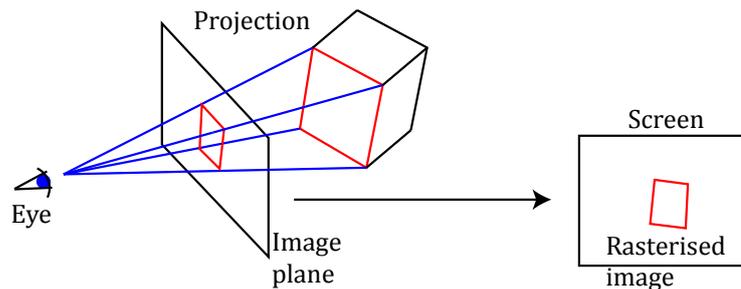


Figure 1.1: Rasterisation: 3D object projection onto a 2D image plane.

Another method, considered more physically accurate, is called ray tracing, where light intensity of each pixel is calculated by traversing rays from a virtual camera through each pixel, and calculating the light contribution from the light sources and surrounding objects at the ray's point of intersection with the virtual scene, see Figure 1.3.

Physically-based rendering aims at solving the rendering equation [Kaj86], see Section 3.3. This calculation considers physical properties of the materials, calculated using Bidirectional Reflectance Distribution Functions (BRDFs). The resultant image is physically correct (see Figure 1.2), but the computational time

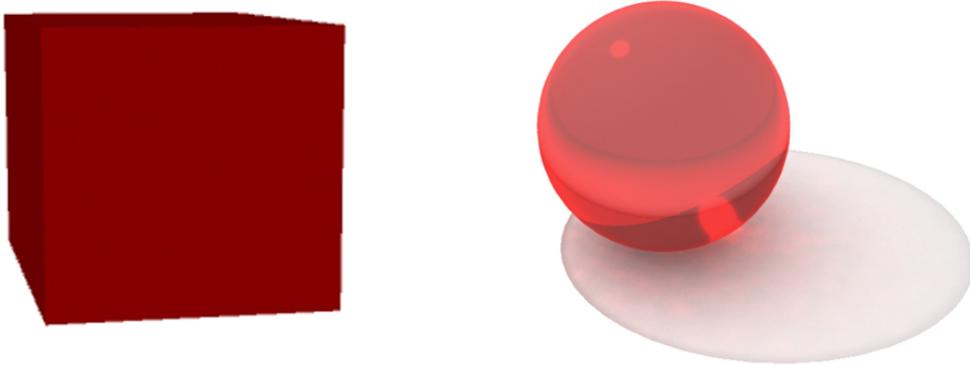


Figure 1.2: An example of different rendering techniques: plain rasterisation (left) and ray tracing technique using Photon mapping. Additional features are included, such as global illumination and caustics (right).

is rather high, making the rendering technique inappropriate for real-time applications. Therefore, the computer graphics research community has worked on improving rendering algorithms and hardware enhancements for several decades, utilising various alternatives including human sensory and attentional limitations.

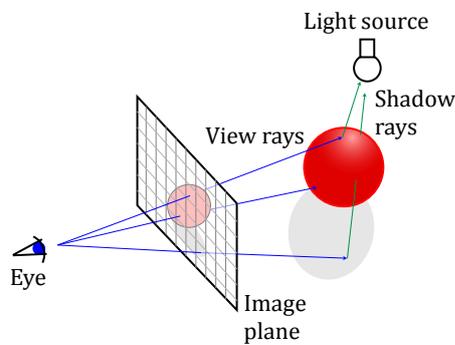


Figure 1.3: Ray tracing: ray traversal through pixels on the image plane.

1.1.1 Applications

Computer graphics have been employed for a wide variety of purposes. However, its utilisation has been restricted due to the inability to render graphics at required spatial or temporal quality. With the advances of rendering hardware,

such as multi-core Central Processing Units (CPUs) and Graphics Processing Units (GPUs), and significant enhancements of rendering algorithms, many limitations have been overcome. This has increased the render quality and expanded the use of computer graphics, which nowadays can be found in many applications, such as:

- **Virtual Worlds:** Recently virtual worlds have become increasingly popular. Second Life, probably the most popular amongst immersive 3D virtual worlds, allows users to socialise, connect and interact with other users in the virtual world, using text or voice chat. It can also be used for training and education [DLFPT09].
- **Computer Games:** One of the leading entertainment industries, along with film industry, is computer games production. The game industry is a significant factor in pushing the limits of the graphics hardware and rendering techniques. This results in a continuous improvement of the user immersion and graphics quality [Mit07].
- **Visual Effects:** The majority of current films use visual effects extensively [TL04]. This can range from applying virtual lighting in a scene to adding virtual characters or particles in form of fire, explosion, water, etc.
- **Data Visualisation:** Computer graphics can also be used for visualising various data when communicating information to a target audience [PNB03]. This data is abstracted and represented with an appropriate visual form.
- **Architecture:** Rendering is frequently employed in architecture for making previews or producing realistic appearance of the design concepts. For the latter, the physically-based rendering is usually used, in order to simulate realistic lighting conditions in a designed space [BHWL99]. This improves

the building process, allowing for design improvements before actually starting the physical construction.

- **Training Simulators:** A good training simulator needs to provide a high level of immersion, which is achievable through high-fidelity stimuli [RMP04]. Its main purpose is to recreate a real life situation and simulate it as if the user was there. The deliverables usually include graphics, audition, haptics, olfaction, and other stimuli.
- **Medicine:** Computer graphics plays an important role in medicine. It is used for medical imaging, such as Computer Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET) and other similar techniques.
- **Archaeology:** Many archaeological sites contain only partial remains of the original constructions and artefacts. Furthermore, very often it is not possible to physically rebuild them and present an authentic appearance. Therefore, virtual reconstructions are made, where 3D representations of the original site are created, giving the opportunity to interact with the model e.g. walk around/through a building, examining artefacts, etc. [HMD*10].

1.2 Cross-modal interaction

The computational expense of creating virtual environments, particularly those involving multiple senses, has led researchers to explore the interactions between different senses within the sphere of the Human Sensory System (HSS). Despite being extremely complex, the HSS is not perfect and it has certain limitations. In particular, limited attentional resources ensure that the HSS cannot attend to all aspects of all the senses concurrently [Jam92]. Perception across modali-

ties is major topic of research in psychology, for example [GM59, Rec03, SS04]. Based on such work, one particular cross-modal effect, investigated in the field of computer graphics is that of the vision and audition [Mas06]. Those results are the initial step towards cross-modal interaction understanding and its utilisation within computer graphics. Since it carries significant potential for enhancing graphics rendering, it needs to be investigated further, which is the topic of this thesis. There are multiple aspects of this phenomenon that have to be carefully studied, in order to be able to design a framework that could adjust various parameters of both auditory and visual stimuli on demand, thus speeding up the rendering process, without a degradation in overall perceived quality. This is possible by tackling all those aspects, such as user experience and immersion, scene dynamics, subjective emotional factors, auditory-visual stimuli relationship, spatio-temporal perceptual influences, etc. individually.

1.3 Research aim and objectives

Until now, the auditory-visual cross-modal interaction in computer graphics has been narrowly investigated, giving an indication of a potential that could benefit the rendering community in the field of computer graphics. The main aim of this thesis is to further explore the direct relationship between vision and audition in both the spatial and temporal domains and provide results that, if harnessed correctly, should make it possible to have a graphics engine that can adjust the audio and visual quality on-demand, to reduce or balance its work load whenever required. This would effectively reduce the computational time of rendering, without the user noticing any perceptual loss in quality.

The research objectives are as follows:

- to provide a comprehensive literature review on auditory-visual cross-modal

interaction, give an overview of the anatomy and limitations of the HSS and main findings and phenomena related to the human attention and perception, all structured in a meaningful framework;

- to investigate the limitations of the human visual system and the impact cross-modal interactions have on perceivable rendering thresholds, in order to speed up the rendering process by reducing the spatial image quality, with perceivable difference in fidelity;
- to investigate the effect of scene related and unrelated audio on spatial visual perception;
- to investigate the effect of audio beat rate on video frame rate in order to decrease the work load by rendering less frames, without any degradation in perceivable rendering quality;
- to investigate the effect of movement related sound effects, such as the sound of footsteps in a walking/running animation, on temporal visual perception;
- to investigate additional effects, such as camera movement speed, familiarity with computer games and/or animation, scene complexity on frame rate perception in computer animations;
- to provide a groundwork and direction for developing an adaptive rendering framework for high-fidelity graphics in real-time.

1.4 Thesis outline

The thesis is organised as follows:

- **Chapter 2: Human Sensory System** provides an overview of features and limitations of the human sensory system, attention and perception.

- **Chapter 3: High-Fidelity Rendering** presents the main rendering techniques, relevant to the work presented in the thesis.
- **Chapter 4: Cross-Modal Interaction** covers the background on cross-modal interaction in both psychology and computer graphics, focusing on auditory influence on vision and visual influence on audition in both auditory and visual rendering.
- **Chapter 5: The Influence of Cross-Modal Interaction of Perceived Rendering Quality Thresholds** describes the study on auditory influence on perceived rendering quality threshold for static images.
- **Chapter 6: Beat Rate Effect on Frame Rate Perception** provides evidence of cross-modal interaction and the potential of utilisation of auditory influence on temporal visual perception for rendering enhancement.
- **Chapter 7: Exploiting Audio-Visual Cross-Modal Interaction** demonstrates the difference in perception of slow and fast animations. Additionally, the influence of the related auditory effects on visual perception is examined and results are presented.
- **Chapter 8: Conclusions and Future Work** concludes the thesis and gives some directions for potential future work.

CHAPTER 2

Human Sensory System

Humans are constantly surrounded by multi-modal stimuli, both relevant and irrelevant. In order to interpret an environment - its content and/or relationships between its elements, we use our external senses: vision, audition, smell, taste, touch, temperature, proprioception and the vestibular system, etc. In addition, the internal senses, e.g. memory, imagination, thirst, hunger, fatigue, etc. are used for informing us about changes within our body. Our sensory organs receive the physical stimulation, which is then transformed into neural signals, and finally interpreted by the brain. However, it is not only the reception of inputs that creates perception. Due to the internal factors, a significant amount of preprocessing and reprocessing is required in order to get the complete perceptual output. This chapter will cover the basics of vision and audition, and the most relevant limitations, including attentional and perceptual shortcomings, that might be utilised in computer graphics for enhancing auditory and visual rendering.

2.1 Human visual system

The human visual system (HVS) is a complex system responsible for receiving the light from the environment, transmitting it through the visual pathways and processing it in the visual cortex. Due to its high importance in computer graphics, the main aspects of the anatomy and functionality of the HVS will be explained in this chapter. More details can be found in [Roo02,BS06,Kai11].

2.1.1 The anatomy of the HVS

The HVS comprises three major parts: the eye, visual pathways and visual cortex. Each part has its own functionality and relies on the functionality of the other two.

The eye

The human eye is the external organ of the HVS, and it functions as an interface to the outer world, see Figure 2.1.

It has a nearly spherical shape and a diameter of approximately 24mm. The outermost layer of the eye is the *sclera*, whose major function is to protect the eyeball. The layer below is the *choroid*, which nourishes the eye cells. The innermost and the most important layer - the *retina*, captures the light and initiates the signal transmission.

The front, transparent, bulged part of the eye is called the *cornea*. When light passes through the cornea, it enters the *lens* through a small opening called the *pupil*. The size of the pupil is mostly determined by the incoming amount of light. The size is controlled by the two sets of muscles, one for decreasing and one for increasing the opening. The size of the pupil also affects the depth of field, similarly to the aperture size in photography. The pupil opening leads to the

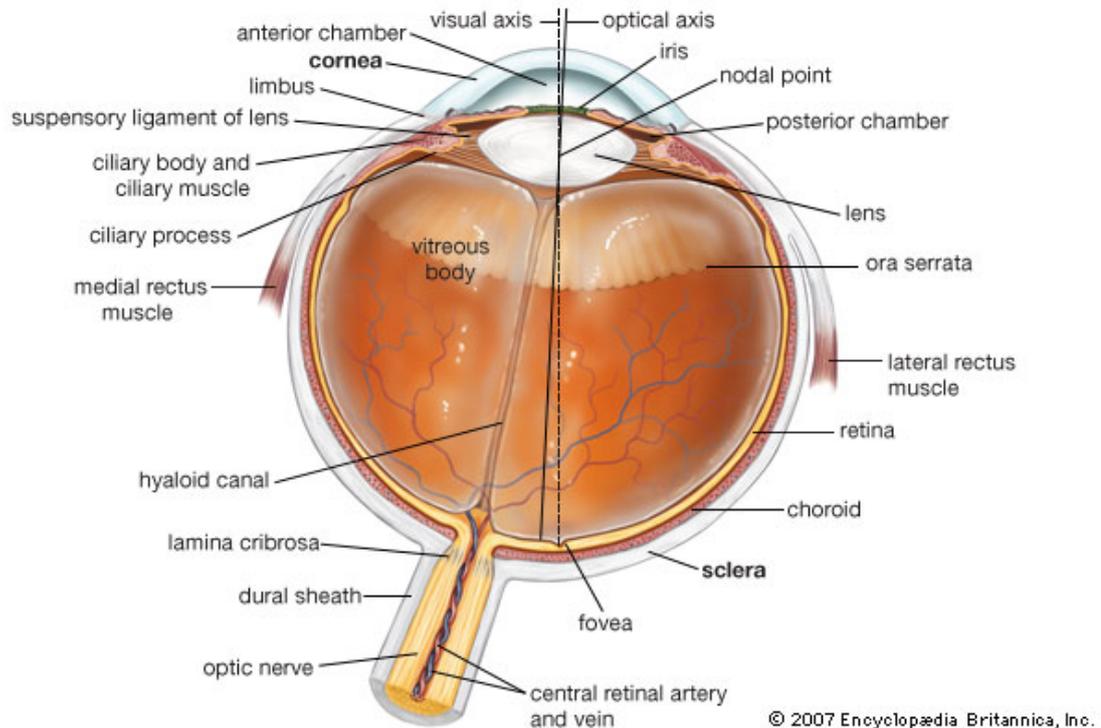


Figure 2.1: The human eye. From [Bri11b]

crystalline lens. The lens is a transparent structure, about 9mm in diameter and 4mm in thickness. Along with the cornea, its function is to refract the incoming light, focusing the light beam onto the retina. The lens can change its shape in order to change the focal distance and focus on an object at a certain distance. This process is called accommodation [BS06].

The retina

The retina is made of the same tissue as the brain and it represents the extension of the central nervous system. The retina itself is comprised of a few layers, but the focus will be kept only on the layer consisting of the photoreceptor cells. There are two types of photoreceptors in humans: *rods* and *cones*, sensitive to wavelengths from about 400 to 700 nm, see Figure 2.2.

In each eye there are about 5 million cones and 100 million rods. The cones

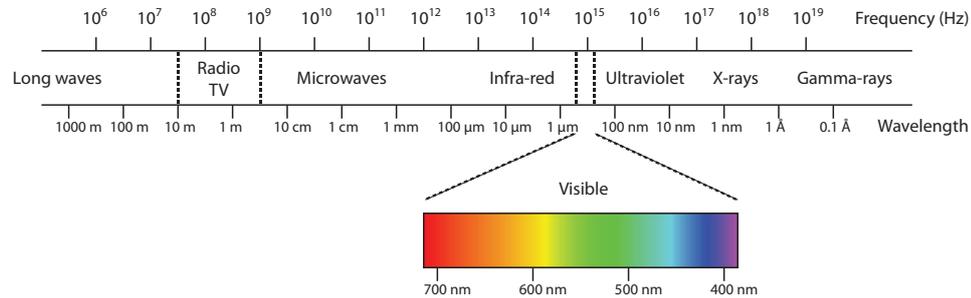


Figure 2.2: The electromagnetic spectrum with visible wavelengths. After [Kai11]

are responsible for colours and they are mostly concentrated in the *fovea*, a small region of the retina with the highest visual acuity. There are three different types of cones (L, M and S-cones), each absorbing different wavelengths, corresponding to red (564nm), green (534nm) and blue (420nm) respectively, see Figure 2.3. The rods, on the other hand, are mainly sensitive to light and benefit vision in low light conditions. The rods are concentrated around the fovea and their density decreases towards the periphery of the eye, see Figure 2.4.

The photoreceptors are connected to the *ganglion cells*, which transmit visual information from the retina to the visual cortex in the brain. The place where the blood vessels enter the eye and where the ganglion cell axons exit the eye forming the optical nerve is called the *optic disc*. The optic disc contains no photoreceptors and therefore, creates “the blind spot”.

The visual pathways

The main objective of the eye is to capture the light. This light is transformed into neural signals and transmitted to the brain. Each eye has its own *optic nerve* constructed from the axons of the ganglion cells. The nerves exit the eyeball through the optic disc. After approximately 5cm they meet at the place called *optic chiasm*, where some fibers cross (contralateral fibers) and some remain on

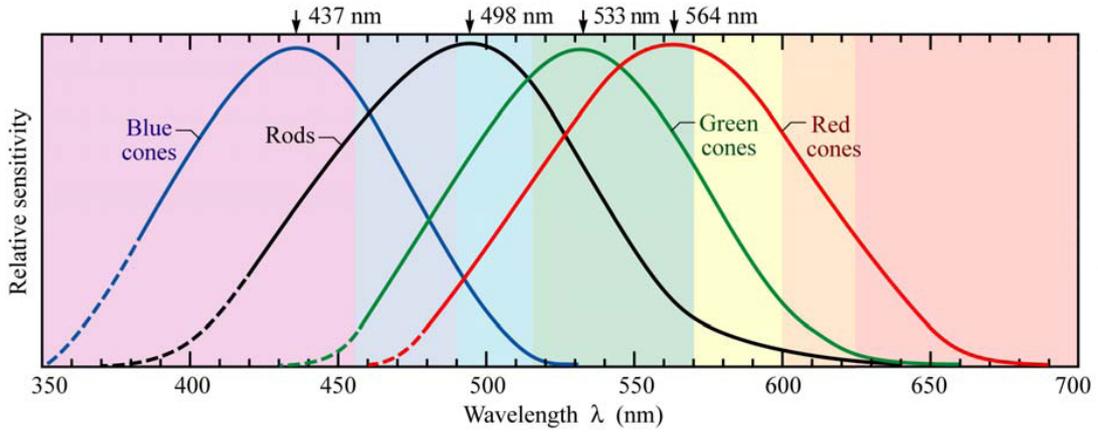


Figure 2.3: Rods and cones normalised spectral sensitivity. From [Sch06].

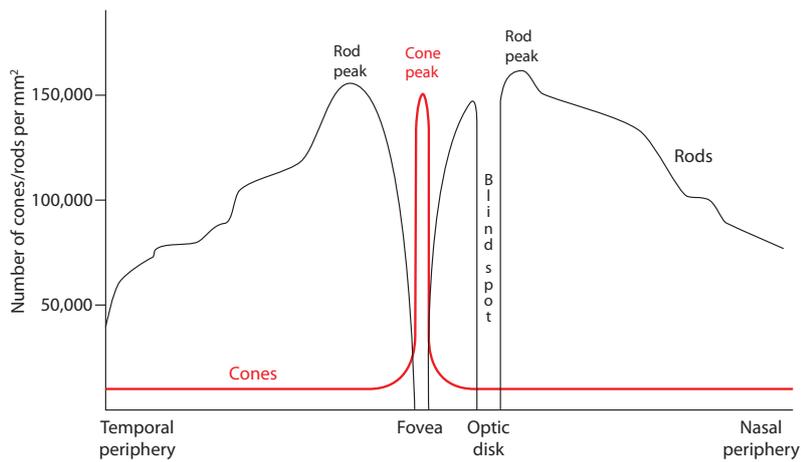


Figure 2.4: Rods and cones distribution across the retina. After [BS06].

the same side of the brain (ipsilateral fibers), see Figure 2.5.

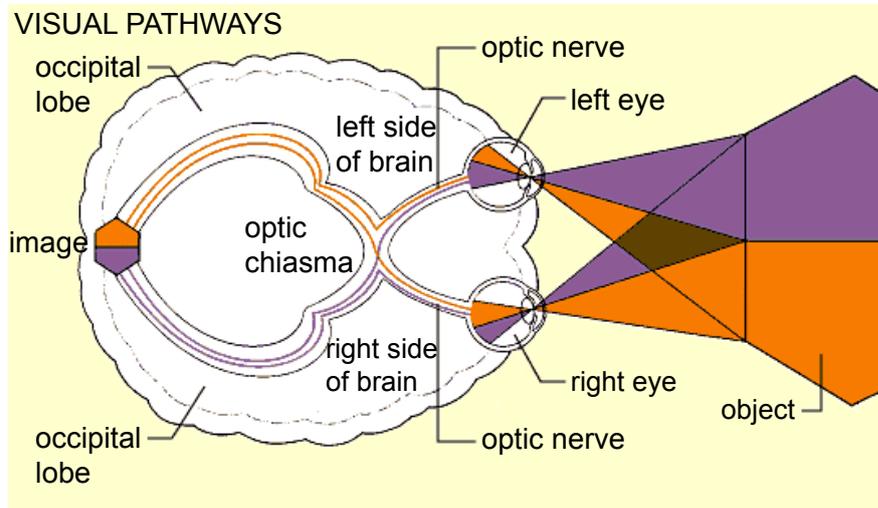


Figure 2.5: The top projection of the optic nerves stretching from the eyeballs to the visual cortex. From [Bri11c]

From the figure we can see that the information from the right visual field, falling on the left side of both eyes' retinas, will be carried and processed by the left hemisphere of the brain, and the information from the left visual field, will be transmitted and processed by the right hemisphere.

The visual cortex

The visual cortex is the main part of the brain responsible for the interpretation of the information covered from the retinal input. It is comprised of several cortical areas: the primary visual cortex (V1), V2, V3, V4 and V5. Each area has a specific function for e.g. contrast and orientation detection, motion recognition, pattern recognition, etc. Besides the visual cortex, there are other parts of the brain that contribute to human vision. One such part is the *superior colliculus*, which represents a primitive visual area, capable of detecting if the eyes see something, but not what that actually is. This area also receives auditory information from the ears, if it originates from the same location where the visual stimulus is

detected, increasing the perceptual experience. Therefore, these cells are called the *multisensory cells*.

2.1.2 Visual perception

As mentioned earlier, the highest concentration of the photoreceptors in the eye is in the foveal region. Consequently, this region has the highest visual acuity, and moving further from the fovea the acuity rapidly decreases. The phenomenon of the *foveal vision* is also known as the *internal spotlight* [Jam90, HB89]. The area of the foveal vision covers only 2 degrees of the visual field. This low angular sensitivity is compensated by the rapid eye movements called *saccades*. These movements are extremely fast (up to 600 degrees per second) and last from 10 to 100ms. However, most of the viewing time is spent on *fixations*, whose length ranges from 150 to 600ms.

Human perception is rather complex and helps guide the cognitive aspect of seeing. It is constructive and content dependent, grouping things together and looking at the elements with regards to their surrounding features, giving sense to what is observed. Visual perception can be divided into two parts: spatial and temporal perception.

Spatial visual perception

As described in [BS06], there are three stages in seeing things: *detection*, *discrimination* and *identification*. When we need to detect if a mug is on the table, we use the first, lowest level process - detection. If we need to distinguish the blue mug from a set of mugs, we need to use discrimination. Furthermore, if we need to find a specific mug, we use a process of identification. However, all these processes happen very quickly and automatically, without us being aware of it. In order to perform the process of identification, our brain needs to process the

input information and provide a feedback which initiates our reactions.

Wavelength sensitivity	Spatial acuity	Temporal acuity
400nm - 700nm	visual angle of 1 minute [BS06]	26Hz [FN05]

Table 2.1: The basic properties of the human visual system.

Spatial perception highly depends on visual attention (discussed in Section 2.3). However, there are some other factors, such as spatial frequency, which might influence the perception [LM00]. In computer graphics, the spatial frequency is particularly important, as it directly affects the level of details or the image sharpness. In humans, a visual angle of one minute is the threshold of spatial visual acuity, see Table 2.1.

Temporal visual perception

Numerous stimuli that we are exposed to on a regular basis, such as TV, computer and mobile displays, cinema, etc., are not continuous. They display visual content at certain rates called temporal frequency (also called just frequency or frame rate). For example, the standard cinematic frame rate is 24 fps (frames per second), standard TV frequency is 25 fps, HDTV works at 50 or 60 fps, while video games use even higher frame rates. Although the threshold of the temporal visual sensitivity has been shown to be 26Hz [FN05] (see Table 2.1), we perceive these stimuli as continuous thanks to the phenomenon called the *flicker fusion*. The reason for this is the *persistence of vision*, which is the ability of the retina to retain an image for a period of 1/20 to 1/5 a second after the exposure [Rog25].

The continuous appearance of the stroboscopic display, also called the apparent motion, where two or more distinct flashing stimuli are perceived as one dynamic stimulus, is explained in [SD83, AA93, SPP00, Get07]. Alterations in visual appearance over time can affect some other aspects of visual perception.

According to Bloch's law, for example, the duration of the stimulus can affect the perception of brightness, even for stimuli with the same luminance [MMC09].

2.2 Human auditory system

Sound is a perceptual experience caused by disturbance in air pressure, that results in the physical energy. In order to audible it requires a sound source and a perceiver [BS06]. Unlike the eyes, that can be shut to block incoming light, our ears are constantly exposed to sound. A sound can differ in many properties, such as location, loudness, rhythm, complexity, duration, etc. It is an important modality which helps us to learn about an environment and to identify surrounding objects and their features. It provides both cognitive and affective information. Furthermore, it has its visual component, resulting in an image creation in our mind[BS06].

This section presents the basic properties of sound, sound perception and the human auditory system. More details can be found in [Moo82,Yos00,Alt04,BS06].

2.2.1 Sound properties

Physical properties

Every sound is created by mechanical oscillations of an object in elastic medium, such as gas, liquid or solid. These oscillations increase pressure in a medium and compress molecules which start to vibrate and transfer the vibration to surrounding molecules. This results in a wave creation, see Figure 2.6.

A simple sound can be represented as a sine wave or a sinusoidal vibration, see Figure 2.7. This type of sound is also called a pure sound and is heard rather rarely in nature. The sinusoid is characterised by three properties: *frequency*, *starting phase* and *amplitude*.

Frequency Every sinusoidal oscillation has two extremes called *compression* and *rarefaction*. The two, on a sine wave, are the top and the bottom peaks of the wave respectively. The distance between two successive compressions or

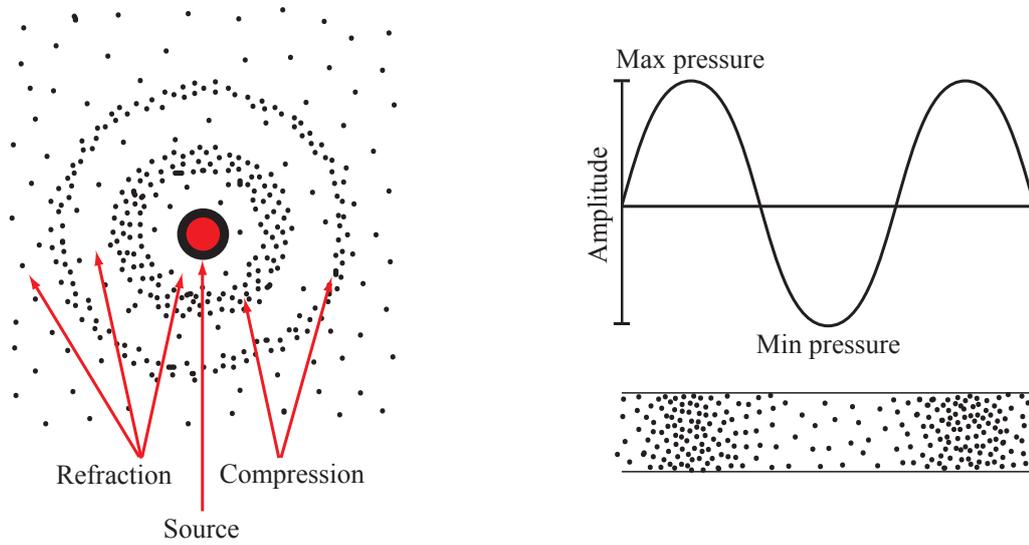


Figure 2.6: An illustration of wave creation: The number of molecules displaced by a vibration determines the amplitude of a sine wave. After [Alt04].

refractions is called the sound *wavelength* (λ). When a wave passes through both peaks and comes to its initial position, it has completed one *cycle*. The number of cycles that a wave passes through in a certain time is called *frequency*. In theory, humans can hear frequencies from 20Hz (20 cycles per second) to 20KHz (20,000 cycles per second), see Table 2.2. However, in practise, the frequency range of most humans is between 35Hz to 17KHz, which decreases further with aging. The time of one cycle is called *period* (P_r). The relationship between frequency and period is given in equation 2.1.

$$f = \frac{1}{P_r} \quad (2.1)$$

The speed of the wave propagation through a medium, called the speed of sound (c), affects the wavelength and can be represented as:

$$\lambda = \frac{c}{f} \quad (2.2)$$

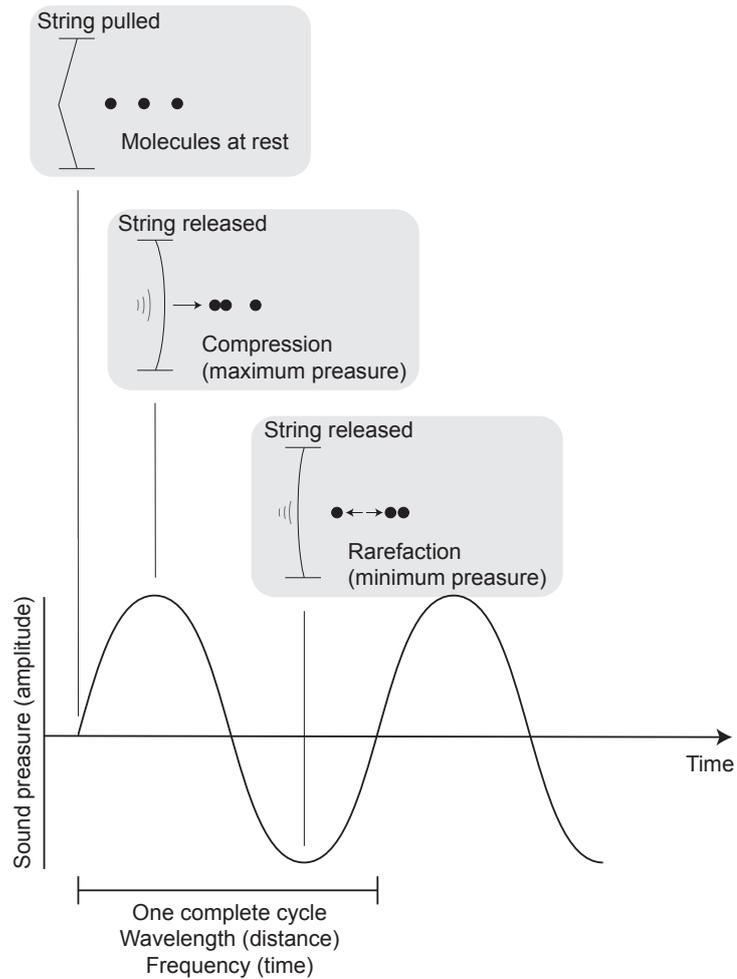
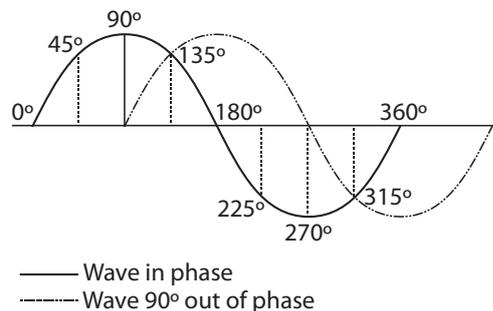


Figure 2.7: Sinusoidal representation of a wave: After the string is released, molecules bump into each other, creating a compression. Subsequently, the string moves inwards pulling the molecules away from each other - rarefaction. Inspired by Figure 2.1 from [Alt04].

Frequency sensitivity	Vertical spatial acuity (elevation)	Horizontal spatial acuity (azimuth)	Temporal acuity
20Hz - 20kHz	angle of 0.97° [PS90]	angle of 3.65° [PS90]	89.3Hz [FN05]

Table 2.2: The basic properties of the human auditory system.

Phase is relative difference in starting points of the cycles of two sound waves. If two waves start their cycles at the same time, they are *in phase*. Since one period corresponds to one full circle, phase is measured in degrees. Therefore, a sine wave which is late for half of the period has the phase of 180° , while the one which is late for a quarter of a period has the phase of 90° , see Figure 2.8. Some other sounds, called periodic sounds, could also have this repetitive feature, and therefore a phase can be defined for them.

Figure 2.8: Two sound waves: The first wave is in phase; the second wave is 90° out of phase. After [Alt04].

Amplitude represents the *intensity* of a vibration. The more molecules that are displaced, the higher the amplitude is. On the sinusoidal representation, amplitude is the difference between the maximum height (crest) and the depth (trough) of the wave.

Although pure tones exist, we are far more often exposed to complex sounds in environments. These sounds cannot be represented by a single sine wave, and they are rather irregularly shaped. If we combine two tones with slightly different

frequencies, the resultant sound wave will look like that shown in Figure 2.9. The wave looks as if it was a single sine wave with changing amplitude. Such a stimulus is called a *beating* sound, and it is perceived as a tone whose loudness is beating. However, this must not be mistaken with the *beat* used in music, which represents a basic time unit of a piece. In this thesis, the term *beat* will be used as in music.

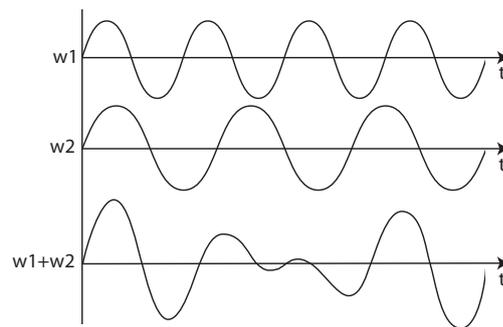


Figure 2.9: Wave interference: w_1 and w_2 interfering waves; w_1+w_2 resultant wave.

Complex sound waves can be decomposed into a series of sine waves with specific frequencies, phases and amplitudes using Fourier analysis. This analysis is used primarily in Digital Signal Processing (DSP) and is beyond the scope of the thesis. More details can be found in [Smi97].

Perceptual properties

The physical properties of a sound can be measured and controlled, but nonetheless, the same sound can be perceived differently by different individuals. This is because the perceptual properties are subjective and require psychophysical evaluation. The components that influence sound perception include: loudness, pitch, harmonics, timbre, rhythm, sound envelope and speed.

Loudness There are various parameters that can influence loudness percep-

tion. The human ear is not equally sensitive to different frequencies (Figure 2.10), and therefore tones presented at 100 Hz, 52 dB; 1000 Hz, 40 dB; and 4000 Hz, 37 dB are perceived as equal in loudness [Yos00]. This phenomenon is called *equal loudness principle*. Another factor, influencing the loudness perception is the *auditory masking*. If a tone at 1000 Hz is masked by a broadband noise, which is a sound perceived as equally loud at all frequencies, its hearing threshold will be shifted for 40 dB. Finally, the duration of the exposure to a tone modifies its loudness perception. This phenomenon, where a sound loudness level is perceived as lower by the time being presented, is called the *loudness adaptation*. This happens because our auditory system automatically adjusts to the noise level in order to protect itself.

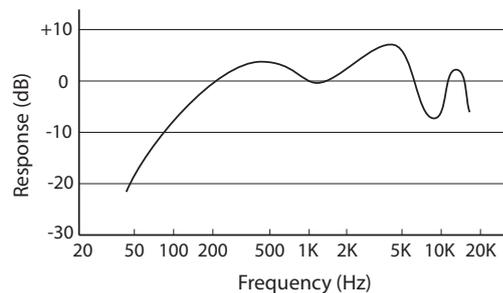


Figure 2.10: Response of the human ear to different frequencies. Inspired by Figure 2.3 from [Alt04].

Pitch is a relative, perceived tonal lowness or highness of a sound, mostly defined by frequency. Frequencies can be divided into three groups: low, midrange and high. For a pure tone, the higher the frequency, the higher a pitch is perceived. However, a complex sounds, such as a note played on a piano, comprises of one predominant frequency called the *fundamental frequency* or the *first harmonics* and its resonant frequencies - *harmonics*. However, if we remove the fundamental frequency, we will still perceive the same pitch.

Timbre is a sound feature that defines the *tone colour*. A sound with the

same pitch, duration and loudness can have different timbre. For example, the same tone played on a piano, guitar or a trumpet will have a distinct sound.

Sound envelope is another factor that influences the timbre of a sound. It is usually represented with three or four stages, see Figure 2.11. *Attack* is the way a sound is initiated by the sound source. *Initial decay* follows the attack and is a slight decrease in the sound intensity. *Sustain* is the stage of a relatively constant dynamics of the audio. After the sound source becomes inactive, the sound starts to diminish. The time from that point until complete silence is termed *release* or *decay*.

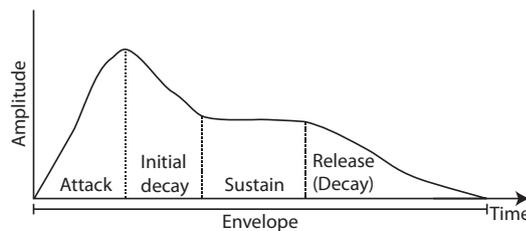


Figure 2.11: Sound envelope components: Attack, Initial decay, Sustain and Release (Decay).

Rhythm Any music can be divided into 2 components: melody and rhythm. Both of them are highly important for “communication” via this medium. Rhythm in music refers to the temporal patterning of sound, and involves beat (pulse), pace (beat rate, tempo) and pattern. It can exist without melody and can be found in a number of contexts, such as heart beat, walking, speech, drums, etc.

2.2.2 Peripheral auditory anatomy

The Human Auditory System (HAS) comprises three parts: the ears; the auditory nerves; and the brain. The ear consists of the outer ear, middle ear and inner ear, see Figure 2.12. In this subsection the functionality of hearing in humans

and the ear's anatomy will be briefly explained. More details can be found in [Moo82, Yos00, Alt04, BS06].

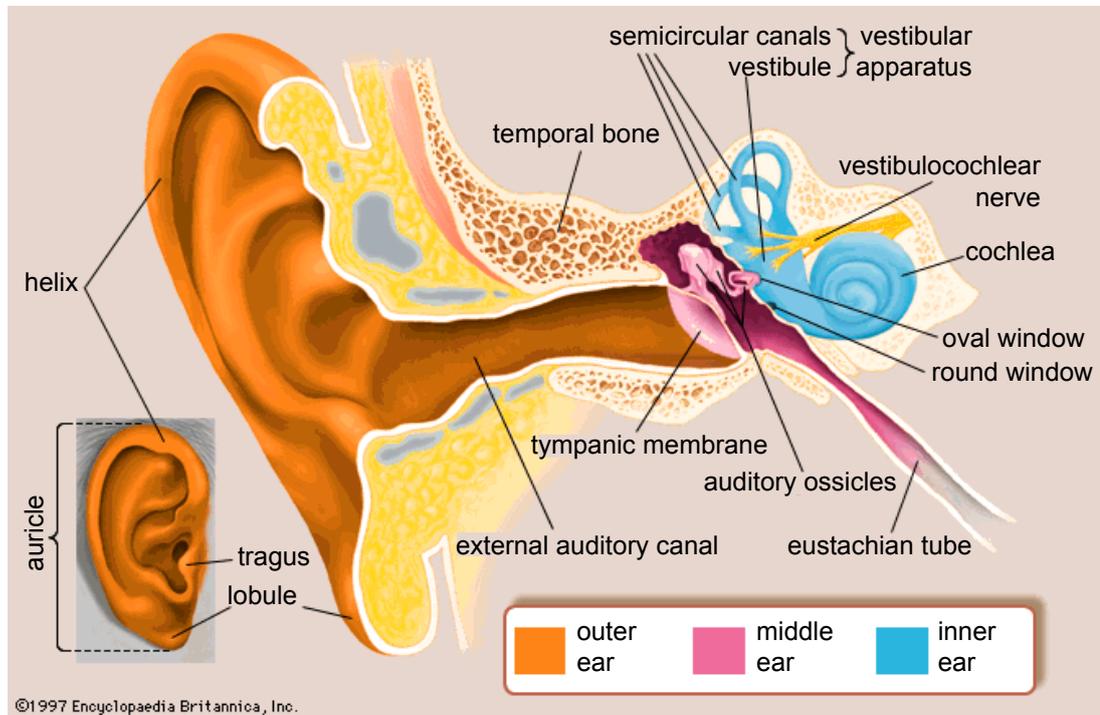


Figure 2.12: The anatomy of the human ear. From [Bri11a].

Outer ear

The outer ear is the visible part of the ear. The most noticeable, a shell-like part, is the *pinna*. The pinna is mostly used for sound localisation. It differs in shape amongst individuals, and therefore, each individual hears each sound differently. A sound, reflected off of the pinna, is further channeled down the *ear (auditory) canal*. The ear canal is around 7mm in diameter, with its resonant frequency around 3,000Hz. Hence, the highest sensitivity at this and nearby frequencies. The ear canal ends with the *tympanic membrane*, which transmits the incoming vibrations to the middle ear.

Middle ear

The middle ear is an air-filled chamber, which connects the outer and the inner ear. On one side, the tympanic membrane closes the “entrance” to the middle ear. Similarly, another tiny membrane, called the *oval window*, separates the middle ear from the liquid-filled inner ear. The three smallest bones in the human body, called *ossicles*, bridge these two membranes. The liquid in the inner ear produces more resistance to the wave movement than the air, because of its higher molecule density. Therefore, the ossicles, besides transmitting, also amplify the vibrations from the outer ear into the inner ear. The ossicles consist of three bones: *hammer* (malleus), *anvil* (incus) and *stirrup* (stapes). The hammer is attached to the tympanic membrane and its vibrations are initiated by the membrane movement. The anvil is further bound to the hammer by ligaments. The last one in the chain is the stirrup, whose footplate is lying on the oval window. In order for the middle ear to function correctly, the air pressure must be equal to the atmospheric pressure in the ear canal. The mechanism for the pressure equalisation is provided by the *Eustachian tube*, the small canal connecting the middle ear and the throat.

Another important mechanism, embedded into the middle ear is the *acoustic reflex*. Its function is to dampen the sound level entering the inner ear. The acoustic reflex includes two tiny muscles: the tensor tympani, attached to the eardrum; and the stapedius, attached to the stirrup. When the ear is exposed to a high level of noise, these two muscles contract, stiffening the eardrum and reducing the movement of the stirrup.

Inner ear

The inner ear consists of few parts and two major functions: maintaining the balance and orientation in space; and frequency and intensity analysis. The

first function is achieved through a specialised sensory system called *semicircular canals*. The three liquid-filled, “looped” canals behave as an accelerometer, used in vehicles, navigation and other engineering systems. It also helps us to maintain the fixation on an object of interest while moving our head in any direction.

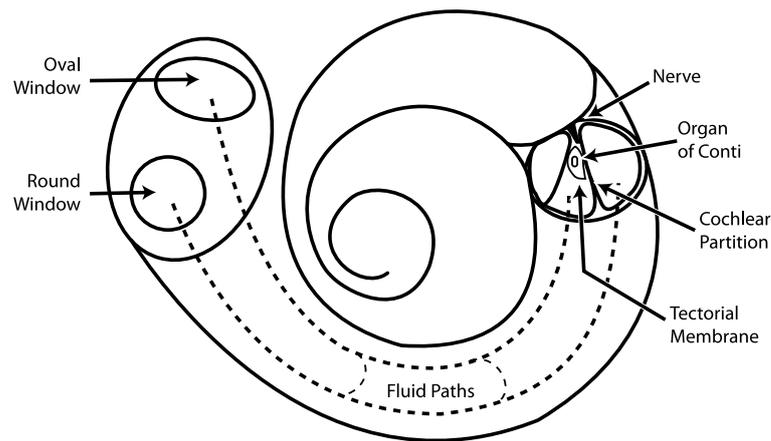


Figure 2.13: Illustration of the human cochlea.

The other part of the inner ear, responsible for hearing, is the **cochlea**, see Figure 2.13. The cochlea is spiral shaped and comprises of three chambers: *vestibular canal*, *cochlear duct* and *tympanic canal*. The first and the last are connected at the end (a place called the apex). The vibrations from the middle ear are transmitted through the oval window, located at the base of the vestibular canal. At the base of the tympanic canal there is another tiny membrane, the *round window*, that compensates the pressure caused by the inward movement of the oval window. The cochlear duct is a separate chamber, containing a different type of liquid. It is separated from the tympanic canal by a basilar membrane. On top of the basilar membrane there is a structure named **the Organ of Corti**, which contains the receptors - hair cells - and transforms the fluid vibrations into neural impulses. There are two types of hair cells: inner hair cells (IHC) and

outer hair cells (OHC). Although the OHCs are more numerous (around three times), 95% of the auditory nerve fibers are attached to the IHCs. A vibration of the oval window creates a travelling wave along the basilar membrane. Depending on the wave frequency, it reaches its peak at a certain distance displacing the basilar membrane which stimulates hair cells at that location, resulting in a tone perception. The relation between the frequency and the length of the basilar membrane is called *tonotopic organisation*. The same principle applies to the intensity of a sound. The higher the intensity, the higher the amplitude of the travelling wave, and the stronger the hair cells stimulation. The neural information from both ears is further carried by the auditory nerve to the auditory cortex.

2.2.3 Auditory perception

In our everyday life, we are constantly exposed to sounds, which are being processed by the brain whether we listen to them or not. There are two major elements of sound that we are interested in, when listening: meaning - *what* we hear; and location - *where* is it coming from. Hearing sensitivity, auditory processing and localisation will be briefly explained below. More details can be found in [Moo82, PISI00, Tsi07, RS08].

Spatial hearing and localisation

Although both sound and light are omnidirectional, our auditory system can receive information from all the direction around. This is different from our visual system. However, the HAS has mechanisms to filter certain sounds and localise them in space. The former is called *auditory masking* or the *cocktail party effect*, and it allows us to pick out and listen to a single sound in a noisy environment [Moo82]. However, sometimes fragments from these sounds could

be missing, without any perceivable effect to a listener. This is possible due to the continuity illusion phenomenon [KT02]. There are two important sound characteristics that influence the auditory localisation: spectral bandwidth and intensity [RS08]. The broader the bandwidth and the louder the sound is, the better our ability for sound localisation. The main factors that affect sound localisation are: binaural and monaural cues, reverberation and inter-sensory interaction.

Interaural intensity difference (IID) is one of the binaural cues for sound localisation. Depending on the azimuth of a sound source, each ear receives the sound at a different intensity level, except for the sounds originating directly ahead or behind us, see Figure 2.14. This cue is stronger for the higher frequencies, since the wave length of the low frequencies are longer than the diameter of the head, which, in that case, does not impede the waves.

Interaural time difference (ITD) is a similar cue to the IID, where one ear receives the sound slightly earlier than the other, see Figure 2.14. Similarly to the IID, the highest difference - around 700 μsec occurs when the sound originates from the side of the head.

Although being a powerful tool for sound localisation, binaural cues do not provide sufficient information about the sound source elevation. Monaural cues, however, can provide us with that information using *head-related transfer functions (HRTFs)*. As the sound travels it reflects off the head, body and pinna. During these reflections some of the energy is lost which leaves the sound spectrum suitable for sound localisation. In certain ambiguous positions, such as from ahead or from the behind of the head, where the IID and ITD are the same, head movement breaks the symmetry and resolves the confusion.

Another important element of sound localisation is distance perception. This ability evolved as we had to know if a prey or a predator is nearby or far away.

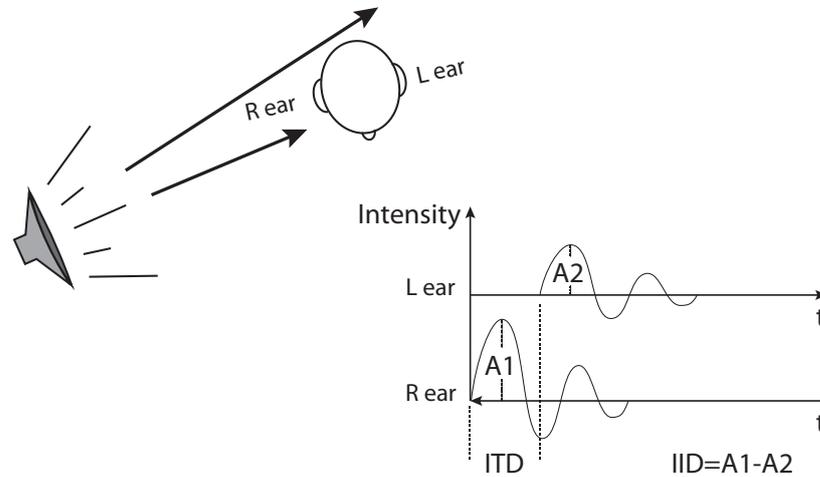


Figure 2.14: Binaural cues: Interaural Intensity Difference (IID) and Interaural Time Difference (ITD).

When listening to a sound indoors, we rely on the reverberation. However, this cue is missing in outdoor environments, and it is substituted by sound intensity and movement of the sound source. Although this can be useful in sound localisation, it behaves rather poorly for unfamiliar sounds.

Despite these localisation techniques, the spatial auditory resolution is very limited. According to Perrott and Saberi, the minimum vertical audible angle without change in elevation is 0.97° and the minimum horizontal audible angle without change in azimuth is 3.65° [PS90]. This makes hearing substantially weaker than vision in spatially related tasks. However, the temporal resolution of the HAS is rather high compared to the HVS, and according to Fujisaki et al. it is 89.3Hz [FN05].

Inter-sensory aspects of perception will be explained in Subsection 4.1.

Temporal auditory processing

Temporal auditory processing can be divided into temporal integration and temporal resolution. Temporal integration is the ability to integrate acoustic features of a particular sound over time. It has been reported that for humans this time varies between 50 and 200ms [RS08]. One example of temporal integration is a forward masking paradigm. If the observer is presented with two sequential sounds, his/her ability to detect the second stimulus will depend on the duration of the first one. This happens because of the adaptation to energy at the frequencies present in the first sound (masker). Additional parameters that influence the masking intensity are: intensity and duration of the masker, duration of the interstimulus interval, duration of the second stimulus (target), the onset interval between the masker and the target (duration of the masker plus interstimulus interval) and interstimulus interval plus duration of the target.

Opposite to temporal integration is temporal resolution, which is the ability to resolve time. Using experiments to measure the ability of gap detection in a sound, it has been shown that humans can detect a silent period lasting a few milliseconds [RS08].

2.3 Attention and Perception

There are three stages in human sensory information processing: sensation, perception and cognition. Sensation is the physical stimulation of the sensory organs. Perception is a set of processes by which we deal with the information sensed in the first stage. Cognition is the most complicated stage in which the information has been fully processed and possibly used for learning, decision making, storing into memory, etc. [MGKP08].

William James in his book *The Principles of Psychology* speaks about attention: “Every one knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and is a condition which has a real opposite in the confused, dazed, scatterbrained state which in French is called *distracted*...”. In his other book *Psychology* he defines perception as “*the consciousness of particular material things present to sense*” [Jam92]. Research in psychology has considered the perception of individual senses separately, and across different modalities. Although the understanding of the perception of individual senses is crucial, in reality, we are rarely exposed to a stimuli affecting solely one modality. Instead, few or all of the senses are stimulated simultaneously, where even if one modality “fails”, the information is received and processed unmistakably, due to the cross-modal integration, see Section 4.2. Additionally, stimulation in one sensory modality can affect the perception in other. This will be discussed in Section 4.1.

Perception can also be affected by other factors, e.g. by user’s beliefs and experience, or by value and need. This was described in 1947 by Jerome Bruner and initiated a movement later named “new look in perception” [Pyl06]. This

paper inspired hundreds of experiments, which proved that e.g. poor children perceive coins as bigger than rich and that a hungry person is more likely to see food.

Our senses are exposed to a number of different stimulations at almost every moment. However, even though they affect our sensory organs, due to attentional limitations they may never get processed so that we experience them [Jam92].

This mostly depends on our consciousness and the focus of the senses and our mind, which is called attention. It can be described as a filter to perception, which helps us to process only relevant information and ignore the rest. The attention can be: completely concentrated, where even the body injuries can remain unnoticed due to the extreme focus of interest; dispersed attention, where the mind is emptied and a person is thinking of nothing - we look and listen but none of what we “see” and “hear” is being absorbed and processed; and the attention that is between these two extremes [Jam92, Jam90]. Depending on the intent, the attention can be intentional, endogenous, top-down attention, where the observers voluntarily orients attention towards a spatial location relevant to the task or action they are undertaking; and unintentional, exogenous, bottom-up attention, in which it is involuntarily captured by a certain event [The91].

2.3.1 Attentional capture

The endogenous attention is selective, which means that we are able to focus our attention in order to process some stimuli more than other. Additionally, Pashner characterised attention as capacity limited and effortful [Pas99]. The latter means that continuous processing of an even stimulus, even if it is enjoyable, may lead to a fatigue. An example of the endogenous attention is seen in the inattentional blindness phenomenon, firstly introduced by Mack and Rock, which demonstrates the inability to detect salient objects in the centre of our gaze, when

performing a task irrelevant to the distracting object [MR98]. In the experiment, participants were asked to judge the size of the arms of a cross briefly presented on a computer screen. The majority of the participants failed to notice unexpected objects appearing on the screen along with the cross. The research was extended with more natural displays by Simons and Chabris in 1999, confirming the same hypothesis [SC99].

The exogenous attention is mostly attracted by a salient objects or their salient features, or by a sudden motion [Yar67, IK01, Sch01]. This means that if there is a red ball on the white background, our gaze will be shifted towards it, or if in the static display an objects starts moving, our attention will unintentionally move towards the moving object. According to Koch and Ullman, exogenous visual attention depends on colour, intensity, orientation and direction of movement, which form topographical, cortical maps called feature maps [KU85]. These maps combined form a saliency map, which can be used to predict the areas of an image which have the highest probability of attracting users' attention. Itti et al. developed a computer implementation of the model for static images [IKN98, IK00, IK01]. An extended version of the framework using Aleph map, which includes the temporal component, was developed by Yee et al. [YPG01].

Neither of the two attentions is constant and they change over the time. Shifting the visual attention, is a particularly important process, as it can significantly affect our visual perception. There are two types of visual attentional shifts: overt and covert shift. If the visual attention is shifted by the eye movement, where the object of interest falls in the foveal region of the eye, the overt attention is used. On the other hand, if we do not require a high level of acuity, we can shift our attention without the gaze shift. This process is known as the covert attention. For a comprehensive overview see [Pos80].

2.3.2 Attentional resources and limitations

As mentioned earlier in this section, our attentional capacity is limited. This means that an attentional pool exists, which, once it gets saturated, cannot receive any further information. However, researchers are not sure on which level this pool operates. There are two parallel, though opposed views on the matter. The first one claims that these resources are inter-modal, shared between modalities, and the second that resources are individual, intra-modal, where each modality has its own attentional pool. However, there are number of parameters affecting the evaluation of this kind, such as detection versus discrimination paradigm and forgetting in short-term memory [MW77]. Furthermore, there is an example of how cross-modal attentional links depend on type of attention, such as covert versus overt and endogenous versus exogenous attention [DS98]. The authors show that shifts of covert attention in one modality induce the attentional shift in other modalities. Similar results can be found in [SM04].

Inter-modal

Some models of attention impose that our attention operates on a global level and is not divided across multiple senses. This means that the performance of a task requiring attention for one modality will be affected by concurrent task in some other modality. For example, speaking on the mobile phone can disrupt the car driving performance, due to the attention diversion [SJ01]. Additionally, there is a difficulty in attending to different locations in the two modalities, e.g. when the audio originates from the left hemisphere and video from the right hemisphere, and vice versa [DS94]. They used recorded audio, played from either left or right side, with active (synchronous lip-movement) and passive (meaningless lip movement) on either same or opposite side of the audio. The same authors

showed that the further the positions of auditory and visual stimuli are, the easier it is to selectively attend to a particular modality [SJD00]. These findings do not explicitly prove the existence of a supramodal attentional systems, but they rather indicate that a spatial link between visual and auditory attentional systems exists.

Intra-modal

On the other hand, Alais et al. in a research dealing with attentional resources for vision and audition [AMB06], claim that there is no attentional dependencies between modalities, at least for low-level tasks, such as discrimination of pitch and contrast. In their experiment, they showed that there was no significant difference in performance between single stimulus and multi-modal dual task. Nevertheless, when two tasks within the same modality were assigned, the performance was significantly reduced, which indicates that there might be some attentional limitations within the modality when performing a dual task. Similar results can be found in [AAR72, BH98, DMW97, BA06].

A possible reason for different findings in these two groups might be the way the experiments were set up. In the first group ([DS94, SJD00]) subjects were directing attention to different spatial locations, while in e.g. [BA06], the attention was distributed across the entire sensory field, with as little change in spatial location of the stimulated regions as possible.

When observing visual and spoken letters presented simultaneously, there is no significant difference in performance when both letters along with the modalities must be reported or when either visual or auditory letter has to be reported regardless of the modality [LMBB03]. Additionally, the *modality confusion* is often experienced, where spoken letter is reported to be seen or visual letter to be heard.

2.4 Summary

This chapter provided an overview of the HSS, attention and perception. The anatomy of both HVS and HAS was briefly presented, followed by the perceptual factors in both spatial and temporal domains. Additionally, sound properties were presented. These concepts are important for the better understanding of the following chapters. In this chapter the different types of attentional captures and shifts have been described, and their implications have been discussed. Furthermore, the attentional limitations were explained, presenting the concepts of inter-modal and intra-modal attentional models.

CHAPTER 3

High-Fidelity Rendering



Figure 3.1: Physically-based rendering examples: images rendered using Path tracing method (*courtesy of Piotr Dubla*) (top left and bottom right); an image rendered using Radiance rendering package [War94] (top right); an image rendered using Mental Ray (*courtesy of Jassim Happa*) (bottom left).

Rendering is the process of digital image generation from a description of 2D or 3D virtual space. This description generally comprises geometry, materials, lighting properties and camera attributes. Rendering can be divided into two major groups: *physically-based* and *non-physically-based* rendering. The former,

considers physical material properties and calculates light emission and propagation through an environment, taking into account light contribution not only from light sources, but from the whole environment, including indirect contribution, see Figure 3.1 for an example. The latter is an approximation of the physically-based approach, generally used in applications where reproduction of the real world is not required, or when higher interactivity, and thus higher rendering speed is needed, see Figure 3.2.

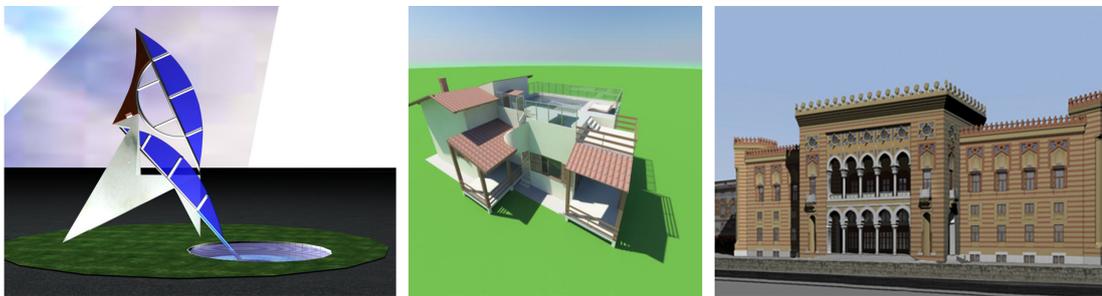


Figure 3.2: Non-physically-based rendering: design concepts (left and middle) and cultural heritage virtual reconstruction example (right).

Since our attention and sensory system are limited (see Chapter 2), there is a threshold above which any rendering quality improvements remain imperceptible. Images that are rendered at this quality level are called high-fidelity images. This is further discussed later in the thesis.

Using high-fidelity rendering, typically based on physically-based quantities, it is possible to reproduce real-world effects such as soft shadows, colour bleeding, caustics, etc. However, these elements increase the rendering time, making it rather impractical for real-time interactive applications. Therefore, some methods use pre-computation, for calculating complex lighting effect offline, for which results are then combined with the real-time rendering, thus producing a high-fidelity outcome interactively. However, this can be used only for static environments, where neither light sources nor objects change.

In this chapter, the basic concept of rendering will be covered, including radiometry, light reflectance models and light transport. Subsequently, rasterisation and ray tracing will be explained, followed by the rendering methods used for the generation of the test stimuli discussed in Chapters 5, 6 and 7.

3.1 Radiometry

Radiometry is the field of study dealing with electromagnetic radiation measurements. For better understanding of the rest of the chapter, basic quantities and principles from radiometry will be explained.

Radiant flux is the amount of energy that passes through a surface in a unit of time. It is usually signified as Φ , and its unit is $\frac{J}{s}$ (Joules per second) or W (Watts).

Irradiance (E) is the radiant flux passing through a unit area. It is measured in $\frac{W}{m^2}$. Irradiance at a point on a surface is calculated as [DBBS06]:

$$E = \frac{d\Phi}{dA}. \quad (3.1)$$

The same equation is used for calculating radiosity, which is a common quantity in radiosity rendering algorithms.

Intensity is the flux density per solid angle, where solid angle describes the area on the unit sphere covered by the projection of the observed object. Solid angle is measured in steradians.

Radiance (L) is the flux density per unit area per unit solid angle [DBBS06]:

$$L = \frac{d^2\Phi}{d\omega dA \cos\theta} \quad (3.2)$$

From this, it is possible to derive the relations between those quantities. Ir-

radiance at point p over a set of direction Ω is given by [DBBS06]:

$$E(p) = \int_{\Omega} L(p, \omega) \cos\theta d\omega \quad (3.3)$$

where $L(p, \omega)$ is the arriving radiance at point p from the direction ω .

3.2 Light reflectance models

Accurate material properties play a major role in the process of digital image synthesis. Every surface has specific characteristics, defining the way it reflects, refracts and absorbs light. There are two major reflectance components that contribute the overall reflectance model: diffuse and specular. However, there are very few materials that can be represented as purely diffuse or specular. The glossy component is a combination of the two, see Figure 3.3.

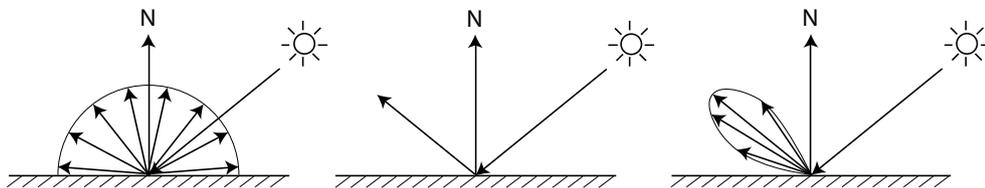


Figure 3.3: Light reflectance: diffuse (left), specular (middle) and glossy (right).

An accurate mathematical representation of the reflectance model is given by the bidirectional scattering distribution function (BSDF). This function is composed of two components: bidirectional reflectance distribution function (BRDF) and bidirectional transmission distribution function (BTDF). However, the majority of materials can be represented using only BRDF, which is an approximation of the BSDF (Bidirectional surface scattering reflectance function) that includes phenomena such as subsurface scattering. This function defines how much radiance leaves a surface from point p in the direction ω_0 , as a result of

incident light falling at the same point from the direction ω_i [PH10], see Figure 3.4.

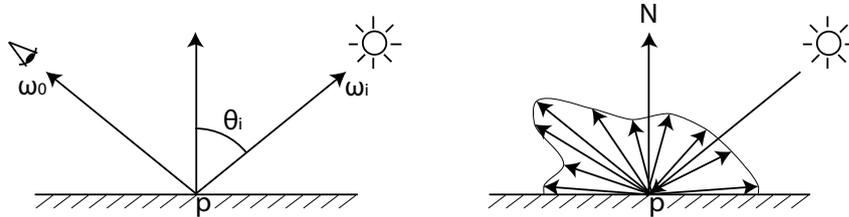


Figure 3.4: The bidirectional reflectance distribution function (BRDF). After [PH10].

The BRDF function f_r is defined by:

$$f_r(p, \omega_0, \omega_i) = \frac{dL(p, \omega_0)}{dE(p, \omega_i)} = \frac{dL(p, \omega_0)}{L(p, \omega_i) \cos\theta_i d\omega_i} \quad (3.4)$$

There are two common properties for BRDFs: reciprocity and energy conservation. The former assumes that for each pair of directions ω_0 and ω_i , $f_r(p, \omega_i, \omega_0) = f_r(p, \omega_0, \omega_i)$, whereas the latter says that the total amount of the reflected light from point p is less than or equal to the energy of the light received at the same point.

3.3 Light transport

The mathematical model for physically-based rendering calculation was first presented by Kajiya [Kaj86]. Assuming no participating media present in the scene, the equation is given by:

$$L(p, \omega_0) = L_e(p, \omega_0) + \int_{S^2} f(p, \omega_0, \omega_i) L_d(p, \omega_i) |\cos\theta_i| d\omega_i \quad (3.5)$$

where $L(p, \omega_0)$ is the reflected radiance on a surface in the point p along the ω_0 direction; $L_e(p, \omega_0)$ is the radiance emitted from the surface; and

$\int_{S^2} f(p, \omega_0, \omega_i) L_d(p, \omega_i) |\cos\theta_i| d\omega_i$ is the scattered radiance from point p as a result of the incident illumination at the point from the other surfaces across the sphere (S^2).

Because of the complexity of the calculation, the rendering equation is not suitable for interactive dynamic environments, composed of large number of surfaces and light sources. Therefore, a convenient integration technique, which can approximate the light transport model, needs to be applied. In the following sections, two popular rendering techniques will be briefly described.

3.4 Rasterisation

For increasing efficiency through parallelism, the rendering process is divided into stages and forms the graphics rendering pipeline. The pipeline can consist of various stages, depending on the implementation. A coarse division into three stages would comprise application, geometry and rasterisation stages [AMHH08]. *Rasterisation* is the rendering method that uses this pipeline, and its name comes from the last stage of the pipeline. More detailed and the conventional pipeline representation is given in Figure 3.5.

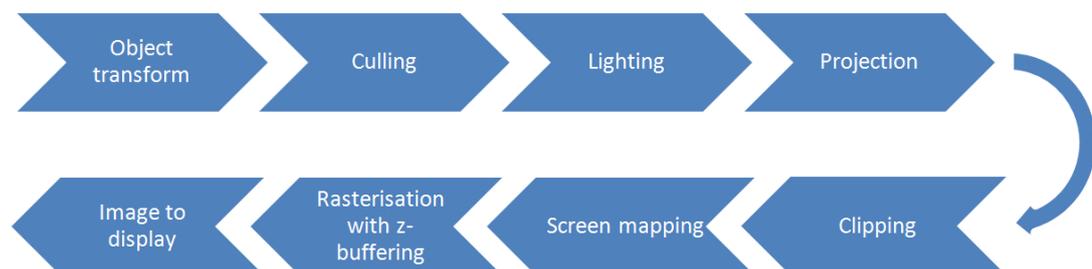


Figure 3.5: Traditional rasterisation pipeline.

The first stage in the pipeline is the application stage. The processes usually implemented in this stage include, for example, collision detection, object and

camera movements or animations of object models. The external inputs (keyboard, mouse, etc.) are also handled in this stage. As the output, the rendering primitives, such as points, lines, triangles or similar are passed to the geometry stage. Since this stage is typically implemented on a CPU, it can be sped up by adding processor cores.

The first step in the geometry stage is to put all objects in the virtual scene into one unique coordinate system - world space. Since the virtual camera is represented as an object in the scene, the origin is then moved to the centre of the virtual camera - a process called view or camera transform. Subsequently, faces pointing away from the camera are removed in the stage called back face culling. In the next stage, lighting is applied to each polygon vertex. Typically, at this stage only the vertex shading is performed. The quality and time depend on the lighting model. There are two types of projections: orthogonal and perspective. The former projection is parallel and the view volume is a rectangular box. The latter projection is more complex, producing a depth cue by scaling the objects according to their distance from the camera - the further the object the smaller it appears. The view volume of the perspective projection is a truncated pyramid. After this stage, all objects and parts of the objects that are outside of the view volume are discarded. This process is known as clipping. The primitives that fall within the view volume are then mapped to the screen (window) coordinates.

Given the transformed and projected vertices with their associated shading data (all from the geometry stage), the goal of the rasteriser stage is to compute and set colours for the pixels covered by the object. This process is called rasterisation or scan conversion, which is thus the conversion from two-dimensional vertices in screen space - each with a z-value (depth-value), and various shading information associated with each vertex - into pixels on the screen.

Finally, in the rasterisation stage of the pipeline, the data from the geometry

stage is converted from two-dimensional vertices in screen space into pixels on the screen. This stage can be further divided into substages, but generally is used for the per-pixel operations, such as triangle traversal, hidden surface removal using z-buffer, per-pixel shading, texture mapping and merging.

The presented model is a substantially simplified version of the physically-based rendering. The fidelity of the resulting images can be increased by implementing additional features, such as global illumination, participating media, caustics, etc. For an overview see [AMHH08]. Rasterisation is used as Autodesk Maya default software renderer, with adjustable rendering features and parameters accessible through the graphics user interface (GUI). This renderer was used for the Kiti scene animation in Chapter 6, see Section 6.2.4.

3.5 Ray tracing

A ray, used in ray tracing methods [Whi80], is an approximation of a photon stream. There are two general ways of tracing rays: forward and backward. Forward ray tracing uses the natural way of light propagation - from the light source through the environment. However, this is highly inefficient as very few rays reach the camera in a finite number of bounces. Therefore, the opposite approach has been proposed, where rays are traced from the camera through the virtual scene, see Figure 3.6. This way there is no wasted computation spent on traversing rays that do not contribute the final image appearance. Some methods, such as photon mapping use both forward and backward ray tracing, as discussed in Section 3.8.

Ray tracing was initially used for hidden surface removal in a form of ray casting [App68]. This method detects the first ray-object intersection by shooting a ray from virtual camera through each pixel on the image plane. In classical ray

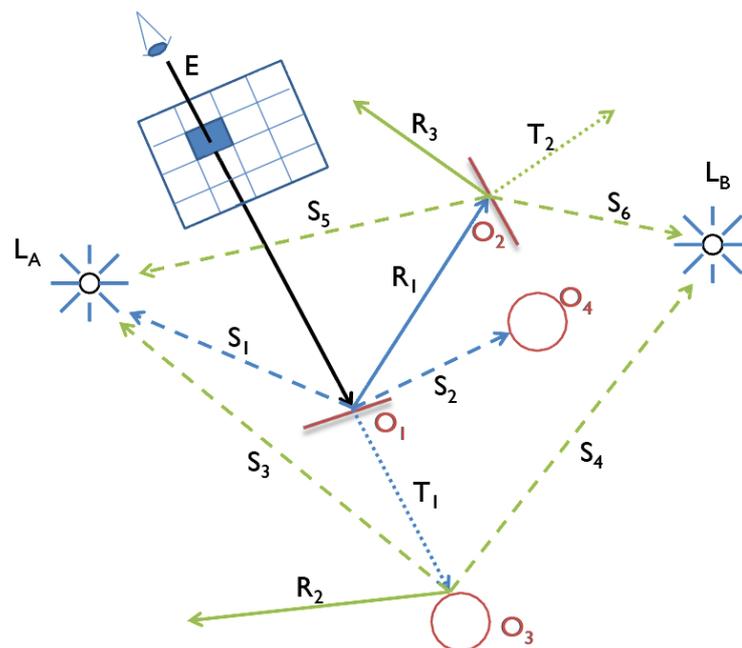


Figure 3.6: Backward ray tracing. A ray is shot through a pixel. At the intersection point, the ray is spawned: shadow rays S_1 and S_2 are shot towards light sources; reflectance ray R_1 is shot off the surface at the calculated angle; transmission ray T_1 is shot through the surface. The process is repeated recursively for R_1 and T_1 .

tracing [Whi80], the first step is to shoot a ray and check for the intersections with the scene objects. The closest point of intersection to the virtual camera is stored. Then, the shadow rays are shot towards each light source to check for the visibility. If there is an object between the intersection point and light source, the point is in shadow. Otherwise, the object is shaded. This process provides the local illumination of the scene. For global illumination, which is the model that combines both direct and indirect illumination, Monte Carlo techniques have been developed. Those include distributed ray tracing [CPC84] and path tracing [Kaj86].

3.6 Path tracing

Path tracing, as defined by Kajiyama's rendering equation (3.5), is an algorithm for high-fidelity rendering [Kaj86]. In this algorithm, only one random ray is traced from the camera to its termination. Therefore, there is no *ray explosion* as in distributed ray tracing, which reduces the memory consumption. Shooting only one ray per pixel produces jagged edges, a phenomenon known as aliasing. Additionally, the resultant image looks grainy. In order to remove those artefacts, multiple rays are traced through each pixel. Producing noise-free images using path tracing is computationally expensive. However, the fidelity of the outcome is extremely high and it is usually used as a referent model for comparison with other rendering techniques, see Figure 3.7.

Path tracing rendering technique was used for the generation of the animations used in Chapter 7 and for the Rabbit scene in Chapter 6.



Figure 3.7: An example of path traced image (*courtesy of Piotr Dubla*).

3.7 Irradiance caching

Distributed ray tracing is an extension of the classical ray tracing algorithm, used for achieving softer appearance of the resultant image. This method shoots multiple rays distributed over an interval, e.g. for achieving soft shadows, multiple shadow rays are shot over the light source area. This method calculates the indirect diffuse component by shooting multiple rays over the hemisphere at the intersection point. Ward et al. noticed that this component is a continuous function over the space, and that the high frequency variations in the environment do not affect its appearance as in case of specular or highly glossy components [WRC88]. Therefore, they introduced the acceleration data structure known as the irradiance caching, which stores the world space irradiance values for a given scene. Each time an irradiance sample is required, the cache is first consulted. If there is another sample calculated within the predefined search radius, the result is interpolated from it. This speeds up the rendering process by reducing the number of rays traversed. This data structure is usually represented by an oc-

tree. The stimuli used in Chapter 5 were rendered using this rendering technique, see Section 5.2.1. Furthermore, the same technique was used for the rendering of the Kalabsha scene in Chapter 6.

3.8 Photon mapping

Another popular rendering technique, used for image generation of some of the test stimuli in Chapter 6 is Photon mapping. Photon mapping is a two-pass rendering algorithm that solves the rendering equation [Jen01]. In the first pass, rays from the light source(s) are traced and stored into photon map, represented by a three-dimensional data structure. In the second pass, ray tracing is used for calculating direct lighting, specular and glossy components. However, for calculating the lighting on diffuse surfaces, the photon map is first consulted for caustics or indirect diffuse illumination. For more realistic results, final gathering is performed for indirect diffuse calculation. Photon mapping is utilised for producing effects such as caustics, diffuse interreflections and subsurface scattering.

Photon mapping is used in the Mental Ray rendering software [Men] found in Autodesk Maya. This renderer was used for the generation of the Kiti-MentalRay, Ball, People and Cars scenes in Chapter 6.

3.9 Summary

This chapter has described the concept of high-fidelity rendering, providing an insight into radiometry and light transport used in physically-based rendering. There is a number of rendering systems available, ranging from commercial packages, such as Mental Ray [Men], Maxwell [Nex], V-Ray [Cha], Lux renderer [Fre], etc. to implementations within game engines (CryEngine 3 [Cry], Unreal Engine

[Epi], Frostbite Engine [EA], etc.). However, in this chapter, only the rendering techniques, which have been used for the generation of the stimuli presented in Chapters 5, 6 and 7, are discussed.

CHAPTER 4

Cross-Modal Interaction

Our senses are almost constantly being stimulated. Furthermore, they are mostly being stimulated simultaneously. This means that when we talk to someone, we are not only listening but also looking the lips or the gesticulation of the speaker. At the same time we smell the environment, feel the temperature, have a sense of direction and balance, etc. All those senses can be examined solely, or the interaction and integration between them can be studied. In this chapter the relationship between vision and audition will be discussed, focusing on both results from psychology and research in computer graphics.

4.1 Auditory-visual cross-modal interaction in psychology

In Chapter 2 we saw mechanisms and limitations of vision and audition. The perception and attention were introduced and basic features explained. However, so far, only individual senses were considered, without an insight on the interaction between different modalities. In this section the cross-modal interaction between audition and vision is reviewed, and the most important and relevant findings are explained and discussed.

Our vision and audition do not have the same spatial and temporal sensitivity.

The temporal resolution of the HVS is found to be lower (26Hz) than for the HAS (89.3Hz) [FN05]. The synchrony detection between auditory and visual stimuli was also investigated using psychophysical experiments. The results revealed that it is not just a temporal lag between stimuli that influences the discrimination task, but also the temporal frequency. For temporal frequencies higher than 4Hz the synchrony-asynchrony discrimination becomes impossible even when the lag between stimuli is large enough to discriminate it with single pulses. Above this frequency the *auditory driving effect* occurs [GM59, Shi64]. This effect is described in Section 4.1.1.

These differences in spatial and temporal sensitivities of vision and audition are the basis of the *modality appropriateness hypothesis* [HT66, WW80]. This hypothesis advocates that the modality that is more appropriate for a certain task will dominate that particular task. In other words, our vision is more accurate in spatial judgements, while audition dominates in temporal domain. However, if e.g. auditory stimuli are presented at a constant frequency, the change in the visual flicker can change the perception of the auditory flutter [WKN03], which is in collision with the modality appropriateness. The modality appropriateness hypothesis might be considered as a foundation of many other perceptual studies in psychology. Furthermore, it has been extensively exploited within computer graphics [MC04, Mas06], including the work presented in Chapters 6, 7.

Research in psychology has shown that strong cross-modal interactions exist [GGB05, Rec03, BA06] and that these cross-modal effects must be taken into consideration when the perception of distinct sensory modalities is investigated [SS01, SS04].

The auditory-visual cross-modal interaction can be divided in two ways: according to target modality into auditory influence on vision and visual influence on audition; and according to the domain, into spatial and temporal domains.

4.1.1 Auditory influence on visual perception

In this section, auditory influence on visual perception is discussed. This is particularly important as it is directly related to the design and procedures of the user studies described in the chapters to follow.

Auditory driving effect

The *auditory driving* was first found in late fifties and investigated ever since [GM59, Shi64]. Although it is possible to discriminate between different flicker and flutter rates individually, cross-sensory rate matching proved to be more difficult. Interestingly, the same observers gave better results when comparing the stimuli successively than when doing the simultaneous comparison [GM59]. Furthermore, if the white light is set to flicker at a certain frequency synchronously with the audio flutter, then increasing or decreasing the flutter rate it is possible to perceive significant perceptual rise or drop of the flicker rate respectively [Shi64]. The same author reported that for one observer a flicker rate, initially set synchronously with flutter to 10 cycles per second, was perceptually decreased to 7 cycles per second and increased to 22 cycles per second by changing the auditory flutter rate. This might be utilised in computer graphics by increasing the beat rate, which should lead to a perceived improvement of the frame rate. Furthermore, it might be possible to maintain the perceived frame rate when the real frame rate drops, due to an increase of computational load. Mastoropoulou et al. studied the effect of music on the perception of frame rate of animated sequences, which is partially based on this phenomenon [MC04]. However, the study by Shipley showed that the flicker does not drive the perception of the flutter rate. The perception of the visual temporal structure is significantly affected by the auditory stimuli, mostly during the encoding phase [GGB05]. This

is mostly because the human perceptual system encodes the visual temporal structure using an essentially auditory code.

Illusory flash effect

Another interesting effect of audition on visual perception is the *illusory flash effect*. This occurs when a single visual flash is accompanied by multiple auditory beeps, resulting in perception of multiple flashes [SKS00, SKS02]. Comparison of the results using this scenario with the ones acquired using multiple flashes and no sound, shows that the illusory flashes are equal with the physical flashes and are not induced due to difficulty of the task, cognitive bias or ambiguity of the visual stimuli. The illusion persisted even when subjects were aware of the fact that the disc was physically flashing only once. This confirms the robustness of the phenomenon to different variations in experimental parameters. When multiple flashes were presented accompanied with a single beep, the perceived number of flashes was equal to the actual number, which shows that responses were based on the visual perception. According to the authors, the reason for this asymmetry lies in the assumption that the modality with the discontinuous stimulus can alter the perception of the modality with the continuous stimulus, much more easily than vice versa.

Temporal ventriloquism

Led by the fact that visual dominance over audition in spatially oriented tasks, e.g. the ventriloquism effect [HT66], is due to the higher acuity of vision, Recanzone conducted a series of experiments in order to find the analogy in the temporal tasks, where the superior temporal acuity of audition should dominate over the vision in multisensory stimulations [Rec03]. He showed that humans have better ability to judge temporal rate using auditory rather than visual cues.

Additionally, he showed that audio as a distracter has a clear influence on visual temporal rate perception. Moreover, this effect persisted after a period of 20 minutes, confirming the existence of the auditory-driven aftereffect, which means that a shift in the visual temporal rate existed. Other parameters were tested in this study: auditory origin location, intensity and spectral bandwidth. None of the conditions influenced the effect of auditory stimulus on perception of temporal rate in the visual domain.

As visual stimuli can change the apparent spatial location of the sound, auditory stimuli can change the apparent temporal location of the visuals [MZSFK03]. More specifically, if two flashes, showed sequentially one after another, are accompanied with audio stimuli (5 ms clicks of 67 dB) presented one before the first and one after the second flash, the video stimuli are perceived as more separated in time from each other, and thus the temporal order judgement (TOJ) task is easier to conduct. In the same manner, two clicks presented between the flashes decrease the apparent time distance between the lights and therefore the TOJ task performance is reduced. The experiments also revealed that only the second click, trailing the light affects the performance and that one click presented between the lights has no effect on the task performance. This phenomenon is known as the *temporal ventriloquism* and has been shown to work only within the temporal window known to support the multisensory integration. The temporal ventriloquism has been further investigated by Bertelson and Aschersleben [BA03, AB03]. A recent study by Burr et al. confirms the existence of the temporal ventriloquism, in case of auditory visual conflict [BBM09].

Temporal ventriloquism has been used for improving the illusion of apparent motion. By presenting two auditory clicks between two successive flashes a user's perception of the interstimuli onset interval (ISOI) is changed and thus the apparent motion illusion enhanced [MZSFK03]. This effect was significant for ISOIs

between 50 and 150 ms. For longer ISOI, Getzmann showed that a single click makes significant facilitation of apparent motion [Get07]. Furthermore, a spatially irrelevant auditory peep can enhance visual search task [VdBOBT08]. This is particularly the case when the visual change and auditory onset are presented simultaneously.

One of the main differences between the experimental setups used in psychology and computer graphics is the temporal frequency of the stimuli. In psychology, the temporal frequency of visuals is usually lower and the frequency of the auditory stimuli is higher than in computer graphics. However, it might be possible to maintain temporal visual quality while decreasing the video frame rate in the presence of auditory stimuli presented at certain beat rate, combining the aforementioned phenomena together. This is further discussed and explored in Chapters 6 and 7.

For an overview of auditory influence on visual perception refer to [SS04]. Another review of auditory spatial and temporal processing is given in [RS08].

4.1.2 Visual influence on auditory perception

Although, according to modality appropriateness hypothesis, audition dominates vision in temporal domain, the perception of ambiguous auditory temporal cues might be affected by changes in visual flicker rate [WKN03]. This might be due to the fact that when a focal modality is temporarily ambiguous, the other modality will overtake the perception, the phenomenon known as the *optimal integration hypothesis* [WKN03, VWH02]. However this influence is significantly weaker than the visual effects on audition in spatial domain.

Ventriloquism effect

One of the most significant examples of the visual influence on the auditory perception is the *ventriloquism effect* [HT66, RS08], where the visual and auditory stimuli are originating from separate locations, but the audio is perceived to be coming from the location of the visual stimulus. One of the most common examples of the ventriloquism effect is watching TV, where the sound is perceived to be originating from actors' mouths, even though it can be emitted from speakers located away from the TV. This illusion is due to higher accuracy of the visual spatial frequency, as described by the modality appropriateness hypothesis [WW80]. The ventriloquism effect can be influenced by both cognitive and non-cognitive factors. *Unity assumption* is one of the cognitive factors. It is caused by the observers assumption that the visual and auditory stimuli represent the same object or event [WW80, Wel99]. There are few non-cognitive factors that can influence the ventriloquism effect: *timing*, as the stimuli have to be synchronous; *compellingness*, which is how well the sound matches what the observer expects by watching the visuals; and the third non-cognitive factor - *spatial influence*, due to which the illusion breaks if the visual and auditory stimuli are too far apart from each other [RS08]. However, even if both the spatial and temporal disparities between visual and auditory stimuli are introduced, where visual information is task-irrelevant, the vision will dominate in spatial related judgements. Bargary et al. explained this as coding of spatial information into vision [BCN07].

An interesting aspect of the ventriloquism effect is its aftereffect. If an observer is presented with the visual and auditory stimuli at different positions for tens of minutes, the auditory stimulus location will be perceived as if it had been shifted towards the visual stimulus [RS08].

Although speech is generally considered as a purely auditory process, the visual influence on auditory perception cannot be neglected. McGurk and MacDonald reported that pronunciation of *ba* is perceived as *da* when accompanied by the lip movement of *ga* [MM76]. This phenomenon is known as the McGurk effect.

4.2 Auditory-visual cross-modal integration

So far, the auditory-visual cross-modal interaction has been examined, showing the spatial dominance of vision over audition, and temporal dominance of audition over vision. Nevertheless, in some situations, e.g. when a stimulus of the dominant sense is ambiguous or corrupted, one modality complements another modality enhancing the perceptual experience [WKN03]. In order to get a full picture of cross-modal processes, it is crucial to investigate the cross-modal integration. One common example of such integration is apparent during a conversation in a noisy environment. In such situation, it might be hard to hear and understand the person you are listening to. However, lip reading, which is a visual process, enhances the hearing ability, enhancing the overall performance. A study by Stein et al. demonstrated that a simultaneous auditory stimulus can increase perceived visual intensity [SLWP96]. The authors showed that the effect is present regardless of the auditory cue location. However, it persisted only at the location of visual fixation. Furthermore, Van der Burg et al. showed that in a visual search task, a single synchronised auditory *pip*, regardless of its position, significantly decreases the search time [VdBOBT08]. Another study demonstrated that a single auditory click can change the meaning of the visual information [SSL97]. When two identical discs, moving towards each other, coinciding and moving apart, are presented on a display with no sound, they are

perceived as they streamed through each other. However, when a brief click was introduced at the time of the collision, the discs appeared as if they bounced off each other. By testing the phenomenon further, changing the parameters of the experiment, the authors showed that the effect is not the result of heightened attention at the moment of collision, but the acoustic event at the time. Although the origin of the effect is unknown, it may be due to users expectation and past experience or as result of the multi-sensory cells stimulation and feedback onto primary visual cortex.

In order to explain this, an approach has been suggested. The authors proposed a framework in which cross-modal information can be optimally combined as a sum of all individual stimulus estimates weighted appropriately [BA06]. The optimal estimate can be calculated following the equation 4.1, where w_A and w_S are weights by which the individual stimuli are scaled, and \hat{S}_A and \hat{S}_V are independent estimates for audition and vision respectively.

$$\hat{S} = w_A \hat{S}_A + w_V \hat{S}_V \quad (4.1)$$

The weights are inversely proportional to the auditory and visual variances (σ^2) of the underlying noise distribution, see equation 4.2.

$$w_A = 1/\sigma_A^2, w_V = 1/\sigma_V^2 \quad (4.2)$$

This has been tested using different visual stimuli with different level of blurriness [AB04]. An example where audition captures the sight occurs when visual stimuli are corrupted by blurring the visual target over a large region. The blurring, however, has to be significant i.e. over about 60° , which makes most scenes unrecognisable. Nevertheless, auditory localisation was performed only by inter-aural timing difference without time varying, which is around one-sixth of the

total cues used in regular hearing, see Section 2.2.

Similarly, there is a mathematical description of the perception of an environment [CD09]:

$$P(\tau, \rho)(t) = \omega_V V(t) + \omega_A A(t) + \omega_S S(t) + \omega_T T(t) + \omega_F F(t) + \omega_\Delta \Delta(t) \quad (4.3)$$

The equation 4.3 is given as a function of time, task and preconditioning, where V, A, S, T and F are functions of vision, audition, scent, touch and feel over time (t) respectively. Δ represents a distraction factor, which indicates the user's level of concentration. Although the equation considers the main factors that could contribute our perceptual experience, it does not necessarily have to be linear. Furthermore, the preconditioning factor is highly complex and might contain multiple contributors. Therefore, this factor should be further investigated and elaborated.

Kording et al. studied the auditory-visual cause-cue relationship, investigating whether different cues are perceived as if they correspond to the same cause or not [KBM*07]. Their approach showed that the causal inference is not only a cognitive process, but it happens effortlessly in perception as well.

4.3 Auditory-visual cross-modal interaction in computer graphics

The need for perceptually-based rendering and cross-modal research in computer graphics came from the inability to achieve high-fidelity rendering in real time. This inability is a result of a high computational complexity and processing requirements of the physically-based rendering process. On the other hand, although human sensory system is extremely complex and precise, it is not perfect, see Chapter 2. This means that there is a threshold beyond which we cannot perceive any quality improvement. Furthermore, in the previous sections it has been shown that stimulation of one modality can change the perception in the other. In this section, the work in computer graphics based on this assumption will be presented.

4.3.1 Auditory rendering

Usually, in virtual environments, it is not enough to deliver only high-fidelity graphics. For a more complete experience and higher degree of immersion, the other senses should be stimulated. Most often, sound is presented along with the video. This has been shown to increase a sense of “presence” in the virtual environment [SUS94, WS98]. However, some auditory stimuli need to be rendered in real-time, which requires significant processing power, especially if multiple sound sources are present in a complex virtual environment. In order to enhance this process, while maintaining equal perceptual quality, different techniques have been explored. For example, auditory events that will attract our attention could be predicted using auditory saliency maps [KPLL05, MBT*07]. These maps are based on three features: intensity, frequency contrast and temporal contrast. Furthermore, spatial rendering of a complex auditory environment with hundreds

of dynamic auditory sources can be significantly simplified using interactive sound masking and spatial LOD, without any perceivable difference [TGD04]. For a complete overview on perceptually-based auralisation see [Tsi07].

Another approach for enhancing the auditory rendering process, that has been briefly investigated, is the cross-modal interaction. The majority of the work on this topic has been done within the CROSSMOD project [CRO]. One of the first studies, conducted by Moeck et al. proposes creation of more sound source clusters within a view frustum [MBT*07]. Since both audio and video stimuli influence the material perception during impact [BSVDD10], the rendering process can be speeded up by using both pre-recorded and impact sounds [GBW*09].

Since the main scope of this thesis is on auditory-visual cross-modal interaction in visual rendering, further discussion on acoustic rendering will be omitted. A complete survey on this topic can be found in [HHT*11].

4.3.2 Visual rendering

Similar to cross-modal auditory rendering, visual rendering performance can be enhanced utilising the limitations of the HVS and human attention. Additionally, the influence of audio on visual perception can be considered, based on the assumptions and findings given in Section 4.1.1.

Perceptually-based visual rendering

Perceptually-based rendering in computer graphics has focused on taking advantage of both exogenous and endogenous visual attention. The former is using saliency maps [YPG01], originally introduced by Itti and Koch [IKN98], while the latter is performed via task maps [CCW03].

Saliency maps were first introduced by Koch and Ulfman [KU85]. A mathematical model, based on feature maps was later developed [IKN98]. Those

feature maps are based on colour, intensity and orientations, and then combined into single topographical map called *the saliency map*. The model was first used by Yee et al., who have adapted this concept for dynamic content, by developing a spatiotemporal error tolerance map, named Aleph map [YPG01]. The map is generated for each frame of the animation, increasing the animation rendering speed in return. It uses the saliency maps with motion features, and spatiotemporal frequency in order to calculate the tolerable error threshold for the observed region. Saliency maps are later used by Chalmers et al. [CDS06] and Longhurst et al. [LDC06].

As opposed to saliency maps, *task maps* use endogenous visual attention model [CCW03]. Using this method, task related objects in the virtual scene are used for the task map creation. The map is used in rendering process so that only task related parts of the scene are rendered in high quality and the remainder in low quality, without perceptual degradation in visual quality. This has been shown to be effective since users' attentional and visual focus is on task related objects, ignoring the rest of the scene. This rendering concept is known as the selective rendering [Deb06].

As a combination of those two approaches, a new map has been developed, using both exogenous and endogenous attention. This map merges saliency and task maps into a new map called *the importance map* [SDL*05]. This technique further enhances the rendering process, by reducing image quality, while maintaining the same perceptual experience.

Rendering process can be enhanced using different perceptually-based metrics, that represent the mathematical model of the spatial and temporal mechanisms of the visual cortex. Some examples of such metrics are: Animation Quality Metric (AQM) [MTAS01, Mys02], perceptual error metric [Dal93], perceptually-based physical error metric [RPG99], illumination components metric [SFWG04]

or visual equivalence predictor (VEP) metric [RFWB07]. These and other such metrics are used for detecting and controlling the maximum tolerable level of error in the rendered images while keeping the perceivable quality above the quality threshold.

For a complete overview on perceptually-based rendering see [HL97, OHM*04, BCFW08].

Cross-modal interactions in visual rendering

Another approach to enhancing the visual rendering performance is to take advantage of auditory-visual cross-modal interaction by introducing auditory stimuli along with the visual content. An early study on auditory-visual cross-modal interaction demonstrated that the quality of the realism in virtual environments depends on both auditory and visual components [Sto98]. The author showed that high-quality audio further increases the perceptual quality of the high-quality video. Furthermore, high-quality video further decreases perceived quality of a low quality audio.

Auditory-visual cross-modal interaction in visual rendering is mostly oriented towards the auditory influence on visual perception. This influence can be divided into two domains: *spatial* and *temporal*. The former investigates how audition can be utilised in order to enhance video rendering by decreasing the spatial quality of the generated imagery, without any perceivable degradation in overall user experience. The latter considers the effect of audio on temporal visual perception.

Using selective rendering and human sensory and attentional limitations, such as angular sensitivity and inattention blindness, it is possible to render only sound emitting objects (SEO) visible in a virtual scene in high quality, while computing lower quality for the rest of the scene, without any perceivable degra-

dation in visual quality. Using abrupt sounds, the animation rendering time can be significantly decreased, by attracting users' attention towards the sound emitting object [MDCCT05a, Mas06]. Additionally, this approach might be used for interactive scenarios using the Aleph map [YPG01].

The studies mentioned above focused on the spatial domain, trying to decrease the image quality while maintaining the perceptual equivalence. However, if the temporal resolution i.e. frame rate, could be decreased, without any degradation in animation smoothness, the rendering process would be further enhanced. The starting point for the research in this domain is the modality appropriateness hypothesis, which claims that audition is the dominant modality in temporal judgements. Hence, researchers tried to find a perceptual model which will allow for lower frame rates, while playing appropriate sound, maintaining the same perceptual visual quality.

Mastoropoulou et al. investigated how music can affect temporal visual perception [MC04, Mas06], based on modality appropriateness hypothesis and the auditory driving effect. For auditory stimuli two music types were used: slow tempo / relaxing and fast tempo / exciting music, both compared with silent animations. The results showed no significant effect for either slow or fast tempo music on perceived frame rate of the observed animations. According to the authors, this may be due to a couple of factors: the frame rate difference between compared animations (4fps) might have been too small; animation clips lasted for 40 seconds, which is far beyond the human working memory (approximately 20 seconds according to Peterson et al. [PP59]). Another factor that could have affected the experiment is the emotional effect of music, which is complex and hard to measure or control. A certain type of music can induce different emotional effects in different people, which could influence our perception, and thus the experimental results.

However, abrupt sounds can attract a part of a viewer's attention away from the visuals and thus allow the frame rate of the presented animated content to be decreased without the user being aware of this reduction [MDCT05b]. This has been proven for both related (sound source visible in the scene) and unrelated sound effects, such as phone ringing or a thunder clap. Furthermore, users familiar with computer graphics were found to have more accurate responses to the frame rate variations. There was no effect of camera movement type found to be significant in the experiments.

Since this is a new area of research in computer graphics, there has not been much work carried out in this field so far. However, all these findings provide a good starting point for the work in this thesis. Therefore, the original work in the thesis represents a continuation of the research by Mastoropoulou [Mas06].

4.4 Summary

This chapter has provided an overview of the auditory-visual cross-modal interaction. First, the findings from psychology were discussed, explaining the phenomena most relevant to the work conducted in this thesis. Although these concepts have been proven to work with simple stimuli, such as video flashes and auditory beeps, more work is required in order to extrapolate these results, using stimuli and scenarios that can be utilised in computer graphics for enhancing rendering process. Secondly, the work that has been done in computer graphics field, until the time of writing this thesis, has been discussed in Section 4.3. Although this work is still in its infancy, it shows the potential that needs to be further investigated.

The next chapters will provide the original work of the thesis author, focusing on both spatial and temporal domains of the auditory-visual cross-modal

interaction in computer graphics.

CHAPTER 5

The Influence of Cross-Modal Interaction on Perceived Rendering Quality Thresholds

5.1 Introduction

The main limitations of the HSS and attentional resources have been introduced in previous chapters. Because of these limitations, it is common that high-fidelity rendered scenes have a greater fidelity than it is possible to perceive. Additionally, when multiple senses are stimulated simultaneously, cross-modal interaction between them exists. In previous work it has been demonstrated that computational performance of rendering algorithms can be improved by decreasing spatial or temporal visual quality without any perceivable quality degradation, when auditory stimuli is introduced (see Section 4.3.2). This study relies on those findings and some existing research in computer graphics. Sundstedt et al. [SDC05] investigated perceived aliasing thresholds by altering the number of rays shot per pixel (rpp) on static scenes and animations. A similar study was conducted by Aranha et al. [ADCH06] using small screen devices. Neither of these studies used

audio as a complementary stimulus. In both studies it was demonstrated that it is possible to decrease the rendering quality and computational costs without a perceivable difference to an observer.

In this chapter, the effect of auditory-visual cross-modal interaction on the perceived rendering threshold for high-fidelity graphics is investigated. The hypothesis was that it would be possible to further reduce the number of rays shot per pixel without perceivable degradation in image quality when using auditory stimuli during scene observation.

A part of this study was published in [HAC08].

5.2 Experiment

In order to quantify the perceived rendering threshold of high-fidelity images when sound was present, a psychophysical study was conducted. Additionally, the perceptual quality of the results was analysed using the Visual Difference Predictor (VDP), which predicts the probability of detecting a visual difference between two images [Dal93]. In the study an independent samples design was used. The dependent variable was the perceived quality of the rendered images (Choice). The independent variables were the audio background (Sound), scene and image quality (rpp).

For the auditory stimuli three conditions were considered: no sound, related sound and noise. Each sound group consisted of 28 pairs of images, where each image pair contained a test stimulus and a gold standard image. Each test stimulus was one of the 28 images, generated as a combination of 4 scenes and 7 image rendering qualities, while the gold standard image was a high-quality image rendered at 49rpp. Images in the pairs were shown one after another in a five-slide sequence, see Figure 5.1. All pairs were ordered randomly in order to

avoid bias.



Figure 5.1: Example of the slide sequence from the experiment.

A few decades ago, Peterson and Peterson, in their psychological studies [PP59] demonstrated that in presence of distractors, humans, on average, have problems with remembering even three elements for more than eighteen seconds. Therefore, every picture in each pair was presented for 5 seconds, so that the whole sequence (Figure 5.1) lasted for about sixteen seconds, which is less than the remembering time threshold.

5.2.1 Stimuli

For carrying out the research, several 3D scenes, used as visual stimuli, were rendered as static images and audio samples for the accompanying sounds were generated. To allow our work to be compared to previous results, we chose a subgroup of scenes used by Sundstedt et al. [SDC05] and Aranha et al. [ADCH06], see Figure 5.2. Three scenes represented realistic environments and the fourth, the checkerboard scene, was used as a stress case, because of its high spatial frequency.

For rendering, a modified version of the *Radiance rpict* renderer [War94], developed by Debattista [Deb06], was used. This renderer was developed for selective, progressive and time-constrained rendering. The renderer uses a distributed ray tracing algorithm to calculate global illumination. The sampling algorithm is based on a hierarchical low-discrepancy (0,2) sampling sequence [KK02] composed of the Sobol and van der Corput sequences. Image reconstruc-



Figure 5.2: Scenes used for the experiment: Checkerboard (top left), Corridor (top right), Kalabsha (bottom left), Library (bottom right). See Appendix A for larger images.

tion is performed using a Gaussian filter. This renderer is an improved version of the one used in the similar experiments [SDC05, ADCH06]. All scenes were rendered with 1, 4, 9, 16, 25, 36 and 49 rays per pixel (rpp), at a resolution of 1024×768 pixels. They were later converted to tiff format using Radiance’s `ra_tiff` command.

Using knowledge from previous studies by Mastoropolouou [Mas06], it was decided to use related sound and noise, but not music. The reason for this is that music can introduce subjective side effects, such as emotions, anxiety, boredom, excitement, or similar. As a result, it could affect the perception and response of the participants in the experiment. Related sounds chosen for the experiment are presented in Table 5.1. Different noise sounds: white, brown and pink noises were used in order to decrease the boredom and to avoid the bias as a consequence of the familiarization with the sound, Figure 5.3. These noise signals differed only in their spectral density, which is a power distribution in the frequency spectrum. All sounds were presented binaurally and provided no spatial information about the virtual sound source. The audio level was not measured, but all the sounds were delivered at a plausible level, in order to avoid discomfort by either loudness or quietness.

Checker board	Corridor	Kalabsha	Library
Footsteps	Background chatting	Sounds of nature	Office noises

Table 5.1: Related sounds used for the experiment.

5.2.2 Visual difference predictor

In order to predict the probability of perceiving difference between the pair of images observed by a human, Daly developed the Visual Difference Predictor

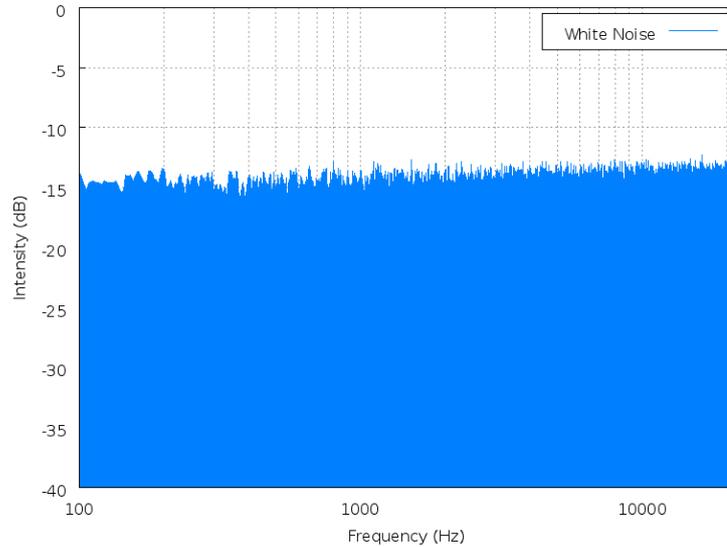


Figure 5.3: White noise frequency spectrum.

[Dal93]. This algorithm uses two images as input - a mask and a target image and compares them, producing a map of perceivable differences and values of detection probabilities, see Figure 5.4. The VDP metric has some shortcomings: it can be used only for low dynamic range (LDR) images and it predicts the global level of adaptation to luminance. With significant progress in HDR applications and its general usage, Mantiuk et al. [MMS04] extended the original VDP algorithm to include the comparison of HDR images, taking into account the local adaptation of the eye to every part of the observed scene and the entire luminance spectrum visible to human eye.

Using HDR VDP, we analysed 24 pairs of 1024×748 images at varying rpp qualities. As in the psychophysical experiments, images with various qualities were compared with the gold standard, except for the image pair 49rpp v 49rpp (for each of the four scenes), because they are identical and there is a zero difference. A suitable gold-standard image was chosen, beyond which there was no significant difference between images. According to this method, it was found



Figure 5.4: An example of the VDP comparison. Top: mask image; Middle: target image; Bottom: difference map with probability of detection - green:0-50%; yellow: 50-75%; red:75-95%; pink:95-100%. See Appendix A for large images.

that at 1024×768 , 49rpp was an ideal gold-standard for the chosen scenes.

5.2.3 Hardware and rendering time

All images were rendered on a PC with a Intel Core2Duo E6650 CPU at 2.33GHz, with 2GB of DDR2 PC6400 RAM memory and GeForce 8800GTS with 640MB graphics card. Rendering times are given in Table 5.2.

The experiment was conducted using Intel Pentium 4 computer running at 3.00GHz, with a Compaq 1825 19" monitor with 1280×1024 pixel resolution and audio stimuli were presented through Sony MDR-V300 sound-insulating Dynamic headphones.

rpp	Checker board	Corridor	Kalabsha	Library
1	13.12	673.99	98.29	130.57
4	51.98	1431.31	170.00	513.40
9	115.81	2471.16	288.62	1153.65
16	206.89	4110.91	409.61	2086.31
25	329.79	5556.64	581.24	3354.27
36	464.08	7807.02	796.40	4627.24
49	812.92	10327.32	1057.14	6277.47

Table 5.2: Rendering times for all scenes presented in seconds.

5.2.4 Procedure

48 volunteers (33 male and 15 female) aged from 18 to 63, with an average age of 25, participated in the user study. Most of them (about 85%) were undergraduate or postgraduate students at the University of Warwick. Other subjects were employees at the same University. Participants were randomly recruited on the University campus and were not paid for participation. All participants reported normal or corrected to normal vision with no hearing impairments. All

use computers in everyday work, and reported average familiarity with computer graphics. All participants were naive about the purpose of the experiment and participated in only one of the randomly selected groups (no-sound, related sound, unrelated sound).

The participants were asked to compare the quality of the images presented in each pair and choose the one they thought contained the higher rendering quality. The test pairs were: 1v49, 4v49, 9v49, 16v49, 25v49, 36v49 and 49v49, where the first image in each pair was the test image and the second the gold standard. This method is known as Two Alternative Forced Choice (2AFC). The display of the stimuli was controlled using a program, developed as part of this research, to provide fixed display time, synchronised audio and collection of the participants decisions. Images were displayed on a black background in a completely dark and noise-isolated room. The observers distance from the monitor was fixed at 60cm and subjects from all three groups were asked to put on headphones. The experiment lasted about 9 minutes.

Prior to the experiment, the participants were asked to sign the consent form, fill in the questionnaire and they were provided with the instructions (see Appendix A). For training purposes, they were then shown a demonstration with two image pairs, which were not used in the study. The first was accompanied by a related sound, while the second was without audio. Images were displayed in pairs showing the same scene and auditory stimulus, with varying rendering levels. After each pair, a question mark appeared (Figure 5.1), in order to cue the participant to press either button 1 or 2 on the keyboard relating to which image, that they believed, contained higher rendering quality. The keyboard was situated within arms-length distance from the observer on the desk in front, so there was no need to move the position of the body or the hand for pressing the buttons. After an answer was provided, the next pair of images was shown. All

results were stored in a text file generated by the program.

Since the 2AFC method requires a user response, even if the difference is not perceivable, after the experiment all participants were asked to say on which features the differences were most obvious and approximately how many of the pairs they perceived as different, to show that they were not guessing.

5.3 Results

In order to analyse the findings, both *statistical analysis* of the psychophysical study as well as *comparison using the VDP* were used.

5.3.1 Statistical analysis of psychophysical experiment

Since all the variables involved in the experiment were categorical (one case falls only into one category, see Table 5.3), it would not make sense to calculate the mean values, because the numeric values attached to different categories are arbitrary, so that the mean of those numeric values would depend only on how many members each category has [Fie09]. Therefore, for categorical variables the frequencies are analysed, which is the number of things that fall into each combination of categories.

Variable	Type	Conditions
Scene	Independent	Checkerboard, Corridor, Library, Kalabsha
Sound	Independent	NoSound, Unrelated, Related
Quality	Independent	1, 4, 9, 16, 25, 36, 49rpp
Score	Dependent	0,1

Table 5.3: Variables used in the experiment.

Since mean values cannot be used, standard Analysis of Variance (ANOVA) could not be used. Instead, a loglinear analysis is used, which is an adaptation

of ANOVA for categorical data. While the same principles are used, the group means model (used in ANOVA) is substituted with the expected frequencies. Therefore, categorical data can be expressed in the form of a linear model provided the log values are used [Fie09]. Loglinear analysis first looks at all the main effects and the interactions between them. Since the predictors can perfectly predict the outcome, there is no error in the model. Therefore, this analysis typically works on a principle of backward elimination, which means that in the next step one predictor (highest-order interaction) from the model is removed and the expected frequencies are calculated to see how well the model fits the data. If the fit of the new model is not significantly different from the previous model, the latter is abandoned. This means that the removed term did not have a significant impact on the ability of the model to predict the observed data. The process is repeated, removing lower-order interactions until an effect that does affect the fit of the model when removed is found.

The loglinear analysis was performed across scenes (all for scenes pooled together) looking at the main effects of Sound, rpp and Choice, and interactions between them. Firstly, the contingency table of the data is created using the crosstabs command in SPSS [SPS]. This table contains the number of cases that fall into each combination of categories. A summary of the contingency table for the sound condition is given in Table 5.4. Graphical representation of the results is shown in Figure 5.5.

Sound condition	TestBetter	GoldBetter	Total
NoSound	12.6%	20.8%	33%
Noise	13.7%	19.6%	33%
Related	13.1%	20.2%	33%

Table 5.4: Summary of the contingency table for the sound condition. TestBetter shows the count of responses preferring the test image over gold standard one. GoldBetter is the count of responses preferring the gold image over the test one.

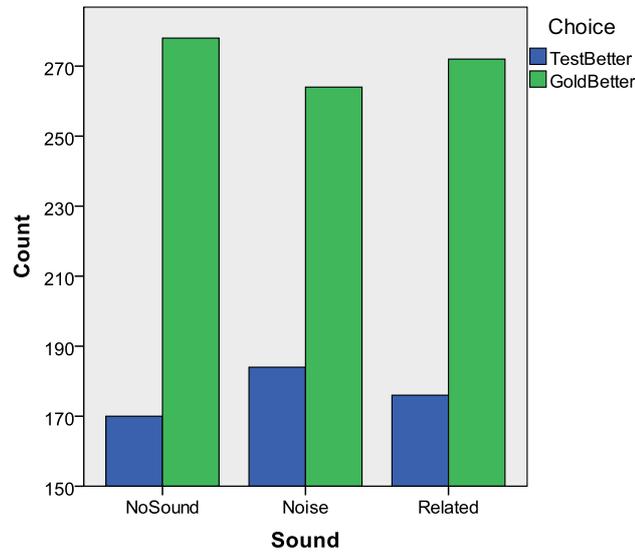


Figure 5.5: The frequencies of the participants' scores across sound conditions.

Since each user group had the same number of participants, the number of responses is divided equally (33% each). The percentages show that more participants rated test images as better quality than gold standard when the background sound was played. The highest count was found for the Noise group (13.7%). Log-linear analysis has one assumption: there should be no expected counts less than 1 and no more than 20% less than 5. Since the sample size was rather high, the assumption was met (the smallest expected count for sound condition was 176.3).

The contingency table for the rpp condition (Table 5.5) revealed the number of cases for each combination in given categories. The table shows that the perceived difference in quality between the test images and gold standard becomes weaker until it completely disappears. The results are illustrated on a chart in Figure 5.6. The lowest expected count in this case was 75.7, which is far more than 5, meaning that the assumption is met.

The results of the K-way effects looks at which components of the model can be removed. There was a significant effect at the first level ($K=1$; $\chi^2=55.14$; $p < .001$), which means that removing the main effects of sound, rpp and choice

RaysPerPixel(rpp) condition	TestBetter	GoldBetter	Total
1v49	1.6%	12.7%	14.3%
4v49	4.5%	9.7%	14.3%
9v49	6.0%	8.3%	14.3%
16v49	6.8%	7.5%	14.3%
25v49	6.4%	7.9%	14.3%
36v49	7.1%	7.1%	14.3%
49v49	7.1%	7.2%	14.3%

Table 5.5: Summary of the contingency table for the rpp condition. TestBetter shows the count of responses preferring the test image over gold standard one. GoldBetter is the count of responses preferring the gold image over the test one.

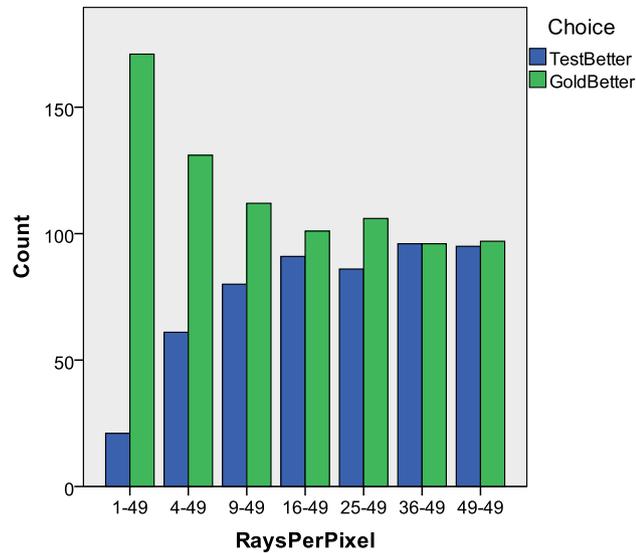


Figure 5.6: The frequencies of the participants' scores across quality conditions (RaysPerPixel).

rpp	Checkerboard		Corridor		Kalabsha		Library	
	χ^2	p	χ^2	p	χ^2	p	χ^2	p
1	7.562	0.0059	14.062	0.0001	7.562	0.0059	14.062	0.0001
4	7.562	0.0059	0.062	0.8033	0.062	0.8033	3.062	0.0801
9	0.062	0.8033	0.562	0.4534	0.062	0.8033	0.562	0.4534
16	0.062	0.8033	0.562	0.4534	0.062	0.8033	3.062	0.0801
25	0.562	0.4534	0.562	0.4534	0.562	0.4534	1.562	0.2113
36	1.562	0.2113	0.062	0.8033	0.562	0.4534	0.562	0.4534
49	0.062	0.8033	0.062	0.8033	0.562	0.4534	0.562	0.4534

Table 5.6: “No sound” group: Chi-Square Analysis (df=1; critical value 3.841 at 0.05 level of significance). Significant results are written in bold.

affects the fit of the model. Similarly, removing the two-way interactions (K=2) between the observed variables has a significant effect on the model ($\chi^2=94.52$; $p < .001$). Finally, removing the three-way interaction between all three factors does not significantly affect the model (K=3; $\chi^2=14.59$; $p = .265$).

Since loglinear analysis is hierarchical, we are interested in the highest level for which the significant effect was found. Therefore, further analysis is required for the two-way interactions (K=2) where $p < .05$. To break down this effect, separate Pearson Chi-square analyses with *Yates’ correction* [Yat34] on the Quality and Score variables were performed separately for scenes and sounds [Fie09]. The analysis compared participants preference for various image pairs (Test v Gold standard). The null hypothesis for each pair of images was that they should have equal preference. Since no bias was assumed, it was expected that the gold-standard self test (49rpp v 49rpp) should deliver equal preference. Computed Chi-square values are presented in Tables 5.6, 5.7 and 5.8 for each scene, divided into no sound, noise and related sound conditions respectively.

According to this probability measure, it was found that the perceivable difference decreased monotonically with the increase in the number of rays-per-pixel, see Figure 5.6. In addition, a significant difference in perceived rendering quality,

rpp	Checkerboard		Corridor		Kalabsha		Library	
	χ^2	p	χ^2	p	χ^2	p	χ^2	p
1	10.562	0.0011	3.062	0.0801	3.062	0.0801	10.562	0.0011
4	7.562	0.0059	0.562	0.4534	1.562	0.2113	3.062	0.0801
9	1.562	0.2113	0.562	0.4534	0.062	0.8033	0.062	0.8033
16	0.062	0.8033	0.062	0.8033	0.062	0.8033	0.062	0.8033
25	0.062	0.8033	0.562	0.4534	0.562	0.4534	0.062	0.8033
36	0.062	0.8033	0.062	0.8033	1.562	0.2113	0.062	0.8033
49	0.062	0.8033	0.562	0.4534	0.062	0.8033	0.062	0.8033

Table 5.7: “Noise” group: Chi-Square Analysis (df=1; critical value 3.841 at 0.05 level of significance). Significant results are written in bold.

rpp	Checkerboard		Corridor		Kalabsha		Library	
	χ^2	p	χ^2	p	χ^2	p	χ^2	p
1	14.062	0.0001	10.562	0.0011	5.062	0.0244	5.062	0.0244
4	1.562	0.2133	0.562	0.4534	1.562	0.2133	5.062	0.0124
9	5.062	0.0244	0.062	0.8033	0.562	0.4534	0.062	0.8033
16	1.562	0.2133	0.062	0.8033	0.062	0.8033	0.062	0.8033
25	0.062	0.8033	0.062	0.8033	0.062	0.8033	1.562	0.2133
36	0.062	0.8033	0.562	0.4534	0.062	0.8033	0.562	0.4534
49	0.062	0.8033	1.562	0.2133	0.062	0.8033	0.062	0.8033

Table 5.8: “Related sound” group: Chi-Square Analysis (df=1; critical value 3.841 at 0.05 level of significance). Significant results are written in bold. Results show that the subjects were looking more closely, and were able to find more differences.

between the test images and the gold standard, occurs in all scenes at less than 4rpp for the *no sound* and *related sound* groups. In the third group, where *noise* was used as an audio stimulus, the subjects were unable to perceive any difference even using the 1rpp comparison for the Corridor and Kalabsha scenes. For the other two scenes the perceived threshold was the same as in the *no sound* group.

The results were consistent across the majority of the scenes. As expected from the Checkerboard scene, the threshold was slightly greater than for the other scenes because of its high spatial frequency characteristics.

By comparing the results of the Chi-square across each of the three groups (no sound, related sound and noise), it can be seen that the *noise* threshold was lower than for the other conditions which confirmed the research hypothesis. However, for the *related sound* group, the research hypothesis cannot be accepted. Furthermore, a higher threshold level was reported, which may indicate that a related sound during the experiment made participants look at the scene more closely.

5.3.2 Comparison using VDP

As discussed in Section 5.2.2, VDP can be used to highlight perceived differences as if viewed by the HVS. A major constraint of the VDP, however, is that it assumes significant viewing time. Using a comparison between the psychophysical study and the VDP analysis, potential differences, which may have resulted from the finite viewing time, were verified.

In order to verify the results of the psychophysical study, the gold-standard (49rpp) was compared with all other images of the same scene. The results of the VDP error measurements are presented in Table 5.9. These results were slightly different from the outcome of the statistical analysis.

The results of the VDP comparison of the Kalabsha scene indicated significant

RaysPerPixel(rpp) condition	Board	Corridor	Library	Kalabsha
1v49	9.35%	2.85%	5.85%	14.36%
4v49	6.35%	1.27%	0.7%	9.1%
9v49	2.49%	0.55%	0.12%	6.64%
16v49	0.59%	0.27%	0.01%	7.22%
25v49	0.92%	0.53%	0.01%	6.62%
36v49	0.12%	0.26%	0.01%	6.74%

Table 5.9: VDP comparison of perceivable differences for all scenes showing the percentage of pixels on which the probability of perceiving that difference is the highest.

variance between pairs of images. However, the Chi-square results showed little statistical preference between images for this scene. The differences between images in this scene are likely to be the result of texturing method of the model and the sampling strategy used by the renderer. This difference was picked up by VDP possibly due to HDR VDP’s increased conservativeness, as described in the research by Ramanarayanan et al. [RFWB07]. For other scenes, the perceptual threshold was significantly higher, in some cases more than ten times. Additionally, Table 5.9 shows that the difference was more likely to be perceived for the Checkerboard scene, as was the case in our psychophysical experiment.

From post-experiment questioning of participants, approximately half of the subjects reported noticeable artefacts occurring on the edges of objects. This is verified by the VDP maps, produced as seen in Figure 5.4. These regions are a consequence of the reduced sampling, caused by tracing fewer rays in regions with high spatial variation.

5.4 Discussion

The results of the statistical analysis indicate that there is a significant main effect of observed factors and a significant two-way interaction between them.

In addition, it has been shown that cross-modal interaction has a significant effect on perceived thresholds. The presence of unrelated sounds, in this case noise, acts as a distractor as expected, producing a reduction in the thresholds. However, it was found that in the case of related sounds, a greater threshold was discovered which may indicate a heightened awareness of differences between rendering qualities. A possible reason for this could be that the visual focus is at specific regions, in which case it may be tolerable to further reduce rendering costs by using selective rendering techniques. Furthermore, different results were found when using different scenes, which indicates that the scene properties should be controlled and could be a subject of further investigations.

Such results confirm that, for scenes with unrelated sounds, it is possible to exploit such accompanying audio by tracing a lower number of rays, and therefore reducing rendering time, without noticeable degradation in image quality. However, due to the nature of the cross-modality, defined by the modality appropriateness hypothesis [HT66, WW80] and auditory driving effect [GM59, Shi64, WKN03, Rec03] the research presented in the following chapters focuses on temporal domain.

CHAPTER 6

Beat Rate Effect on Frame Rate Perception

6.1 Introduction

In current interactive virtual environments, such as video games, typically, more computation time is spent on computing compelling visuals than calculating audio. This is because much of the audio may be pre-recorded. This is particularly true when considering physically-based illumination and character animation, which may have to be computed on a frame by frame basis. In virtual environments, audio, for example music, may be used to create an emotional involvement with the simulation. If the influence of such audio can be exploited, it may be possible to reduce the computation required for the visuals without reducing the perceived visual quality. However, the full emotional influence on a listener caused by music is a very complex phenomenon [PCK07]. Therefore, it is more convenient to investigate the more straightforward relationship between specific aspects of audio and video that function in the temporal domain, in particular the beat rate and the frame rate. If understood and harnessed correctly, this relationship should make it possible to have a graphics engine that can change the

beat rate and the frame rate on-demand to reduce or balance its work load when required. This would effectively reduce the computational time of rendering, without the user noticing any perceptual loss in quality.

In this chapter, the experimental procedure, results and discussion of this work, published in [HCD*09], are presented.

6.2 Experiments

The aim of the study is to investigate how audio and beat rate can influence frame rate perception. In the evaluation of smoothness perception of video animation, multiple independent variables were used. The first is scene complexity which involves many dimensions, such as geometry, materials and lighting complexity, as described by Gero and Kazakov [GK04] and Ramanarayanan et al. [RBFW08]. Gero and Kazakov use semantic vertex graphs for visual complexity calculation. A scene is first turned into an outline model representation from which a semantic vertex graph can be generated. From this resulting graph, visual complexity can be calculated using structural information in the form of labels, circuits and edges [GK04]. Ramanarayanan et al. looked at the dimensionality of visual complexity, showing that each subject seems to approach the problem of visual complexity differently and had different complexity spaces. However, they inferred that generally the complexity can be represented using two or three dimensions: numerosity and material/lighting complexity [RBFW08]. Secondly, in computer graphics the virtual camera can be fixed or moving throughout the virtual environment. Camera movement can be in the form of translation, rotation or panning, or any combination of these movement types. Furthermore, oscillating motions of the camera can be added to improve the sensation of walking/running in a virtual environment [LBHD06, HDAC10, HDAC11]. Another

parameter that is of a great importance in determining the complexity of the visual images presented to the subjects is the speed of the camera movement. All this applies to both static and dynamic scenes, which doubles the number of experimental conditions.

For the auditory stimuli, sounds which should not introduce any strong subjective side effects to the participants, such as emotions, anxiety, excitement, boredom, or similar were chosen. Any of these factors could influence the perception and response of a participant during the psychophysical experiment [MC04].

There are some other factors that could have been considered, such as age, sex, familiarity with the depicted content, etc. However, these have not been included as they have not been reported as significant in similar studies. Furthermore, inclusion of these factors would lead to significant increase in the experimental design and analysis complexity.

6.2.1 Design

In the study, the perceptual responses of participants were evaluated in a complete randomised design. Static and dynamic scenes were studied separately, with four static and two dynamic scenes (Figure 6.1), using a within-participant design. The former has a fixed scene with a moving camera, while the latter contains some moving object(s) with fixed camera. All the scenes were generated by the thesis author. For both scene groups, the following effects were examined: frame rate, to see if the selected frame rates were appropriate; scene, to see how it can affect the smoothness perception; and beat rate, to see if introduction of audio with different beat rate can affect perceived temporal visual quality when observing an animated content. The scene complexity contains three major elements: scene content, rendering technique used for image generation and camera movement type (for static scenes). This shall be referred to as the *scene* factor,

containing four static scenes named: *Kalabsha*, *Kiti*, *Kiti-mentalRay* and *Rabbit*; and two dynamic scenes: *Cars* and *People*. Four different frame rates were considered: 10, 15, 20 and 60fps (frames per second). These conditions will be referred to as *FR10*, *FR15*, *FR20* and *FR60* respectively. For static scenes three different audio conditions were used: no sound, 2 and 6bps (beats per seconds) while for dynamic scenes we used an additional audio condition - 4bps. These conditions will be referred to as *BR0*, *BR2*, *BR4* and *BR6*. To measure the perceptual responses of participants a single stimulus non-categorical method was used [ITU], also known as Interactive Rating Scale method in sound quality studies [GJF*06]. This method uses a slider bar with a scale using numbers from 0 to 100 and adjectives on both ends, such as *smooth* and *jerky*. Additionally, a post-hoc pairwise comparison was conducted, to see the difference between the ratings for a given effect.

Finally, the interaction between the independent variables was investigated. The results of this analysis can provide information about the occurrence of ranking patterns when using certain combinations of the tested factors, e.g. if users gave a higher ranking for all scenes when using high beat rates, or if they performed lower rankings when viewing one scene at any frame rate.

The hypothesis was that each factor (frame rate, scene and beat rate) affects the perception of the frame rate.

6.2.2 Participants

99 people volunteered in two sets of experiment, 87 of whom were university students studying a variety of subjects, and the rest were university staff. Out of the 99 participants 75 were male and 24 female. The participants were aged between 18 and 46. The average age was 23. All of them reported normal or corrected to normal vision and no hearing impairments. All participants were



(a) Static scenes: Kiti (top left), Rabbit (top right), Kalabsha (bottom left), Kiti-mentalRay (bottom right).



(b) Dynamic scenes: People (left) and Cars (right).

Figure 6.1: A sample frame from each of the animations used in the study. See Appendix B for larger images.

recruited randomly at the Universities in Sarajevo, Bosnia and Herzegovina and Vila Real, Portugal. The participants were not paid for their participation.

6.2.3 Apparatus

The experiments were conducted in a dark, quiet room. In the first experiment, the visual stimuli were presented on a Dell E198FPB 19" monitor with 1280×1024 pixel resolution and a refresh rate of 60 Hz. In the second set of experiments an LG W2234S 22" monitor with a refresh rate of 60 Hz at 1680×1050 was used. Different monitor types were used since the experiments took place at different locations. However, the physical size of the presented stimuli and the experimental environments were the same. The stimuli were positioned at eye level, 60-70 cm from the participants' eyes. The resultant physical stimuli dimensions were the same in both experiments. A LTB Magnum 5.1 AC97 Headphone set was used for audio stimuli.

6.2.4 Stimuli

In order to control for fatigue and to analyse the behaviour of the participants according to scene variations, in the first experiment four different animations were used, see Figure 6.1a. In the second experiment another two dynamic scenes: (*Cars* and *People*) were used, see Figure 6.1b. For static scenes, four different camera movements were considered: translation for Kiti scene, panning for Kiti-mentalRay scene, rotation around own axis for Kalabsha scene, rotation around the object for Rabbit scene. For all translational types of camera movement the same speed was used, which corresponds to a young person's average normal walking speed of 1.425 m/s [AONI04]. The same speed was used, because different speeds could affect motion perception, and thus bias the results [SPP00].

Additionally, for dynamic scenes, a static camera was used, with one or more moving objects. For both dynamic scenes the same virtual environment was used, with cameras at different locations. The Cars scenes consisted of a bus and a few cars moving down a street. The bus stopped within the frame for the last 3.5 seconds of the animation. In the People scene three people entered the frame from the right side and walked towards the camera until they eventually stopped in front of it. All scenes were modeled using commercial modeling software Autodesk Maya 8.5.

For auditory stimuli, sounds which should not have an emotional effect on participants, but do have a rhythmical significance were presented. In order to minimise the synthetic nature of the sound, a looped rhythmical sample created using Propellerhead Reason 4 software with two different conga kick sounds was used. All sounds were produced using two channels (stereo), a sample rate 44100Hz and a bit rate of 1411kbps. These audio samples varied just in the beat rate. Both audio and video files were uncompressed. The correlations between beat and frame rates are shown in Figure 6.2.

6.2.5 Procedure

Both sets of experiments were conducted using the procedure described below. Prior to the experiment the participants were asked to read and sign a consent form. They were then asked to read the instructions, and the experimental procedure and how to rate the smoothness of the animations was verbally explained (see Appendix B). Figure 6.3 was used to illustrate the effect of frame rate to the participants.

After clarifying that participants understood the nature and the purpose of the experiment, they were shown a sample animation (Figure 6.4) at 10 and 60fps with no sound, and told that these are the worst and the best cases respectively.

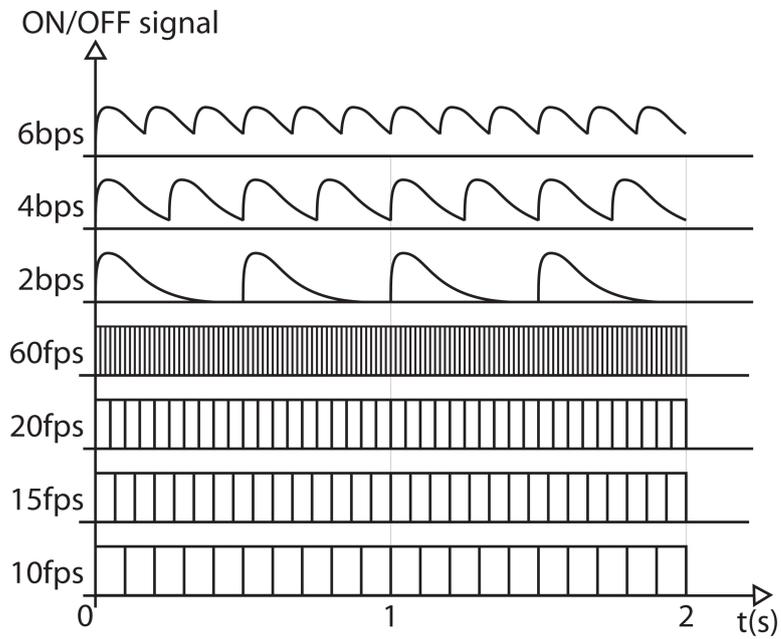


Figure 6.2: Correlation between beat rates (bps) and frame rates (fps). The diagram shows the number of frames that fit within a beat.

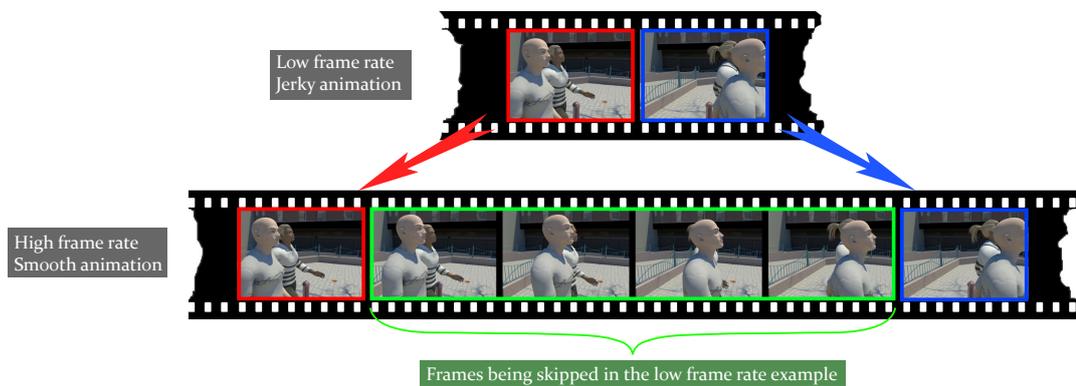


Figure 6.3: An illustration of the frame rate concept used in the instructions.

Participants were not told what frame rates were involved in either the experiment or the sample animations. After the training phase they were asked to rate the smoothness of each animation using a slider bar, see Figure 6.5.

In the first experiment 49 participants were shown 60 randomly ordered animations, while in the second experiment another 50 participants evaluated 32 randomly ordered animations. The results from the first five animations in each session were shown to control for the effects of practise. The data from the dummy presentations was not taken into account when analysing the results of the experiments. A slider bar, ranging from 0 to 100, as recommended in [ITU], was shown at the bottom of the screen as shown in Figure 6.5. After each animation, the slider bar was set to the middle of the bar, i.e. to the value 50. The time for the evaluation of each animation was restricted to 5 seconds. The next animation started automatically after the evaluation period. If a participant has not moved the slider, its initial value was recorded. The total trial time for the first experiment was 16 minutes and 15 seconds and for the second experiment 9 minutes and 15 seconds.

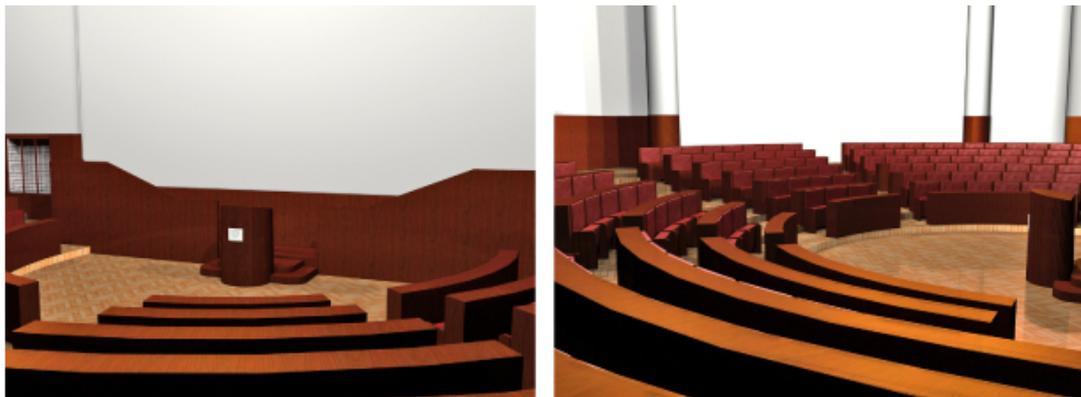


Figure 6.4: Two frames from the sample animation.

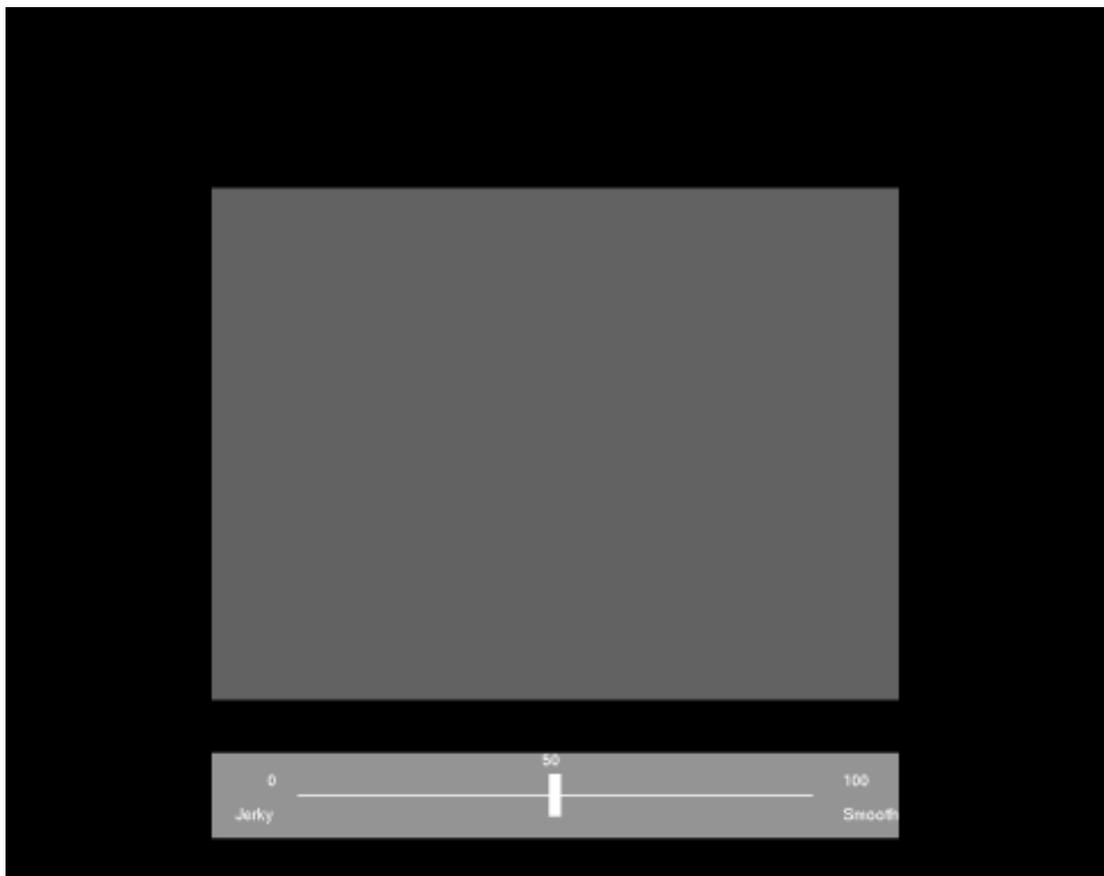


Figure 6.5: Preview of the slider bar used in the experiment.

6.2.6 Analysis methods

A commonly used quality and perception evaluation method is pairwise forced choice comparisons. However, this method has some disadvantages, such as the reported subjective preference may depend on the order of the presented videos and the experiment is time demanding (the number of comparisons increases rapidly with the inclusion of new levels of factors e.g. with 2, 3, 4, 5 scenes we need 2, 6, 12, and 20 comparisons respectively). Despite these shortcomings, the forced choice pairwise comparison is often used (especially in static pictures with no camera movement) and the statistical methodology is well established for such comparisons.

In this study we used a slider bar to measure perceived smoothness of the animations. The main challenge of slider bars is that participants may use it in a different way introducing another source of between participant variability into data. In our study about 5-10% of participants tended to use only the middle 50-80 of the 0-100 scale, while others used the whole scale. The reported analysis of slider bar data, reported by Allman-Ward et al. [AWWDJ04] gives a good example of how the results can be analysed via exploratory data analysis, using statistics such as the mean and standard deviation. However, this does not show whether the observed differences are significant i.e. the differences are due to the effect of a factor or are due to the random variability in the subjective feedback collected. Subsequent sound perception studies suggest using Analysis of Variance (ANOVA) [GJF*06].

For the data analysis the repeated-measures ANOVA method was used in SPSS 17.0 [SPS]. Generally, the analysis of variance is used for testing differences between several means. This method assumes similarity between scores in different treatment conditions, or, put in other words, it assumes that the vari-

ances between pairs of scores are roughly equal. This assumption is called the sphericity assumption, and if violated, requires a correction of F-values. Most often Greenhouse-Geisser correction is recommended [Fie09]. This correction (usually denoted as $\hat{\epsilon}$) varies between $1/k-1$ and 1 (k equals to the number of repeated-measures conditions). The closer the $\hat{\epsilon}$ is to 1, the more homogeneous the variances of differences are, and hence the closer the data are to being spherical.

The statistical analysis was conducted on each user group (static and dynamic scenes) separately. The same procedure was undertaken in both analyses. The effect of each factor on the perception of smoothness was studied in two steps. First, a main effect of the factor was tested using the within participants test. This test reveals if a factor affects the dependent variable, assuming the other factors are fixed. Within the same test the interaction between the independent variables was analysed. The following combinations were considered: scene-beat rate, scene-frame rate and frame rate-beat rate.

In the second step a post-hoc pairwise comparison was conducted, looking at the within factor relationship and the variability of the influence of each condition. For the multiple comparisons between the independent variables, where all combinations of groups have to be tested, a familywise error arises. In order to control for this error, by correcting the level of significance for each test, such that the overall Type I error rate across all comparisons remains at .05, Bonferroni correction was used. Although there are other post-hoc tests, including Tukey's test, the Bonferroni method is the most robust and controls the alpha value regardless of manipulation.

6.3 Results

The statistical analysis is conducted onto two data sets, acquired separately from the two experiments. In Section 6.3.1 the results from the first experiment, using static scenes, are presented. Section 6.3.2 presents the results of the second experiment, where dynamic scenes were utilised. In both analyses the main effect of frame rate (FR), beat rate (BR) and scene were analysed. In addition, the post-hoc pairwise comparison was utilised to see the interaction between the conditions for each factor.

6.3.1 Static scenes

For analysing the data, a $4(\text{FR}) \times 3(\text{BR}) \times 4(\text{scene})$ repeated-measures ANOVA was used. Since all independent variables completely cross over, there were 48 experimental conditions for each participant.

The results of Mauchly's sphericity test, which is one of the assumptions for repeated-measures ANOVA, showed that the assumption was met only for the effect of beat rate ($p = .286$) and for the interaction between scene and beat rate ($p = .968$). For the other two effects and interactions the assumption was violated ($p < .05$), so Greenhouse-Geisser correction was applied.

The within participant test, with corrected F-values, showed a significant main effect of Scene ($F=52.087$; $p < .001$). Figure 6.6 shows a mean values for each Scene, across frame rates and beat rates. From this graph, it is clear that Rabbit scene was rated the highest (59.96) and Kalabsha the lowest (43.59). Furthermore, the pairwise comparison for the main effect of Scene, corrected using a Bonferroni adjustment, shows a significant difference between ratings of all scenes, except for the Kiti and KitiMR pair ($p = .894$).

The corrected significance values from the ANOVA test indicate that main

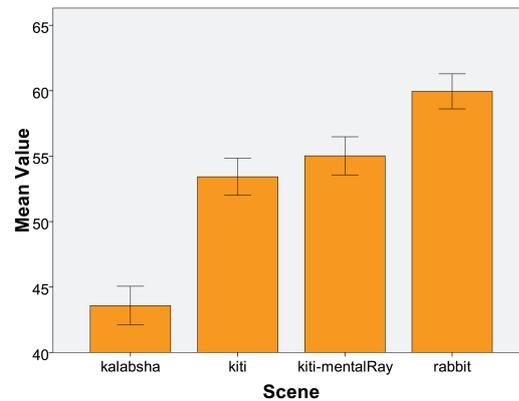


Figure 6.6: Mean values of subjective scores across static scenes with standard error. All frame and beat rates are pooled.

effect of frame rates was also significant ($F=478.54$; $p < .001$). This means that if the scene and BR are ignored, participants' ratings would differ according to the frame rate used. From the graph in Figure 6.7 it is clear that higher ratings were given to higher frame rates. This is further confirmed by the pairwise comparison for the main effect of frame rate, corrected using the Bonferroni adjustment. The comparison showed a significant difference between each level ($p < .001$). This confirms that the chosen frame rates were suitable, i.e. that the differences between each frame rate was distinguishable.

Since Mauchly's sphericity test showed no significant for the effect of beat rate, no correction was needed. The within participants test showed a significant main effect of beat rate ($F=8.23$; $p = .001$). This means that if the frame rate and type of scene are ignored, participants would have significantly different ratings for each audio condition. Figure 6.8 shows the mean ranking values for each beat rate when scenes and frame rates are pooled. The graph shows that BR2 increased and BR6 decreased the perceived animation smoothness comparing to the no sound condition. The pairwise comparison for the main effect of beat rate showed a significant difference between ratings of BR2 and BR6 ($p = .002$), but

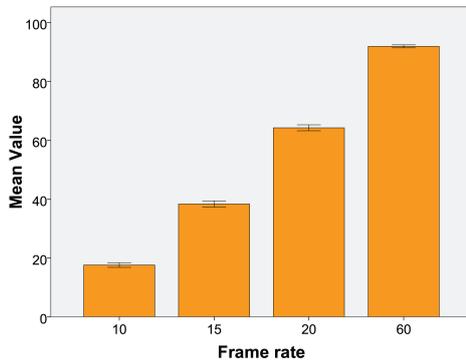


Figure 6.7: Mean values of subjective scores across frame rates with standard error. All static scenes and beat rates are pooled.

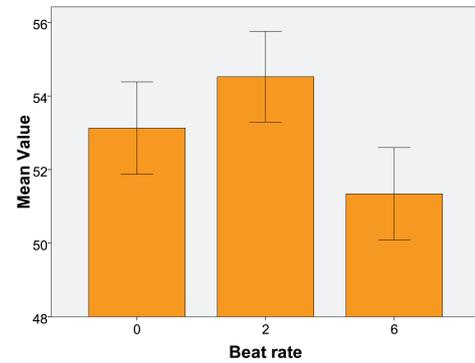


Figure 6.8: Mean values of subjective scores across static scenes with standard error. All static scenes and frame rates are pooled.

not between no sound and BR2 ($p = .172$) or BR6 ($p = .074$).

Looking at the interactions between the variables, only the interaction between FR and Scene was found to be significant ($F=35.85$; $p < .001$), see Figure 6.9. This means that for various scenes, different frame rates are ranked differently. Although, the ranking pattern exists, we can see that the FR15 was ranked significantly higher for Rabbit than for Kalabsha scene. Furthermore, for the Rabbit scene, FR10 and FR15 were ranked higher than FR15 and FR20 respectively, when observing the Kalabsha scene.

Not finding significant interaction between either Scene and Beat rate or Beat rate and Frame rate means that a ranking pattern exists. Looking at the graph in Figure 6.10, it can be seen that, when all frame rates are pooled, the BR2 was always ranked higher than BR0 (no sound). In addition, for all the scenes, BR6 was ranked lower than BR2, while only for the Kalabsha scene it was not ranked lower than BR0.

Similarly, on the graph in Figure 6.11, almost identical patterns can be seen. There are two differences: for FR60, the no sound condition was ranked the highest; for the FR10, BR6 was ranked as lower than BR0.

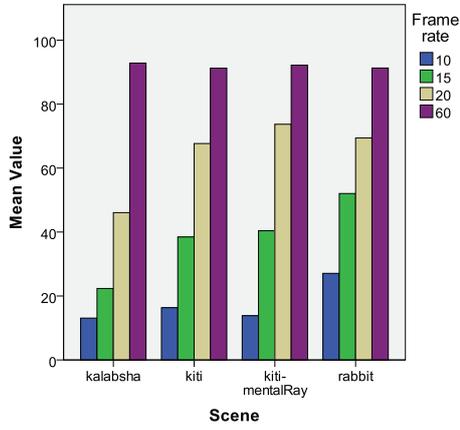


Figure 6.9: Mean values of subjective scores across static scenes and frame rates with standard error.

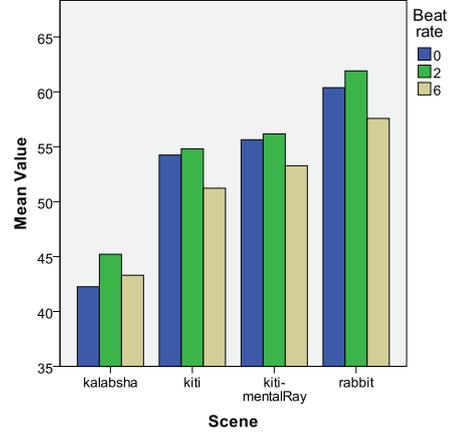


Figure 6.10: Mean values of subjective scores across static scenes and beat rates with standard error.

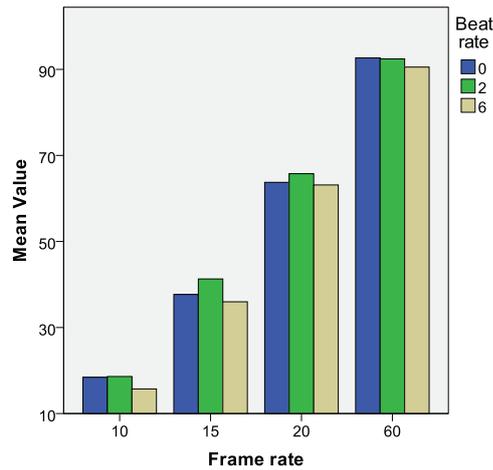


Figure 6.11: Mean values of subjective scores across frame rates and beat rates with standard error.

6.3.2 Dynamic scenes

The data of the second experiment was analysed using a $4(\text{FR}) \times 4(\text{BR}) \times 2(\text{scene})$ repeated-measures ANOVA. As in the first experiment, there were four frame rates (10, 15, 20 and 60FPS). Since, in the first experiment, a significant difference in users' ratings was found between BR2 and BR6, in the second experiment another beat rate was added, to see the behaviour between the two extremes (BR2 and BR6). Therefore, there were four beat rates in total (0-no sound, 2, 4 and 6BPS). The main difference between the second and the first experiment was the nature of the scenes. In this experiment two dynamic scenes were used (Cars and People) with moving objects and fixed camera. There were 32 experimental conditions for each participant. The main effect of each factor was investigated. Additionally, a post-hoc pairwise comparison was conducted, to see the difference between the ratings for each condition of a given effect. Again, the dependent variable was the perceived animation smoothness.

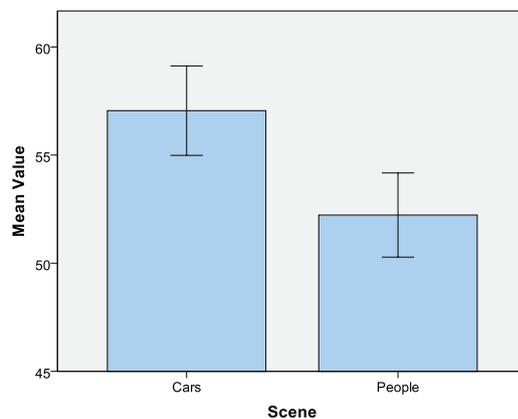


Figure 6.12: Mean values of subjective scores across dynamic scenes with standard error. All frame and beat rates are pooled.

The results of Mauchly's sphericity test, for dynamic scenes, showed that the assumption was violated for the effect of frame rate ($p < .001$), beat rate ($p < .001$) and for the interaction between frame rate and beat rate ($p = .023$).

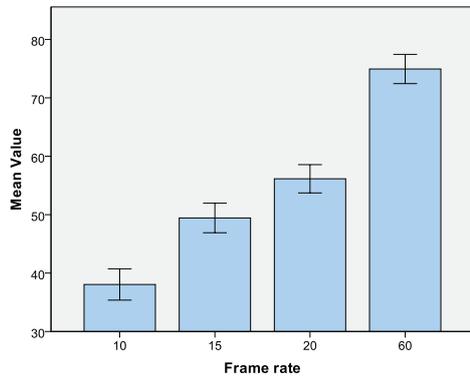


Figure 6.13: Mean values of subjective scores across frame rates with standard error. All dynamic scenes and beat rates are pooled.

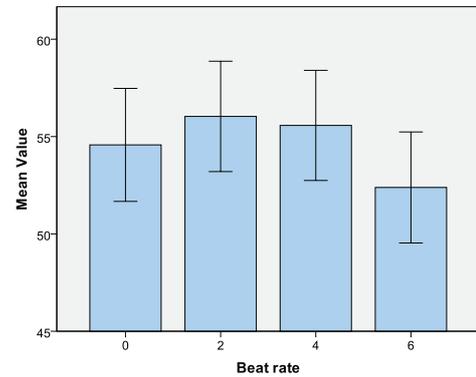


Figure 6.14: Mean values of subjective scores across beat rates with standard error. All dynamic scenes and frame rates are pooled.

For these cases Greenhouse-Geisser correction was applied. The assumption was met for the other effects and interactions.

The within participant test showed the significant main effect of frame rate ($F=51.497$; $p < .001$) and scene ($F=4.628$; $p = .036$). In case of the frame rate effect the F-value was corrected using Greenhouse-Geisser method. From Figure 6.12, showing a mean values for both scenes across frame rates and beat rates, it is clear that Cars scene (Mean=57.06) was more preferred than People scene (Mean=52.23). The pairwise comparison, since there were only two scenes, showed the same significance level as the test for the main effect of scene ($p < .05$).

Testing the main effect of frame rate, with corrected F-values, it was found that the frame rates significantly affect user perception of the animation smoothness ($F=51.497$; $p < .001$), meaning that if the scene and beat rates are ignored, participants' ratings would differ according to the frame rate. The graph in Figure 6.13 shows that the higher the frame rate was the value of the dependent variable. The pairwise comparison, corrected using the Bonferroni adjustment, confirmed this showing the significant difference ($p < .05$) between each frame rate pair.

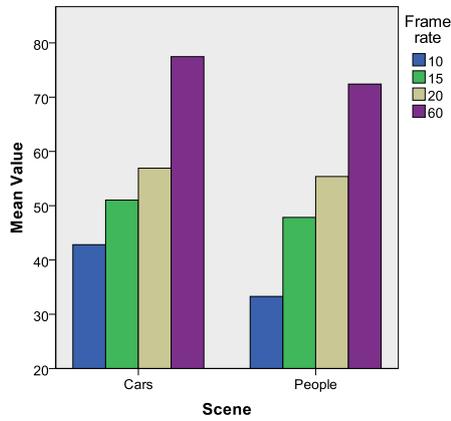


Figure 6.15: Mean values of subjective scores across dynamic scenes and frame rates with standard error.

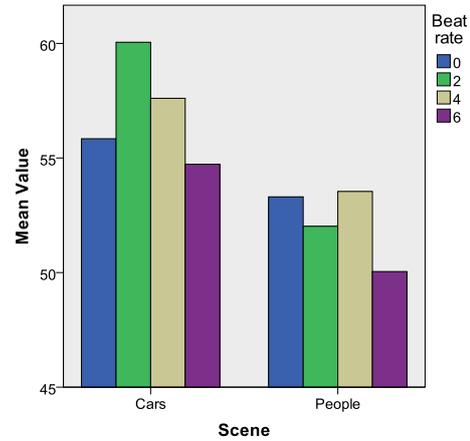


Figure 6.16: Mean values of subjective scores across dynamic scenes and beat rates with standard error.

Although Figure 6.14 shows the same ranking pattern for the mean values of different beat rates as for the static scenes, when scenes and frame rates are pooled, the within participants test, using Greenhouse-Geisser test showed no significance for the effect of beat rate ($F=1.592$; $p = .208$). The graph shows that BR2 and BR4 increased the perceived animation smoothness comparing to the no sound condition, while BR6 had the opposite effect, decreasing the perceived smoothness. The pairwise comparison for the main effect of beat rate showed no significant difference between any beat rates ($p > .05$).

Using the same test, no significant interaction between the variables was found. Figure 6.15 illustrates the interaction between the scenes and frame rates. Although we can see a slight tendency towards Cars scene, especially for FR10, this difference was found as insignificant ($F=2.512$; $p = .06$).

As for the static scenes, there was no significant interaction found between scenes and beat rates ($F=1.442$; $p = .23$). While for the case of Cars scene, the same ranking pattern was found (low frame rates increase and high beat rates decrease the perceived animation smoothness), the mean values for People scene indicate that both low and high beat rates decrease the perceived smoothness,

see Figure 6.16. Mean rankings for the no sound condition and BR4 were almost identical.

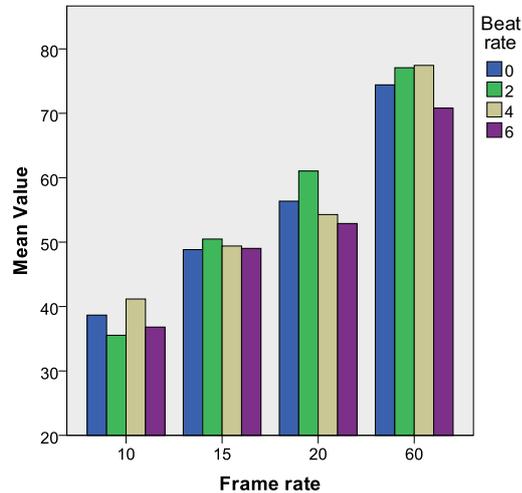


Figure 6.17: Mean values of subjective scores across frame rates and beat rates with standard error.

Finally, the interaction between frame rates and beat rates was not found as significant ($F=1.711$; $p = .08$). In Figure 6.17 a similar pattern can be identified, except for the FR10 where both low and high beat rates decreased the perceived smoothness.

6.4 Discussion

In this study the main effects of beat rate, frame rate and scene on the perception of animation smoothness were investigated. Additionally, the effect within and between them was investigated. The investigation was done by conducting two sets of experiments with two different participant groups, which resulted in having two separate within participant statistical analyses: for static and for dynamic scenes.

Since multiple factors and a correlation between them existed (each partici-

pant gives multiple scores), it was important to choose a statistical method which allows testing of one factor while controlling for the effect of the others. The analysis for each of the groups (static and dynamic) consisted of two stages. In the first stage main effects of all three factors were tested using the within participants tests. Additionally, the interaction between the factors was analysed. In the second stage, a post-hoc pairwise comparison for each factor was conducted. This was explained in more details in Section 6.2.6.

The results of the analysis for static scenes showed a significant effect for each factor (frame rate, beat rate and scene). Although a different rating of the animations played at different frame rates might be considered as obvious, it was important to make sure that the chosen frame rates were adequate, i.e. distinguishable. The motivation partially lied in the Wertheimer theory, in which it is claimed that the HVS tends to blend the successive images shown with a short delay between presentations (between 1/10 and 1/40 of a second) - a phenomenon known as *the apparent motion* [BS06]. Another reason for choosing these frame rates is the fact that computer games usually run at 60fps and more, and that standard computer display refresh rate is 60Hz. All this indicates that frame rates between 10fps and 60fps are the ones to be considered. Therefore, the significance of the main effect of frame rate, found by analysing the data, means that ignoring the effect of beat rate and scene, participants were able to distinguish all frame rates used in the experiments. Having this in mind, the interaction between the beat rate and frame rate was investigated.

Since beat rate was found as significant factor, a further analysis was performed, looking at the pairs of beat rates and the interaction between them. Looking back at Section 6.3.1 it can be seen that there is a significant difference in rating animation smoothness when using a slow beat rate (BR2) and a fast beat rate (BR6). Interestingly, the audio with slow beat rate tended to

increase the perceived animation smoothness, while BR6 had the opposite effect. Although it is hard to explain this behaviour with certainty, the effect of low beat rates might be a consequence of the attentional limitations, i.e. inter-modal attentional model [SJ01]. This model proposes that our attention operates on a global level and is not divided across multiple senses. This means that the performance in a task requiring attention for one modality will be affected by concurrent tasks in some other modality. On the other hand, the negative effect of high beat rate on smoothness perception might be a result of the auditory driving effect [GM59, Shi64] or temporal ventriloquism [MZSFK03, BA03, AB03], or the combination of the two. These phenomena advocate that the auditory stimuli can drive the temporal perception of the visual stimuli.

For the third independent variable, the scene factor, a significant main effect was found. The results indicate different perception of animation smoothness for different scenes. As explained in Section 6.2 this factor has many dimensions, such as virtual scene complexity (geometry, materials and lighting) and virtual camera behaviour (movement type and speed). Therefore, although these findings are important, showing that smoothness perception highly depends on the scene being observed, a further investigation that would look at all these mentioned parameters separately would be useful to further understand the complexities of these variables. A segment of this investigation, showing that the camera movement speed affects the animation smoothness perception is presented in Chapter 7.

For the dynamic scenes, no significance of the beat rate effect was found. This again confirms the importance of the scene factor, as discussed in the text above. The other two factors (frame rate and scene) were found as significant.

The overall results indicate that the effect of frame rate, beat rate and scene on the smoothness perception of an animation exists. However, since this is the

first study to investigate the influence of beat rates on smoothness perception, more work is needed in the future to find a direct relationship between them.

CHAPTER 7

Exploiting Audio-Visual Cross-Modal Interaction

7.1 Introduction

As seen in the previous chapters, related and unrelated sound effects differently affect visual perception (Chapter 5), while camera movement might be a significant factor in perceiving animation smoothness (Chapter 6). Based on these findings, a follow up study, presented in this chapter, has been conducted, focusing on the following factors:

- the influence of camera movement speed on temporal visual perception
- the effect of movement-related sound effects and frame rate perception threshold
- the efficacy of such cross-modal interaction using the sound effect of footsteps associated with the movement being performed

Part of the work presented in this chapter has been published in [HDAC10, HDAC11].

Exp.	Test	Observed effect	Compared frame rates	
Exp. 1	Test 1	Camera movement speed	Audio	NoAudio
			60r vs 60w	60r vs 60w
	Test 2	Perceived smoothness threshold	Audio	
			10r-60w, 20r-60w, 30r-60w vs 60r-60w	
Exp. 2	Test 3	Movement related sound effect	Audio vs NoAudio	
			Run	Walk
			10-20, 10-30, 10-60 20-30, 20-60, 30-60	10-20, 10-30, 10-60 20-30, 20-60, 30-60

Table 7.1: The details of the experimental design for each test. Numbers represent frame rate, while “r” and “w” stand for *running* and *walking* animations respectively.

7.2 Experiments

In this study, the effect of two factors on visual perception were investigated, by conducting two sets of experiments (*Exp1* and *Exp2*), see Table 7.1. Two tests were performed with the data from the first experiment and one test with the data from the second experiment. In the first test (*Test 1*) the effect of camera movement speed (walking and running) on temporal visual perception was investigated. The research hypothesis was that the speed of the camera movement will affect smoothness perception. Additionally, in *Test 2*, the perceived smoothness threshold for the animations accompanied by the audio effects was investigated. The research hypothesis was that there will be difference between the preference of the discrepant frame rates and preference of the control condition (60r-60w). In the second experiment (*Test 3*), the movement related sound effect on the running and walking animations were investigated separately.

The research hypothesis was that scene related audio effects will increase visual smoothness perception.

7.2.1 Design

Both experiments used a within-participant design with three independent variables: camera movement, frame rate and auditory condition. Two camera conditions: *walking* (slow) and *running* (fast), and four different frame rates: 10, 20, 30 and 60 frames per second (fps) in different combinations were used, see Table 7.1. Audio conditions were: *Audio* (footsteps sound effect) and *NoAudio* (silent animation). The dependent variable was the perceived smoothness of the animations. This was measured using the Two Alternative Force Choice (2AFC) method in a complete randomised design. To control for fatigue and familiarity, 10 different animations of the same scene were used. For the *Audio* condition the footsteps sound effects were always synchronised with the visual stimulus.

7.2.2 Participants

In the experiments 86 people volunteered, 71 of whom were university students studying a variety of subjects, and the rest from university staff. The participants' age varied from 17 to 58 with an average age of 26. Out of 86 participants, 61 were male and 25 female. All of them had normal or corrected to normal vision. None of the participants reported any hearing impairments.

7.2.3 Apparatus

The experiment was conducted in a dark, quiet room with no distracters. In the first experiment the visual stimuli were presented on a calibrated 17 inch Philips 170B6 monitor with 1280×1024 pixel resolution and a refresh rate of 60 Hz. In

the second set of experiments, an LG W2234S 22 inch monitor with a refresh rate of 60 Hz and resolution 1680×1050 was used. Although different types of the monitors were used in the experiments, the physical size of the presented stimuli was the same. The stimuli were positioned at eye level, 60-70 cm from the participants eyes. For auditory stimuli an LTB Magnum 5.1 AC97 Headphone set was used.



Figure 7.1: Four frames taken from the walk-through animations. The top two frames are from the animations with camera moving from the corridor to the conference hall, and the bottom two from the animations where the camera is moving from the conference hall to the corridor. See Appendix C for larger images.

7.2.4 Stimuli

The visual stimuli were based only on one scene, see Figure 7.1. To control for familiarity, two animations, at a resolution of 800×600 , were rendered along the same path but in opposite directions, see Figure 7.2. The animations were rendered using our own implementation of path tracing, see Section 3.6. All scenes were static with only frontal camera movement and no rotation relative to the motion path. For each of them a curved motion path with the oscillating motion of the camera along the vertical axis was used, see Figure 7.3. The oscillating motion was used to improve the sensation of walking in the experiment [LBHD06]. The strides in walking and running animations were 0.8m and 1.5m respectively. A young subjects' average normal walking speed of 1.425 m/s [AONI04] was used for the walking condition. For the running condition a speed of 4 m/s was used. All videos were compressed using XviD MPEG-4 Codec (single-pass encoding, target quantizer: 3.00). The animations were divided into three *walking animations* and two *running animations* in both directions, each lasting for five seconds.

For audio, an animation related sound effect was used in both camera conditions (walking and running). This was the sound of footsteps, produced by capturing the sound of leather soled shoes against a firm tiled floor. To synchronise the sound effects with the animation, the length of the silence between the ON signals was varied. The amount of echo in the effect was adjusted according to the nature of the scene and did not change during the animations. Sounds were delivered uncompressed, using two channels (stereo), sample rate 44100Hz and bit rate of 1411kbps. No background music was played in order to avoid any subjective side effects.

For the audio-visual presentation, a framework with support for frame rate

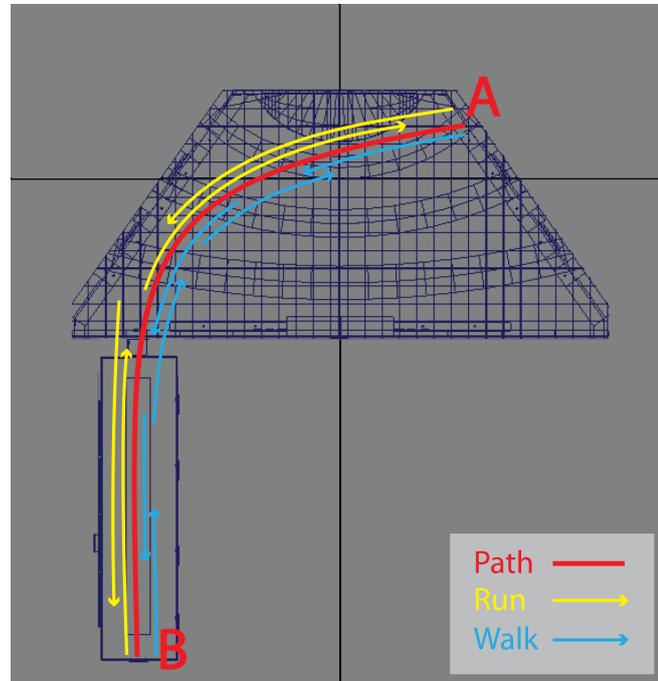


Figure 7.2: Camera path used for the animations (red). Four running animation sequences (yellow) and six walking animation sequences (blue) were used.

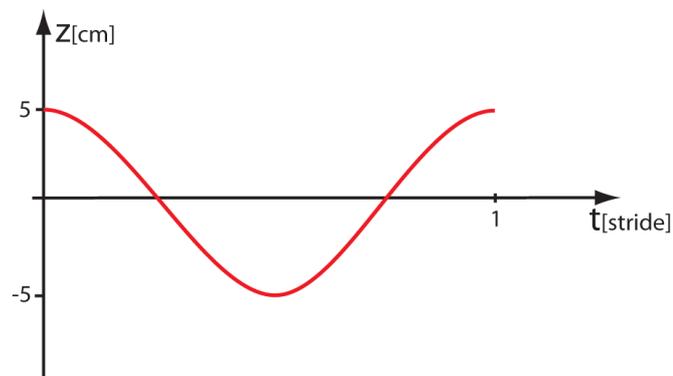


Figure 7.3: Oscillating camera motion along the vertical (z) axis.

and audio control was developed. All results from each trial were saved in separate text files.

7.2.5 Procedure

Prior to the experiment each participant was asked to read and sign a consent form, in which they agreed to voluntarily and anonymously participate in the experiment. They were told that they could withdraw from it at any time. Additionally, all the participants were given a questionnaire to fill in and they were presented with instructions (see Appendix C), followed by two sample animation pairs played at 10 and 60fps. They were told that these were the worst and the best cases respectively, but not what frame rates the animations were. In the instructions they were further explained what frame rate is using Figure 6.3, and that they were going to watch pairs of animations for which they will have to evaluate their smoothness. They were shown 22 and 37 pairs of animations in the first and second experiment respectively.

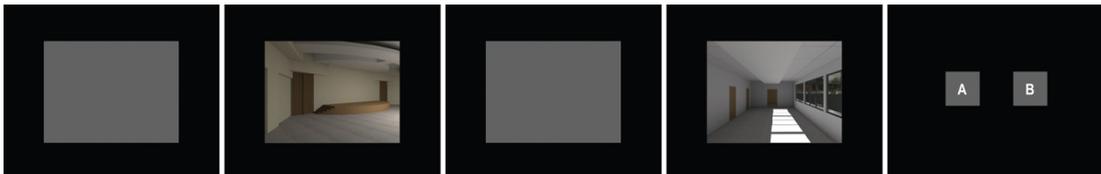


Figure 7.4: The experimental procedure. From left to right: grey box, first animation, grey box, second animation and A/B evaluation screen.

The randomly ordered animation pairs were presented sequentially, see Figure 7.4. Each animation was preceded by a grey box (RGB: 0.3, 0.3, 0.3) lasting for two seconds. The length of each animation was five seconds. After each pair, the A and B boxes were shown on the screen. Participants were instructed to choose the smoother animation by clicking on one of the two boxes, after which the next cycle started automatically. Each trial in the first experiment lasted for about

six minutes and in the second experiment for about 10 minutes. The participants were debriefed on the nature of the study after the experiment.

7.3 Results

In order to test the research hypotheses, the data was analysed using descriptive statistics and non-parametric tests. For both tests Chi-square test was used. Wilcoxon 2 Related Samples test was used for the Test 3.

7.3.1 Test 1: Camera movement speed influence on animation smoothness perception

The first research hypothesis was that the speed of the camera movement will affect the smoothness perception. The null hypothesis was that camera movement speed will have no effect on visual smoothness perception. The hypothesis was tested comparing the walking (Walk) and running (Run) animations, both played at 60fps, which was the gold standard. *NoAudio* (silent) and *Audio* condition were tested separately. The test had a single independent variable - camera movement speed.

Tabular data with observed and expected frequencies, χ^2 and p - values are given in Table 7.2. The Chi-square test found the relationship between Running and Walking animations for *NoAudio* condition as significant ($p = .020$). Therefore, the null hypothesis was rejected in favour of the research hypothesis that the speed of the camera movement affects the animation smoothness perception. The participants preferred the walking rather than the running animation.

The same test for *Audio* condition showed no significance ($p = .182$), and thus the null hypothesis cannot be rejected. The fact that the results were different for the *Audio* and *NoAudio* condition, indicates that audio might affect perception

NoAudio

	Observed N	Expected N	Residual
Run	11	18.0	-7.0
Walk	25	18.0	7.0
Total	36		

$$\chi^2(1) = 5.444, df = 1, p = .020$$
Audio

	Observed N	Expected N	Residual
Run	14	18.0	-4.0
Walk	22	18.0	4.0
Total	36		

$$\chi^2(1) = 1.778, df = 1, p = .182$$

Table 7.2: Test 1: Observed and expected frequencies for the Run - Walk animation smoothness perception comparison.

of animation smoothness. Therefore, the influence of sound effect of footsteps on temporal visual perception was further investigated in Section 7.3.2 and Section 7.3.3.

7.3.2 Test 2: Sound effect's influence on perceived smoothness threshold

In this test, the perceived smoothness threshold was investigated, while watching the animations accompanied by the audio effects. The difference in the preferences between the discrepant frame rate pairs and the control group (60r-60w fps) was compared, see Table 7.1. Lower frame rates were used with the running animation in each test pair. The null hypothesis for each test pair was that animations in discrepant frame rate pairs played at 60fps will not be perceived as smoother. Since no bias was assumed, this means that each test pair will have

	Mean value	$p - value$
10r-60w	1.83	.032
20r-60w	1.52	N/A*
30r-60w	1.63	.5
60r-60w	1.61	N/A

Table 7.3: Test 2: Mean and p values for *Audio* condition. $p - value$ is given for difference in preference between the test pairs and 60r-60w condition. Lower and upper bounds were 1 (first animation preferred) and 2 (second animation preferred) respectively. *Not inline with a 1-tailed test.

the same preference compared to the control group.

For the analysis a one-tailed Chi-square test was used. Therefore, in order to test the validity of the one-tailed hypothesis, corresponding means were compared. The mean values for test pairs 10r-60w, 20r-60w and 30r-60w were 1.83, 1.52 and 1.63 respectively, where lower and upper bounds were 1 and 2, see Table 7.3. The mean value for the control group was 1.61. Since the mean value of 20w-60w condition was lower than the mean value of the control group, the null hypothesis for this pair cannot be rejected. For the 10r-60w and 30r-60w pairs, the difference was in line with our research hypothesis such that we can carry on with the test.

The results show that only for 10r-60w pair there was significant difference in preference ($p = .032$), and thus the null hypothesis can be rejected. For 30r-60w ($p = .5$) pairs there was no significant difference in preference comparing to the 60r-60w control group. Hence, the null hypothesis in this case cannot be rejected. These results appear to show that the perceived smoothness threshold when watching the animations with movement-related sound effects is somewhere between the 10fps and 20fps. This further indicates that walking animations, rendered at 60 frames per second, when accompanied by the movement-related sound effects, were not perceived significantly smoother than the same animation

RUN

	10-20	10-30	10-60	20-30	20-60	30-60
Mean (Audio)*	1.70	1.78	1.76	1.70	1.74	1.52
Mean (NoAudio)*	1.80	1.78	1.72	1.58	1.82	1.52
<i>p</i> - value	.098	N/A**	N/A**	N/A**	.173	N/A**

WALK

	10-20	10-30	10-60	20-30	20-60	30-60
Mean (Audio)*	1.92	1.78	1.78	1.50	1.52	1.32
Mean (NoAudio)*	1.90	1.92	1.88	1.78	1.68	1.40
<i>p</i> - value	N/A**	.010	.029	.002	.044	.197

*Lower bound = 1; Upper bound = 2

**Not inline with a 1-tailed test

Table 7.4: Test 3: Mean and p values for *Audio* condition. *p* - value is given for difference in preference between the test pairs and 60vs60 condition. Lower and upper bounds were 1 (first animation preferred) and 2 (second animation preferred) respectively.

rendered at 20fps and 30fps.

7.3.3 Test 3: Sound effect's influence on animation smoothness perception

The results from the first experiment showed that movement related sound effects could affect the perception of animation smoothness. Therefore, to test for that effect, the second experiment was conducted. The research hypothesis was that movement related sound effects (i.e. footsteps) will increase the perception of smoothness. For the analysis, the data was first divided into two groups: run and walk. Then the Wilcoxon 2 related samples test was performed on each group separately, comparing the animation pairs given in Table 7.1. The animation played at higher frame rate in each pair was always played without sound (*NoAudio* condition), while the sound effect of footsteps was used for the

animation played at lower frame rate. For example, for the pair 10-30 we compared users' preference between the two test pairs: *10/Audio vs 30/NoAudio* and *10/NoAudio vs 30/NoAudio*.

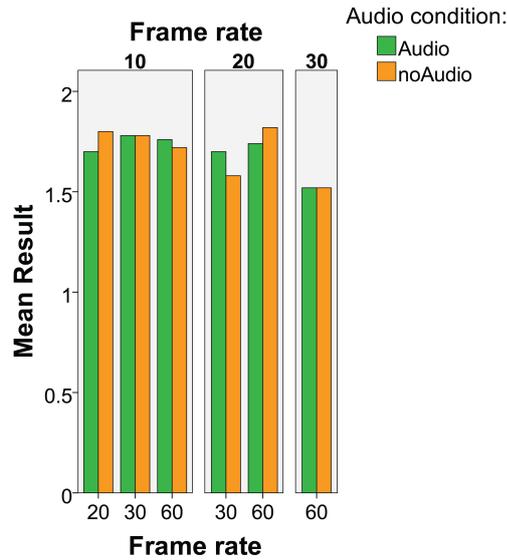


Figure 7.5: Mean values of the compared running animation test pairs.

The results showed that for the fast camera movement (running) there is no significant effect of audio on the perceived visual quality, see Table 7.4. The mean values for this test are shown in Figure 7.5.

Statistical analysis revealed different results for the walking condition. Animations played at 10fps with sound effects have been thought of as smoother than animations played at 30fps ($p = .010$) or 60fps ($p = .029$) with no audio, see Table 7.4. Additionally, the same test showed that animations presented at 20fps with audio were rated as smoother than silent animations played at 30fps ($p = .002$) or 60fps ($p = .044$). For these animations pairs, the null hypothesis can be rejected, which means that movement related sound effects increase the temporal visual perception. Figure 7.6 depicts the mean values for the walking animations across *Audio* and *NoAudio* sound conditions.

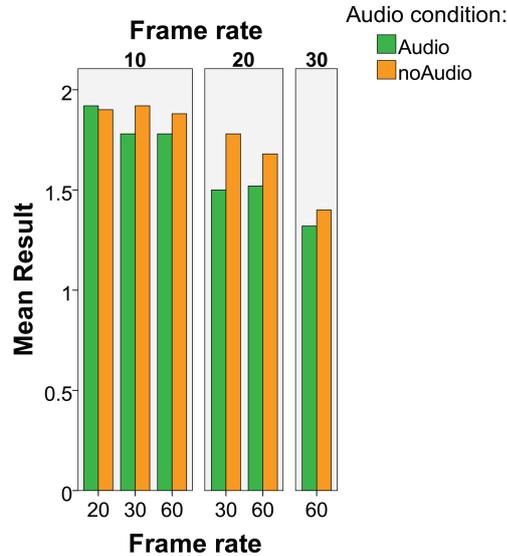


Figure 7.6: Mean values of the compared walking animation test pairs.

7.4 Discussion

In this study, the effect of three factors was investigated: the influence of camera movement speed on perceived smoothness quality of animation; the influence of movement-related audio effects on perceived smoothness threshold of video animations; and movement-related sound effect's influence on animation smoothness perception. The three factors that could influence our results: camera movement speed, frame rate and auditory condition were considered.

The results from the analyses showed that both camera movement speed and movement-related audio effects influence animation smoothness perception. In Section 7.3.1, it has been shown that camera speed significantly affects the perception of animation smoothness when presented with no audio. However, there was no significance of this effect in case of the *Audio* condition. In both audio conditions, the slower (walking) animation was preferred over the faster (running) animation. The different results for the *Audio* and *NoAudio* conditions indicate that audio effects could indeed influence smoothness perception. The results of

Test 2 showed that the perceived smoothness threshold of the pre-rendered animations, when accompanied by a movement-related sound effects, lies between 10 and 20 frames per second. Lastly, in Section 7.3.3 the significance of the auditory influence on perceived smoothness of the animations was investigated. The results showed that movement-related sound effects do increase the animation smoothness perception. The effect was, however, significant only for the slow animations. The reason for this might be in the fact that in running animations, the vertical oscillating motion of the camera introduces a jitter effect, making the frame rate of the animation hard to distinguish.

These results represent another step towards understanding auditory-visual cross-modal interaction and its possible uses in computer graphics. Properly exploited in interactive scenarios, related sounds could be introduced or emphasised to maintain the same perceptual quality when the computational resources are insufficient and required frame rate cannot be delivered.

CHAPTER 8

Conclusions and Future Work

While graphics hardware and rendering algorithms are constantly improving, there is a consistent demand for quality enhancement in auditory-visual rendering, such as geometry complexity, lighting complexity, physics, interaction, etc. Additionally, a complete user experience will require rendering for other modalities (smell, taste, touch). Due to this demand, high-fidelity rendering of complex scenes is still not achievable in real time. Therefore, an alternative solution for achieving this goal is the utilisation of related findings from other disciplines. In this thesis, auditory-visual cross-modal interaction findings, from the field of psychology, have been employed in order to increase or maintain spatial or temporal visual perception when computational resources are insufficient, which results in a speed up for the visual rendering process.

Although multi-modal interaction has been a research topic in psychology for decades and several highly valuable findings have been presented (see Section 4.1), it has been only recently exploited in computer graphics. Different findings on auditory-visual cross-modal interactions in psychology and corresponding studies in computer graphics are presented in Table 8.1.

As can be seen, the main focus of interest in computer graphics so far has been on the perceptual and attentional limitations, such as angular sensitivity,

Phenomenon/Psychology	Used in/Computer Graphics
Angular sensitivity [Jam90, HB89]	[YPG01, CCW03, MDCT05a, Mas06, CDS06, LDC06]
Inattentional blindness [RLGM92, MR98, SC99]	[CCW03, MDCT05a, Mas06]
Modality appropriateness hypothesis [HT66, WW80]	[MC04, Mas06], Chapters 6, 7
Auditory driving effect [GM59, Shi64, WKN03, Rec03]	[MC04, Mas06], Chapters 6, 7
Temporal ventriloquism [MZSFK03, BA03, AB03, BBM09]	Chapters 6, 7
Illusory flash induced by sound [SKS00, SKS02]	Chapters 6, 7
Stimuli weighting [BA06]	[CD09]
Ventriloquism effect [HT66, CWGJ75, VBdG98, VdG04]	[MBT*07]

Table 8.1: The cross-modal phenomena found in psychology and the main studies within the computer graphics that were inspired by this work.

inattentional blindness or modality appropriateness hypothesis. However, the results presented in Chapter 5 indicate that significant potential lies in investigating the perceptual and attentional limitations in auditory-visual cross-modal interaction. In this study, exposure to broadband noise as the auditory stimuli decreased the perceived rendering quality threshold. Furthermore, the work presented in Chapters 6 and 7 showed that audio has a significant influence on frame rate perception. In particular, the beat rate can have strong effect, increasing the perceived frame rate of animations when observing walk-through animations with no objects moving. In addition, movement related sound effect significantly increase perception of animation smoothness when observing slow (walking) animations.

The main aim of this thesis was to further investigate the direct relationship between vision and audition in both spatial and temporal domain which should give a better understanding of this interaction, making it possible to effectively

reduce the computational time of rendering, without the user noticing any perceptual loss in quality. This was achieved by introducing auditory stimuli as an addition to the visuals. The behaviour and effectiveness of the auditory-visual cross-modal interaction experienced in such scenarios were investigated by conducting multiple psychophysical experiments, looking at auditory influence on spatial and temporal visual perception.

The investigation of auditory influence on the perceived rendering threshold (Chapter 5) showed that the cross-modal interaction between the two modalities indeed exists. This was first confirmed by analysing the data via loglinear analysis, which confirmed the main effects of Sound, Quality and Score ($\chi^2=55.14$; $p < .001$), along with the two-way interactions between these factors ($\chi^2=94.52$; $p < .001$). To interpret these interactions, Chi-square analysis was used across sound conditions and scenes. This analysis compared the user preference between the images, rendered at different quality, presented with related sound, unrelated sound or without sound. The results revealed that the related and unrelated sounds had an opposite effect on visual perception: the unrelated sounds reduced the perceived rendering quality threshold, while the threshold was increased by the related sounds.

For two out of four scenes (Corridor and Kalabsha) that were used in the study, participants could not notice any quality difference between the images rendered at 1rpp or 49rpp when presented with unrelated sounds ($p = .081$). At the same time, the comparison for *no sound* and *related sound* groups showed that this difference was significant (all $p - values < 0.05$). However, in the experiments, noise sounds (white, brown and pink noise) were used, which might have had a distracting effect. Therefore, this might be useful only in situations when we intentionally want to distract a user by some external, unexpected event, while, at the same time, decreasing the spatial visual rendering quality.

When using related sounds, the perceived rendering threshold was found to be either the same or higher than in *no sound* condition. For the Corridor and Kalabsha scenes no significant effect of related sound was found, while for Checkerboard and Library scenes the threshold was higher - at 9rpp and 4rpp respectively. This was in conflict with the research hypothesis, stating that by introducing sound, it would be possible to deliver lower quality images without any perceivable difference in visual appearance. This might be partially attributed to attentional attraction towards certain regions of the observed image by the presented sound. Therefore, it would be necessary to investigate this further using eye-tracking techniques, which should reveal if a user focuses on certain parts of the presented image when listening to accompanying audio. This way, users' gaze could be analysed, and only those regions attended to, need to be selectively rendered in higher quality.

In the following study, presented in Chapter 6, the main effects of beat rate, frame rate and scene on the perception of animation smoothness were investigated. Additionally, the effect within and between these factors was studied. The investigation was done by conducting two sets of experiments with two different participant groups. This resulted in having two separate within participant statistical analyses: for static scenes, where no objects in the virtual scene were moving, except for the virtual camera which changes in position and/or orientation; and for dynamic scenes, in which one or more objects were moving while the camera was fixed.

The results of the analysis for static scenes showed a significant effect of each factor (frame rate, beat rate and scene). Although a different rating of the animations played at different frame rates might be considered as obvious, it was important to make sure that the chosen frame rates were adequate, i.e. distinguishable. The motivation partially lied in the Wertheimer theory, in which

he claims that the HVS tends to blend successive images shown with a short delay between presentations (between 1/10 and 1/40 of a second) - a phenomenon known as *the apparent motion* [BS06]. Another reason for choosing these frame rates is the fact that computer games usually run at 60fps and more, and that standard computer display refresh rate is 60Hz. All this indicates that frame rates between 10fps and 60fps are the ones to be considered. Therefore, the significance of the main effect of frame rate ($F=478.54$; $p < .001$), found by analysing the data, means that ignoring the effect of beat rate and scene, participants were able to distinguish all frame rates used in the experiments. Having this in mind, the interaction between the beat rate and frame rate effects was investigated.

Since beat rate was found as significant factor ($F=8.23$; $p = .001$), a further analysis was performed, looking at the pairs of beat rates and the interaction between them. Looking back at Section 6.3.1 it is clear that there is a significant difference in rating animation smoothness when using a slow beat rate (BR2) and a fast beat rate (BR6). Interestingly, the audio with slow beat rate tended to increase the perceived animation smoothness, while BR6 had the opposite effect. The effect of low beat rates might be a consequence of attentional limitations, i.e. the inter-modal attentional model [SJ01]. This model proposes that our attention operates on a global level and is not divided across multiple senses. This means that the performance of a task requiring attention for one modality will be affected by concurrent tasks in some other modality. On the other hand, the negative effect of high beat rate on smoothness perception might be a result of the auditory driving effect [GM59, Shi64] or temporal ventriloquism [MZSFK03, BA03, AB03], or the combination of the two. These phenomena advocate that the auditory stimuli can drive the temporal perception of the visual stimuli.

For the third independent variable, the scene factor, a significant main effect was found ($F=52.087$; $p < .001$). The results indicate a different perception of

animation smoothness for different scenes. As explained in Section 6.2 this factor has many dimensions, such as virtual scene complexity (geometry, materials and lighting) and virtual camera behaviour (movement type and speed). Therefore, although these findings are important, showing that smoothness perception highly depends on the scene being observed, a further investigation that would look at all these mentioned parameters separately would be useful to further understand the complexities of these variables. A segment of this investigation is presented in Chapter 7.

For the dynamic scenes, no significance of the beat rate effect was found ($F=1.592$; $p = .208$). This again confirms the importance of the scene factor, as discussed in the text above. The other two factors: frame rate ($F=51.497$; $p < .001$) and scene ($F=4.628$; $p = .036$) were found to be significant.

Since the study presented in Chapter 6 indicated that the camera movement might affect the perceived frame rate of the animations, this was further investigated in Chapter 7. Namely, the influence of a specific aspect of the camera movement - its speed, on frame rate perception was investigated. It has been shown that slow camera movement, i.e. walking animations, is perceived as smoother than the fast camera movement (running animations) when presented at 60fps with no audio ($p = 0.20$). The same comparison, using animations accompanied with footsteps sound effects, revealed no significance ($p = .182$). The former might be due to the smaller distance interval Δs of the slow camera movement, resulting in a lesser image shift between successive frames. Another possible cause for this user preference might be the fast vertical oscillating motion of the camera that could introduce a jitter effect, making the frame rate of the animation hard to distinguish. The latter indicated that there might be a direct influence of footsteps sound effect on animation smoothness perception.

Furthermore, the same study (see Chapter 7) investigated the influence of the

movement-related sound effect on perceived smoothness threshold. The results showed that this threshold lies between 10fps and 20fps, meaning that the viewers could not perceive significant difference in quality between 20fps and 60fps when accompanied by the movement-related sound effects. Therefore, rendering animations at 20fps instead of 60fps, while maintaining the same perceptual quality, could lead to a computational speedup of three times.

Finally, the direct effect of the footsteps sound on animation smoothness perception was investigated. The results show a significant affect for the walking animations. More specifically, animations played at 10fps and 20fps with sound effects were perceived as smoother than animations played at 30fps or 60fps with no audio. The cross-modal effect found in the second and third tests indicate that it would be possible to achieve a significant speedup in rendering (up to six times) when using movement-related sound effects.

This knowledge could be used to control a rendering system, saving computation while maintaining the same perceptual experience within a multi-modal environment, for example in interactive systems when scene complexity increases rapidly such that the frame rate is affected. In such a system the beat rate could be manipulated interactively, somewhat akin to time-constrained level-of-detail manipulation [FS93], such that the drop in frame rate is not perceivable by the viewer. An alternative scenario where the beat rate and frame rate relationship is useful is when computing high-quality animations which are highly computationally expensive and typically require huge resources such as render farms. Under such conditions, if the animation score is known beforehand, the minimal required frame rate could be calculated for each audio segment, significantly reducing the total rendering time without compromising the visual experience.

8.1 Contributions

In this thesis, a comprehensive overview of the auditory-visual cross-modal interaction has been presented, including the main findings in psychology and a detailed survey of the corresponding work carried out in the field of computer graphics (Chapter 4). The novel work, presented in Chapters 5, 6 and 7 build on these findings. Through multiple user studies, this thesis investigates some particular features of audition and vision that could affect human perception, and thus benefit visual rendering performance.

There are several major outcomes of the thesis which confirm that:

- a relationship between the vision and audition in both the spatial and temporal domain exists (Chapters 5, 6 and 7).
- the perceived rendering quality threshold can be reduced by sound. When accompanied by an unrelated sound, spatial quality of the presented image can be decreased without any perceivable difference (Chapter 5) The effect is dependant on the scene properties.
- audio beat rate can significantly affect video frame rate perception. Low beat rate significantly increases perceived frame rate (6).
- movement related sound effects influence temporal visual perception.
- camera movement speed affects temporal visual perception.
- the temporal visual perception is affected by the scene content and complexity.
- familiarity with computer games and/or animation has no influence on visual perception.

Additionally, so far there has been no established statistical framework for video animation quality comparison using numerical scores assigned from a slider bar for both dependent and independent sample design. Therefore, the methodology described in Chapter 6 can be used as a framework for similar data analysis in such studies.

The contributions of the thesis provide an initial step for the future development of the rendering framework, that could enable real-time high-fidelity rendering at high frame rates.

8.2 Impact

The novel findings from this thesis could be used for various multimedia application, where both rendering quality and time constraints need to be satisfied. In many virtual reality scenarios, such as entertainment, simulations and training, virtual reconstructions and museums, etc. audio is used as a complementary stimulus, to achieve a richer experience and a higher level of immersion and presence in a virtual space. As shown in this thesis, the addition of audio does not have to lead to an increase of computation, but it can be rather used for maintaining the same perceptual quality while decreasing the spatial and/or temporal frequency of graphics.

Producing realistic virtual environments requires high-quality stimulation of multiple senses, aiming at the full immersion of a viewer as if he or she was present in the real scene being depicted [CDMdS07]. Although sound utilisation in virtual environments is not novel, this thesis shows that knowing the cross-modal effects between vision and audition can help in developing better multi-modal environments and simulations.

Since this is one of the first studies in computer graphics to investigate this

phenomenon, more work is needed in order to create a framework that could use the full potential of the auditory-visual cross-modal interaction. This framework could adjust the audio and visual quality on-demand, reducing or balancing the work load whenever required.

Finally, this thesis may be used as a guideline for a future cross-modal research investigating other modalities, such as smell, taste and touch.

8.3 Directions for future work

Using findings on multi-modal interaction from psychology and extrapolating them for the computer graphics scenarios requires significant changes in test stimuli. These changes lead to a substantial increase of the complexity level and number of variables that need to be carefully controlled. Therefore, individual variables need to be isolated and observed separately.

Although various psychological phenomena are directly mapped and some of them extrapolated into computer graphics applications (see Table 8.1), there are still some to be investigated, and potentially utilised in computer graphics, such as:

- modality confusion [LMBB03]
- McGurk effect [MM76]
- audio effect on visual intensity [SLWP96]
- audio effect on visual search [VdBOBT08]
- bouncing targets / circles [SSL97]

The results presented in this thesis could be extended by focusing on the conditions that indicated significant cross-modal influence. This might include

higher granularity of rays-per-pixel (rpp), beat rates and frame rates. Furthermore, direct comparison of the audio-visual content at lower frame rates with ones presented at higher frame rates could be investigated. Additionally, a deeper investigation into scene complexity, different audio types and sounds, camera movement is required.

Although the study on spatial visual perception indicates that unrelated sounds decrease perceived rendering quality threshold and that related sounds may require an increase in rendering quality when compared to a no-sound scenario, a further investigation of the manner in which this affects user's attention is needed. As with any such statistical study, the validity of the result is dependent upon the sample size. Therefore, increasing the number of participants would improve the level of significance. Additionally, the results might be further improved by using selective rendering techniques [Deb06, Mas06].

Temporal visual perception could be investigated and potentially enhanced by using music as the auditory stimulus in multi-modal environments. Music, however, carries the effect of emotions, which has rather subjective impact on a user, and is hard to control [MC04]. Nevertheless, this would benefit not only virtual worlds and video games, but also the film and advertising industries.

Another segment of the cross-modality that has not been tackled in this thesis is the spatial audio. It has been recently shown that using stimuli that is precisely defined and positioned in virtual space, it is possible to attract users' attention to one side of the observed screen, or even to a specific object in the screen [MDCT05a, HWBR*10]. Using selective rendering with 3D audio could lead to a significant rendering speed-up.

In addition, the impact of auditory influence on visual perception has yet to be applied to interactive scenarios. Although this brings another set of variables into consideration, making the environment harder to control, the results could

be directly mapped to 3D virtual scenarios. Such results also entail the possibility of building decision-theoretic systems [DPF03] based on cross-modal effects that ensure a constant perceived frame rate rather than the commonly used fixed frame rate [FS93].

Finally, other senses, such as smell, taste and touch, need to be included in such research. Adding those senses could heighten the immersion in virtual worlds and enhance the overall user experience.

8.4 Final remarks

While high-fidelity rendering of simple scenes might be performed interactively at low resolutions, markets and technology are imposing increased demands, requiring higher frame rates, image resolutions and interactivity level, more complex virtual scenes and containing additional modality stimulations. This means that, as the hardware and graphics algorithms evolve, consumer expectations grow simultaneously. Therefore, additional techniques for enhancing the rendering process need to be developed.

The cross-modal interaction in computer graphics has been investigated for less than a decade, and therefore, there is a substantial amount of work still to be done. Although this is a long and complex process, the findings presented in this report promise a bright future for the field. This thesis represents another step towards a detailed understanding of auditory-visual cross-modal interaction and its possible uses in computer graphics.

References

- [AA93] ANDERSON J., ANDERSON B.: The Myth of Persistence of Vision Revisited. *Journal of Film and Video* 45, 1 (1993), pp. 3–12.
- [AAR72] ALLPORT D. A., ANTONIS B., REYNOLDS P.: On the division of attention: a disproof of the single channel hypothesis. *Q J Exp Psychol* 24, 2 (May 1972), pp. 225–235.
- [AB03] ASCHERSLEBEN G., BERTELSON P.: Temporal ventriloquism: crossmodal interaction on the time dimension: 2. evidence from sensorimotor synchronization. *International Journal of Psychophysiology* 50, 1-2 (2003), pp. 157 – 163.
- [AB04] ALAIS D., BURR D.: The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology* 14, 3 (February 2004), pp. 257–262.
- [ADCH06] ARANHA M., DEBATTISTA K., CHALMERS A., HILL S.: Perceived Rendering Thresholds for High-Fidelity Graphics on Small Screen Devices. In *Theory and Practice of Computer Graphics 2006* (July 2006), EG, pp. 133–140.

- [Alt04] ALTEN S. R.: *Audio in Media*, 7th ed. Wadsworth Publishing, 2004.
- [AMB06] ALAIS D., MORRONE C., BURR D.: Separate attentional resources for vision and audition. *Proc Biol Sci* 273, 1592 (Jun 2006), pp. 1339–1345.
- [AMHH08] AKENINE-MÖLLER T., HAINES E., HOFFMAN N.: *Real-Time Rendering 3rd Edition*. A. K. Peters, Ltd., 2008.
- [AONI04] ARIF M., OHTAKI Y., NAGATOMI R., INOOKA H.: Estimation of the effect of cadence on gait stability in young and elderly people using approximate entropy technique. *Measurement Science Review* 4 (2004).
- [App68] APPEL A.: Some techniques for shading machine renderings of solids. In *AFIPS '68: Proceedings of the April 30–May 2, 1968, Spring Joint Computer Conference* (1968), ACM, pp. 37–45.
- [AWWDJ04] ALLMAN-WARD M., WILLIMAS R., DUNNE G., JENNINGS P.: The evaluation of vehicle sound quality using an nvh simulator. In *Proceedings of the 33rd International Congress and Exposition on Noise Control Engineering 2004* (2004).
- [BA03] BERTELSON P., ASCHERSLEBEN G.: Temporal ventriloquism: crossmodal interaction on the time dimension: 1. evidence from auditory-visual temporal order judgment. *International Journal of Psychophysiology* 50, 1-2 (2003), pp. 147 – 155.
- [BA06] BURR D., ALAIS D.: Combining visual and auditory information. *Prog Brain Res* 155 (2006), pp. 243–258.

- [BBM09] BURR D., BANKS M., MORRONE M.: Auditory dominance over vision in the perception of interval duration. *Experimental Brain Research* 198 (2009), pp. 49–57.
- [BCFW08] BARTZ D., CUNNINGHAM D., FISCHER J., WALLRAVEN C.: The role of perception for computer graphics. In *Eurographics State-of-the-Art-Reports* (2008), pp. 65–86.
- [BCN07] BARGARY G., CHAN J., NEWELL F.: Seeing where the ears hear: visual encoding of auditory spatial sequences. In *8th Annual Meeting of the International Multisensory Research Forum* (2007).
- [BH98] BONNEL A. M., HAFTER E. R.: Divided attention between simultaneous auditory and visual signals. *Percept Psychophys* 60, 2 (February 1998), pp. 179–190.
- [BHWL99] BASTOS R., HOFF K., WYNN W., LASTRA A.: Increased photorealism for interactive architectural walkthroughs. In *Proceedings of the 1999 symposium on Interactive 3D graphics - SI3D '99* (1999), ACM Press, pp. 183–190.
- [Bri11a] BRITANNICA E.: ear: structure of the human ear. art. encyclopedia britannica online. <http://www.britannica.com/EBchecked/media/530/Structure-of-the-human-ear>, July 2011.
- [Bri11b] BRITANNICA E.: eye, human: horizontal cross section of the human eye. art. encyclopedia britannica online. <http://www.britannica.com/EBchecked/media/100415/A-horizontal-cross-section-of-the-human-eye-showing-the>, July 2011.

- [Bri11c] BRITANNICA E.: optic chiasm: visual pathways. art. encyclopedia britannica online. <http://www.britannica.com/EBchecked/media/53283/Visual-pathways>, July 2011.
- [BS06] BLAKE R., SEKULER R.: *Perception*, 5th ed. McGraw-Hill Higher Education, 2006.
- [BSVDD10] BONNEEL N., SUIED C., VIAUD-DELMON I., DRETTAKIS G.: Bimodal perception of audio-visual material properties for virtual environments. *ACM Transactions on Applied Perception* 7, 1 (Jan. 2010), pp. 1–16.
- [CCW03] CATER K., CHALMERS A., WARD G.: Detail to attention: exploiting visual tasks for selective rendering. In *EGRW '03: Proceedings of the 14th Eurographics Workshop on Rendering Techniques* (2003), Eurographics Association, pp. 270–280.
- [CD09] CHALMERS A., DEBATTISTA K.: Level of realism for serious games. *Games and Virtual Worlds for Serious Applications, Conference in 0* (2009), pp. 225–232.
- [CDMdS07] CHALMERS A., DEBATTISTA K., MASTOROPOULOU G., DOS SANTOS L.: There-reality: Selective rendering in high fidelity virtual environments. *International Journal of Virtual Reality* 6, 1 (2007).
- [CDS06] CHALMERS A., DEBATTISTA K., SANTOS L. P.: Selective rendering: computing only what you see. In *GRAPHITE '06: Proceedings of the 4th international conference on Computer graph-*

- ics and interactive techniques in Australasia and Southeast Asia* (2006), ACM Press, pp. 9–18.
- [Cha] CHAOS GROUP: V-ray render. <http://www.chaosgroup.com/en/2/vray.html>.
- [CPC84] COOK R. L., PORTER T., CARPENTER L.: Distributed ray tracing. *ACM SIGGRAPH Computer Graphics* 18, 3 (July 1984), pp. 137–145.
- [CRO] CROSSMOD: CROSSMOD project, cross-modal perceptual interaction and rendering.
- [Cry] CRYTEK: Cry engine 3. <http://mycryengine.com/>.
- [CWGJ75] CHOE C. S., WELCH R. B., GILFORD R. M., JUOLA J. F.: The "ventriloquist effect": Visual dominance or response bias? *Perception and Psychophysics* 18, 1 (1975), pp. 55–60.
- [Dal93] DALY S.: *The visible differences predictor: an algorithm for the assessment of image fidelity*. MIT Press, 1993, pp. 179–206.
- [DBBS06] DUTRE P., BALA K., BEKAERT P., SHIRLEY P.: *Advanced Global Illumination*. AK Peters Ltd, 2006.
- [Deb06] DEBATTISTA K.: *Selective Rendering for High Fidelity Graphics*. Phd in computer science, 2006.
- [DLFPT09] DE LUCIA A., FRANCESE R., PASSERO I., TORTORA G.: Development and evaluation of a virtual campus on second life: The case of seconddmi. *Comput. Educ.* 52 (January 2009), pp. 220–233.

- [DMW97] DUNCAN J., MARTENS S., WARD R.: Restricted attentional capacity within but not between sensory modalities. *Nature* 387, 6635 (June 1997), pp. 808–810.
- [DPF03] DUMONT R., PELLACINI F., FERWERDA J. A.: Perceptually-driven decision theory for interactive realistic rendering. *ACM Trans. Graph.* 22, 2 (2003), pp. 152–181.
- [DS94] DRIVER J., SPENCE C.: *Attention and Performance XV*. MIT Press, 1994, pp. 311–331.
- [DS98] DRIVER J., SPENCE C.: Crossmodal attention. *Curr Opin Neurobiol* 8, 2 (April 1998), pp. 245–253.
- [EA] EA DIGITAL ILLUSIONS CE: Frostbite engine.
<http://www.dice.se/>.
- [Epi] EPIC GAMES: Unreal engine.
<http://www.unrealtechnology.com/>.
- [Fie09] FIELD A.: *Discovering Statistics Using SPSS (Introducing Statistical Methods)*, 3rd ed. Sage Publications Ltd, 2009.
- [FN05] FUJISAKI W., NISHIDA S.: Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *Exp Brain Res* 166, 3-4 (Oct 2005), pp. 455–464.
- [Fre] FREE SOFTWARE FOUNDATION: Lux render.
<http://www.luxrender.net>.
- [FS93] FUNKHOUSER T., SÉQUIN C.: Adaptive display algorithm for interactive frame rates during visualization of complex virtual envi-

- ronments. In *SIGGRAPH '93: Proceedings of the 20th annual conference on Computer graphics and interactive techniques* (1993), ACM, pp. 247–254.
- [GBW*09] GRELAUD D., BONNEEL N., WIMMER M., ASSELOT M., DRETTAKIS G.: Efficient and practical audio-visual rendering for games using crossmodal perception. In *I3D '09: Proceedings of the 2009 symposium on Interactive 3D graphics and games* (2009), ACM, pp. 177–182.
- [Get07] GETZMANN S.: The effect of brief auditory stimuli on visual apparent motion. *Perception* 36, 7 (2007), pp. 1089–1103.
- [GGB05] GUTTMAN S. E., GILROY L. A., BLAKE R.: Hearing what the eyes see: Auditory encoding of visual temporal sequences. *Psychological Science* 16, 3 (March 2005), pp. 228–235.
- [GJF*06] GIUDICE S., JENNINGS P., FRY J., DUNNE G., WILLIAMS R., ALLMAN-WARD M.: The evaluation of vehicle sound quality using an nvh simulator. In *INTER-NOISE 2006* (2006).
- [GK04] GERO J. S., KAZAKOV V.: On measuring the visual complexity of 3d objects. *Journal of Design Sciences and Technology* 12, 1 (2004), pp. 35–44.
- [GM59] GEBHARD J. W., MOWBRAY G. H.: On discriminating the rate of visual flicker and auditory flutter. *Am J Psychol* 72 (Dec 1959), pp. 521–529.
- [HAC08] HULUSIC V., ARANHA M., CHALMERS A.: The influence of cross-modal interaction on perceived rendering quality thresholds.

- In *WSCG 2008 Full Papers Proceedings* (2008), Skala V., (Ed.), pp. 41–48.
- [HB89] HUMPHREYS G. W., BRUCE V.: *Visual Cognition: Computational, Experimental and Neuropsychological Perspectives*. Lawrence Erlbaum Associates Ltd, 1989.
- [HCD*09] HULUSIC V., CZANNER G., DEBATTISTA K., SIKUDOVA E., DUBLA P., CHALMERS A.: Investigation of the beat rate effect on frame rate for animated content. In *Spring Conference on Computer Graphics 2009* (2009), Hauser H., (Ed.), Comenius University, Bratislava, pp. 167–174.
- [HDAC10] HULUSIC V., DEBATTISTA K., AGGARWAL V., CHALMERS A.: Exploiting audio-visual cross-modal interaction to reduce computational requirements in interactive environments. In *Proceedings of the IEEE conference on Games and Virtual Worlds for Serious Applications* (2010), IEEE Computer Society.
- [HDAC11] HULUSI V., DEBATTISTA K., AGGARWAL V., CHALMERS A.: Maintaining frame rate perception in interactive environments by exploiting audio-visual cross-modal interaction. *The Visual Computer* 27 (2011), pp. 57–66.
- [HHT*11] HULUSIC V., HARVEY C., TSINGOS N., DEBATTISTA K., STEVE W., DAVID H., ALAN C.: Acoustic rendering and auditory-visual cross-modal perception and interaction. In *Eurographics State-of-the-Art-Reports* (2011).
- [HL97] HORVITZ E., LENGYEL J.: Perception, attention, and resources: A decision-theoretic approach to graphics rendering. In *1997, Pro-*

- ceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI-97 (1997)*, pp. 238–249.
- [HMD*10] HAPPA J., MUDGE M., DEBATTISTA K., ARTUSI A., GONALVES A., CHALMERS A.: Illuminating the past: state of the art. *Virtual Reality 14* (2010), pp. 155–182.
- [HT66] HOWARD I. P., TEMPLETON W. B.: *Human spatial orientation [by] I.P. Howard and W.B. Templeton*. Wiley, London, New York,, 1966.
- [HWBR*10] HARVEY C., WALKER S., BASHFORD-ROGERS T., DEBATTISTA K., CHALMERS A.: The Effect of Discretised and Fully Converged Spatialised Sound on Directional Attention and Distraction. In *Theory and Practice of Computer Graphics* (2010), Collomosse J., Grimstead I., (Eds.), Eurographics Association.
- [IK00] ITTI L., KOCH C.: Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention. *Vision Research 40* (2000), pp. 1489–1506.
- [IK01] ITTI L., KOCH C.: Computational modelling of visual attention. *Nat Rev Neurosci 2*, 3 (March 2001), pp. 194–203.
- [IKN98] ITTI L., KOCH C., NIEBUR E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. Pattern Anal. Mach. Intell. 20*, 11 (1998), pp. 1254–1259.
- [ITU] INTERNATIONAL-TELECOMMUNICATION-UNION: Methodology for the subjective assessment of the quality of television pictures. Recommendation ITU-R BT.500-11.

- [Jam90] JAMES W.: *The principles of psychology*. Holt, 1890.
- [Jam92] JAMES W.: *Psychology*. H. Holt and Co, 1892.
- [Jen01] JENSEN H. W.: *Realistic Image Synthesis Using Photon Mapping*. AK Peters, 2001.
- [Kai11] KAISER P.: The Joy of Visual Perception (Web book). <http://www.yorku.ca/eye/thejoy.htm>, April 2011.
- [Kaj86] KAJIYA J. T.: The rendering equation. In *SIGGRAPH '86: Proceedings of the 13th annual conference on Computer graphics and interactive techniques* (1986), ACM, pp. 143–150.
- [KBM*07] KORDING K. P., BEIERHOLM U., MA W. J., QUARTZ S., TENENBAUM J. B., SHAMS L.: Causal inference in multisensory perception. *PLoS ONE* 2, 9 (09 2007).
- [KK02] KOLLIG T., KELLER A.: Efficient multidimensional sampling. *Computer Graphics Forum* 21, 3 (2002), pp. 557–563.
- [KPLL05] KAYSER C., PETKOV C. I., LIPPERT M., LOGOTHETIS N. K.: Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology* 15, 21 (2005), pp. 1943–1947.
- [KT02] KELLY M. C., TEW A. I.: The continuity illusion in virtual auditory space. In *Proceedings of AES 112th Convention* (May 2002).
- [KU85] KOCH, C., ULLMAN, S.: Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology* 4, 4 (1985), pp. 219–227.

- [LBHD06] LECUYER A., BURKHARDT J.-M., HENAFF J.-M., DONIKIAN S.: Camera motions improve the sensation of walking in virtual environments. In *VR '06: Proceedings of the IEEE conference on Virtual Reality* (2006), IEEE Computer Society, pp. 11–18.
- [LDC06] LONGHURST P., DEBATTISTA K., CHALMERS A.: A gpu based saliency map for high-fidelity selective rendering. In *AFRIGRAPH '06: Proceedings of the 4th international conference on Computer graphics, virtual reality, visualisation and interaction in Africa* (2006), ACM, pp. 21–29.
- [LM00] LOSCHKY L. C., MCCONKIE G. W.: User performance with gaze contingent multiresolutional displays. In *ETRA '00: Proceedings of the 2000 symposium on Eye tracking research & applications* (2000), ACM, pp. 97–103.
- [LMBB03] LARSEN A., MCILHAGGA W., BAERT J., BUNDESEN C.: Seeing or hearing? perceptual independence, modality confusions, and crossmodal congruity effects with focused and divided attention. *Percept Psychophys* 65, 4 (May 2003), pp. 568–574.
- [Mas06] MASTOROPOULOU G.: *The Effect of Audio on the Visual Perception of High-Fidelity Animated 3D Computer Graphics*. Phd in computer science, 2006.
- [MBT*07] MOECK T., BONNEEL N., TSINGOS N., DRETTAKIS G., VIAUD-DELMON I., ALLOZA D.: Progressive perceptual audio rendering of complex scenes. In *I3D '07: Proceedings of the 2007 symposium on Interactive 3D graphics and games* (2007), ACM, pp. 189–196.

- [MC04] MASTOROPOULOU G., CHALMERS A.: The effect of music on the perception of display rate and duration of animated sequences: An experimental study. In *TPCG '04: Proceedings of the Theory and Practice of Computer Graphics 2004 (TPCG'04)* (2004), IEEE Computer Society, pp. 128–134.
- [MDCT05a] MASTOROPOULOU G., DEBATTISTA K., CHALMERS A., TROSCIANKO T.: Auditory bias of visual attention for perceptually-guided selective rendering of animations. In *GRAPHITE '05: Proceedings of the 3rd international conference on Computer graphics and interactive techniques in Australasia and South East Asia* (2005), ACM Press, pp. 363–369.
- [MDCT05b] MASTOROPOULOU G., DEBATTISTA K., CHALMERS A., TROSCIANKO T.: The influence of sound effects on the perceived smoothness of rendered animations. In *APGV '05: Proceedings of the 2nd symposium on Applied perception in graphics and visualization* (2005), ACM Press, pp. 9–15.
- [Men] MENTALIMAGES: Mental ray rendering software. <http://www.mentalimages.com/products/mental-ray.html>.
- [MGKP08] MARAGOS P., GROS P., KATSAMANIS A., PAPANDREOU G.: *Cross-Modal Integration for Performance Improving in Multimedia: A Review*. Springer-Verlag, 2008.
- [Mit07] MITTRING M.: Finding next gen: CryEngine 2. pp. 97–121.
- [MM76] MCGURK H., MACDONALD J.: Hearing lips and seeing voices. *Nature* 264, 5588 (December 1976), pp. 746–748.

- [MMC09] MACKNIK S., MARTINEZ-CONDE S.: *Encyclopedia of Perception*. SAGE Press, 2009, ch. Vision: Temporal factors, pp. 1060–1062.
- [MMS04] MANTIUK R., MYSZKOWSKI K., SEIDEL H.-P.: Visible difference predictor for high dynamic range images. In *2004 IEEE International Conference on Systems, Man & Cybernetics* (October 2004), Thissen W., Wieringa P., Pantic M., Ludema M., (Eds.), vol. 3, IEEE, pp. 2763–2769.
- [Moo82] MOORE B. C.: *An Introduction to the Psychology of Hearing*, 2nd ed. Academic Press, 1982.
- [MR98] MACK A., ROCK I.: *Inattentional Blindness*. The MIT Press, 1998.
- [MTAS01] MYSZKOWSKI K., TAWARA T., AKAMINE H., SEIDEL H.-P.: Perception-Guided Global Illumination Solution for Animation Rendering. In *{SIGGRAPH} 2001, Computer Graphics Proceedings* (2001), Fiume E., (Ed.), ACM Press / ACM SIGGRAPH, pp. 221–230.
- [MW77] MASSARO D. W., WARNER D. S.: Dividing attention between auditory and visual perception. *Perception & Psychophysics* 21, 6 (1977), pp. 569–574.
- [Mys02] MYSZKOWSKI K.: Perception-based global illumination, rendering, and animation techniques. In *{SIGGRAPH} 2002, Proceedings of the 18th spring conference on Computer graphics* (2002), ACM Press / ACM SIGGRAPH, pp. 13–24.

- [MZSFK03] MOREIN-ZAMIR S., SOTO-FARACO S., KINGSTONE A.: Auditory capture of vision: examining temporal ventriloquism. *Brain Res Cogn Brain Res* 17, 1 (Jun 2003), pp. 154–163.
- [Nex] NEXT LIMIT TECHNOLOGIES: Maxwell render. <http://www.maxwellrender.com/>.
- [OHM*04] O’SULLIVAN C., HOWLETT S., McDONNELL R., MORVAN Y., O’CONOR K.: Perceptually adaptive graphics. In *Eurographics State-of-the-Art-Reports* (2004).
- [Pas99] PASHLER H.: *The psychology of attention*. The MIT Press, 1999.
- [PCK07] PARKE R., CHEW E., KYRIAKAKIS C.: Quantitative and visual analysis of the impact of music on perceived emotion of film. *Comput. Entertain.* 5, 3 (2007).
- [PH10] PHARR M., HUMPHREYS G.: *Physically Based Rendering: From Theory To Implementation*, second ed. Morgan Kaufmann, 2010.
- [PISI00] PAINTER T., IEEE S. M., SPANIAS A., IEEE S. M.: Perceptual coding of digital audio. In *Proceedings of the IEEE* (2000), pp. 451–515.
- [PNB03] POST F. H., NIELSON G. M., BONNEAU G.-P. (Eds.): *Data Visualization: The State of the Art* (2003), Kluwer.
- [Pos80] POSNER M. I.: Orienting of attention. *The Quarterly Journal of Experimental Psychology* 32, 1 (1980), pp. 3–25.

- [PP59] PETERSON L. R., PETERSON M. J.: Short-term memory retention of individual items. *Journal of Experimental Psychology* 58 (1959), pp. 193–198.
- [PS90] FERROTT D. R., SABERI K.: Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the Acoustical Society of America* 87, 4 (1990), pp. 1728–1731.
- [Pyl06] PYLYSHYN Z. W.: *Seeing and Visualizing: It's not what you Think*. MIT Press, March 2006.
- [RBFW08] RAMANARAYANAN G., BALA K., FERWERDA J. A., WALTER B.: Dimensionality of visual complexity in computer graphics scenes. Rogowitz B. E., Pappas T. N., (Eds.), vol. 6806, SPIE.
- [Rec03] RECANZONE G. H.: Auditory influences on visual temporal rate perception. *Journal of neurophysiology* 89 (Feb 2003), pp. 1078–1093.
- [RFWB07] RAMANARAYANAN G., FERWERDA J., WALTER B., BALA K.: Visual equivalence: towards a new standard for image fidelity. *ACM Trans. Graph.* 26, 3 (2007).
- [RLGM92] ROCK I., LINNETT C. M., GRANT P., MACK A.: Perception without attention: results of a new method. *Cognit Psychol* 24, 4 (October 1992), pp. 502–534.
- [RMP04] ROBINSON A., MANIA K., PEREY P.: Flight simulation. In *Proceedings of the 2004 ACM SIGGRAPH international conference on Virtual Reality continuum and its applications in industry - VRCAI '04* (2004), ACM Press.

- [Rog25] ROGET P. M.: Explanation of an Optical Deception in the Appearance of the Spokes of a Wheel Seen through Vertical Apertures. *Philosophical Transactions of the Royal Society of London (1776-1886)* 115, -1 (1825), pp. 131–140.
- [Roo02] ROORDA A.: *Human Visual System - Image Formation*, vol. 1. 2002, pp. 539–557.
- [RPG99] RAMASUBRAMANIAN M., PATTANAIK S. N., GREENBERG D. P.: A Perceptually Based Physical Error Metric for Realistic Image Synthesis. In *Siggraph 1999, Computer Graphics Proceedings* (1999), Rockwood A., (Ed.), Addison Wesley Longman, pp. 73–82.
- [RS08] RECANZONE G. H., SUTTER M. L.: The biological basis of audition. *Annual Review of Psychology* 59, 1 (2008), pp. 119–142.
- [SC99] SIMONS D. J., CHABRIS C. F.: Gorillas in our midst: sustained inattentive blindness for dynamic events. *perception* 28 (1999), pp. 1059–1074.
- [Sch01] SCHOLL B. J.: Objects and attention: the state of the art. *Cognition* 80, 1-2 (June 2001), pp. 1–46.
- [Sch06] SCHUBERT E.: *Light-emitting diodes*. Cambridge University Press, 2006.
- [SD83] STAAL H. E., DONDERI D. C.: The Effect of Sound on Visual Apparent Movement. *The American Journal of Psychology* 96, 1 (1983), pp. 95–105.

- [SDC05] SUNDSTEDT V., DEBATTISTA K., CHALMERS A.: Perceived Aliasing Thresholds in High-Fidelity Rendering. In *APGV 2005 - Second Symposium on Applied Perception in Graphics and Visualization (poster)* (August 2005), ACM.
- [SDL*05] SUNDSTEDT V., DEBATTISTA K., LONGHURST P., CHALMERS A., TROSCIANKO T.: Visual attention for efficient high-fidelity graphics. In *SCCG '05: Proceedings of the 21st spring conference on Computer graphics* (2005), ACM Press, pp. 169–175.
- [SFWG04] STOKES W. A., FERWERDA J. A., WALTER B., GREENBERG D. P.: Perceptual illumination components: a new approach to efficient, high quality global illumination rendering. *ACM Trans. Graph.* 23, 3 (2004), pp. 742–749.
- [Shi64] SHIPLEY T.: Auditory flutter-driving of visual flicker. *Science* 145 (Sep 1964), pp. 1328–1330.
- [SJ01] STRAYER D. L., JOHNSTON W. A.: Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone. *Psychol. Sci.* 12, 6 (2001), pp. 462–466.
- [SJD00] SPENCE C., JANE R., DRIVER J.: Cross-modal selective attention: on the difficulty of ignoring sounds at the locus of visual attention. *Perception & Psychophysics* 62, 2 (2000), pp. 410–424.
- [SKS00] SHAMS L., KAMITANI Y., SHIMOJO S.: What you see is what you hear. *Nature* 408 (2000).
- [SKS02] SHAMS L., KAMITANI Y., SHIMOJO S.: Visual illusion induced

- by sound. *Brain Res Cogn Brain Res* 14, 1 (Jun 2002), pp. 147–152.
- [SLWP96] STEIN B. E., LONDON N., WILKINSON L. K., PRICE D. D.: Enhancement of Perceived Visual Intensity by Auditory Stimuli: A Psychophysical Analysis. *J Cog Neurosci* 8, 6 (1996), pp. 497–506.
- [SM04] SPENCE C., McDONALD J.: *The Handbook of Multisensory Processes*. MIT Press, Cambridge, MA, 2004, pp. 3–25.
- [Smi97] SMITH S. W.: *The scientist and engineer’s guide to digital signal processing*. California Technical Publishing, 1997.
- [SPP00] STEINMAN R. M., PIZLO Z., PIZLO F. J.: Phi is not beta, and why Wertheimers discovery launched the Gestalt revolution. *Vision Research* 40, 17 (2000), pp. 2257–2264.
- [SPS] SPSS: Spss statistics 17.0. <http://www.spss.com>.
- [SS01] SHIMOJO S., SHAMS L.: Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion in Neurobiology* 11, 4 (August 2001), pp. 505–509.
- [SS04] SHAMS L. K. Y., SHIMOJO S.: Modulations of visual perception by sound. in the handbook of multisensory processes (eds. calvert, g.a., spence, c. and stein, b.e.). pp. 27–33.
- [SSL97] SEKULER R., SEKULER A. B., LAU R.: Sound alters visual motion perception. *Nature* 385, 6614 (January 1997).
- [Sto98] STORMS R. L.: *Auditory-Visual Cross-Modal Perception Phenomena*. Phd thesis, 1998.

- [SUS94] SLATER M., USOH M., STEED A.: Depth of presence in virtual environments. *Presence* 3, 2 (1994), pp. 130–144.
- [TGD04] TSINGOS N., GALLO E., DRETTAKIS G.: Perceptual audio rendering of complex virtual environments. *ACM Trans. Graph.* 23, 3 (2004), pp. 249–258.
- [The91] THEEUWES J.: Exogenous and endogenous control of attention: the effect of visual onsets and offsets. *Perception & psychophysics* 49, 1 (1991), pp. 83–90.
- [TL04] TABELLION E., LAMORLETTE A.: An approximate global illumination system for computer generated films. *ACM Transactions on Graphics* 23 (2004), pp. 469–476.
- [Tsi07] TSINGOS N.: Perceptually-based auralization. In *19th Intl. Congress on Acoustics* (sep 2007).
- [VBdG98] VROOMEN J., BERTELSON P., DE GELDER B.: A visual influence in the discrimination of auditory location. *Terrigal*.
- [VdBOBT08] VAN DER BURG E., OLIVERS C. N., BRONKHORST A. W., THEEUWES J.: Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of experimental psychology. Human perception and performance* 34, 5 (October 2008), pp. 1053–1065.
- [VdG04] VROOMEN J., DE GELDER B.: Perceptual Effects of Cross-modal Stimulation: Ventriloquism and the Freezing Phenomenon. in *The Handbook of Multisensory Processes* (eds. Calvert, G.A., Spence, C. and Stein, B.E.). pp. 140–150.

- [VWH02] VAN BEERS R. J., WOLPERT D. M., HAGGARD P.: When feeling is more important than seeing in sensorimotor adaptation. *Current biology* 12, 10 (2002), pp. 834–837.
- [War94] WARD G. J.: The RADIANCE lighting simulation and rendering system. In *SIGGRAPH '94: Proceedings of the 21st annual conference on Computer graphics and interactive techniques* (1994), ACM Press, pp. 459–472.
- [Wel99] WELCH R. B.: Chapter 15 meaning, attention, and the "unity assumption" in the intersensory bias of spatial and temporal perceptions. In *Advances in Psychology: Cognitive Contributions to the Perception of Spatial and Temporal Events*, Gisa Aschersleben T. B., M?sseler J., (Eds.), vol. 129. North-Holland, 1999, pp. 371 – 387.
- [Whi80] WHITTED T.: An improved illumination model for shaded display. *Communications of the ACM* 23, 6 (June 1980), pp. 343–349.
- [WKN03] WADA Y., KITAGAWA N., NOGUCHI K.: Audio-visual integration in temporal perception. *Int J Psychophysiol* 50, 1-2 (October 2003), pp. 117–124.
- [WRC88] WARD G. J., RUBINSTEIN F. M., CLEAR R. D.: A ray tracing solution for diffuse interreflection. In *SIGGRAPH '88: Proceedings of the 15th annual conference on Computer graphics and interactive techniques* (1988), ACM Press, pp. 85–92.
- [WS98] WITMER B. G., SINGER M. J.: Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoper. Virtual Environ.* 7, 3 (June 1998), pp. 225–240.

- [WW80] WELCH R. B., WARREN D. H.: Immediate perceptual response to intersensory discrepancy. *Psychological bulletin* 88, 3 (November 1980), pp. 638–667.
- [Yar67] YARBUS A. L.: *Eye Movements and Vision*. Plenum Press, 1967.
- [Yat34] YATES F.: Contingency table involving small numbers and the x^2 test. In *Journal of the Royal Statistical Society (Supplement)* 1 (1934), Blackwell Publishing, pp. 217–235.
- [Yos00] YOST W. A.: *Fundamentals of hearing : an introduction*, 4th ed. Academic Press., 2000.
- [YPG01] YEE H., PATTANAİK S., GREENBERG D. P.: Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Trans. Graph.* 20, 1 (2001), pp. 39–65.

CHAPTER 9

Appendix A: Additional materials from
the study presented in Chapter 5

Questionnaire

Please fill in the following questionnaire:

Age: _____ Sex: _____

Occupation: _____

Country of origin? _____

Do you have normal or corrected to normal vision? Yes No

Do you have any hearing impairments? Yes No

How much do you use computers?

Not at all A little bit Moderately Very much

How familiar are you with Computer Graphics?

Not at all A little bit Moderately Very much

How much do you play computer games?

Not at all A little bit Moderately Very much

THANK YOU!

Instructions

INSTRUCTIONS:

We are doing a research in Computer Graphics on visual perception. This experiment is about image quality perception. You will be presented with 28 pairs of images. Each image pair will be shown in a sequence of A and B image. Following each pair, you will be prompted to select the image with a better quality (A or B). You should respond using keyboard.

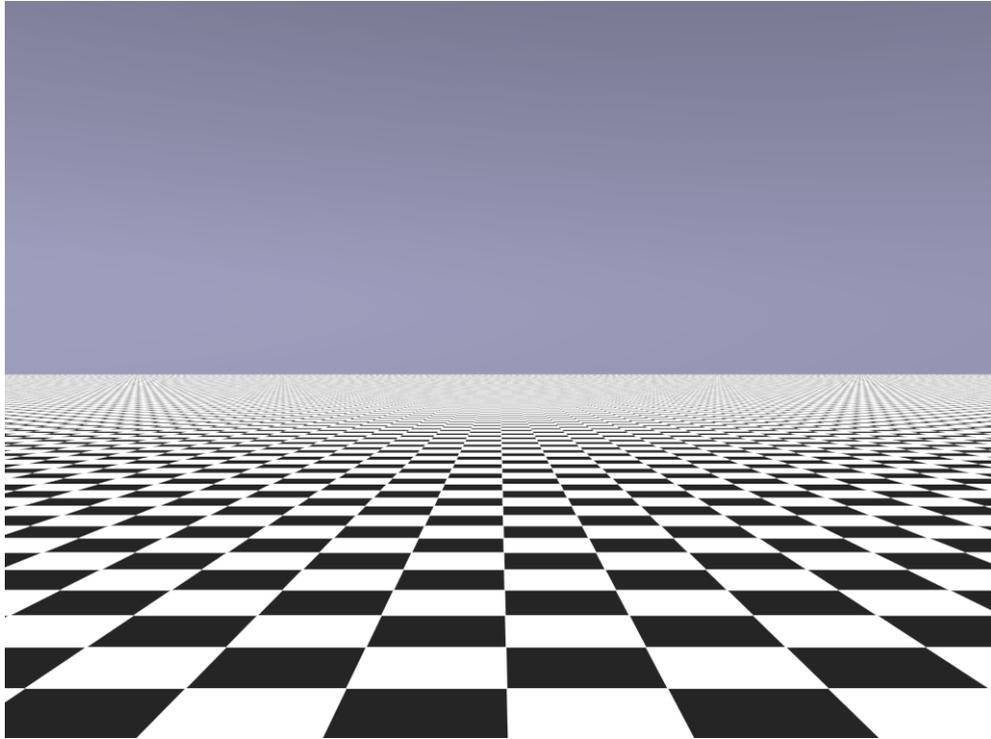
If you think the first image (A) had better quality, press 1.

If you think the second image (B) had better quality, press 2.

Do not worry about pressing any other buttons on the keyboard. Those will be ignored. After pressing 1 or 2 a new image sequence will start automatically.

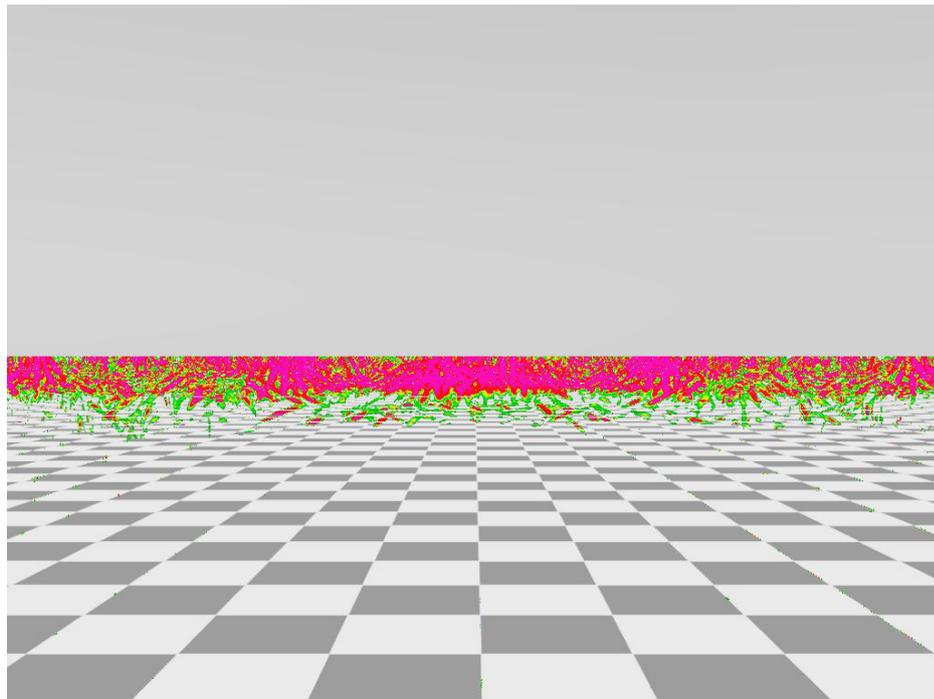
Here is the example of the experiment with only one image pair. Watch them and choose the better (higher quality) one.

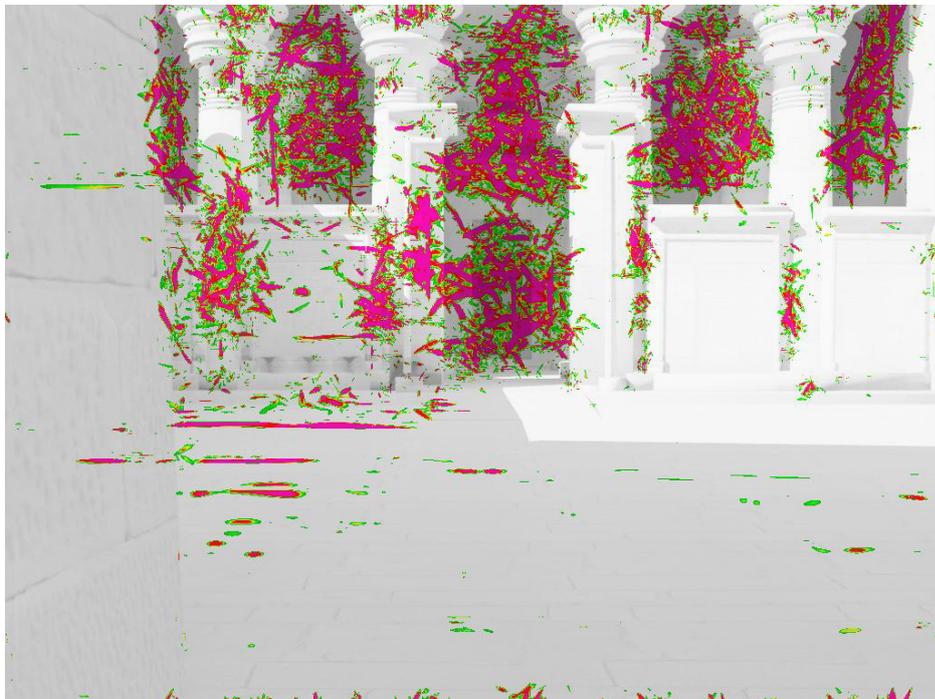
Visual stimuli

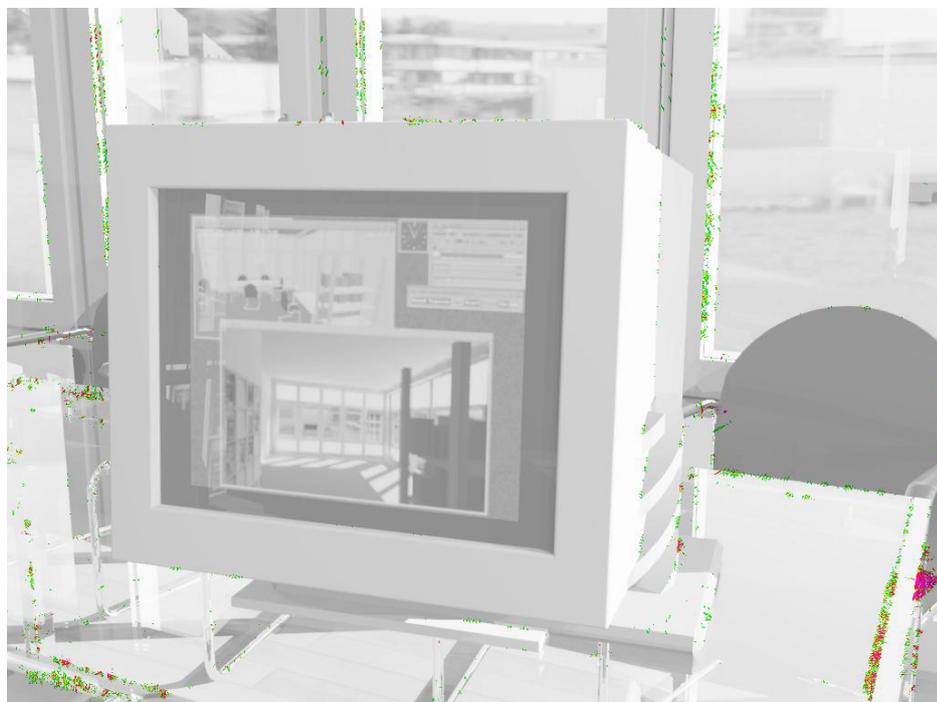




VDP comparison: Examples of the VDP comparison between rendered images generated using 9rpp and 16rpp. From top to the bottom: Checkerboard, Corridor, Kalabsha and Library.







CHAPTER 10

Appendix B: Additional materials from
the study presented in Chapter 6

Instructions

Instructions:

We are doing research in Computer Graphics and this one in particular is about the visual perception. The best way of evaluating the results and testing the human visual system and visual perception is conducting the psychophysical experiments.

This experiment is about the frame rate perception. Every animation/video that we see in the cinema, on the TV or computer screens consists of series of consecutive images, called frames. If these images are delivered quickly enough the animation will appear as smooth, while if there are just few frames shown in a time unit, the animation will look jerky. That frequency of image delivery in an animated content is called the frame rate. More frames we show in a time unit the higher frame rate we get (smoother), while less frames shown means lower frame rate (jerky).

In this experiment, your task is to judge the smoothness of the animations you see, moving the slider shown after each animation. The slider range is from 0 to 100, where 0 represents a jerky animation and 100 the smooth one. Each animation lasts for ten seconds. Sliders will be shown for 5 seconds after which the next animation will follow automatically.

Here is the example of the experiment with just two animations. The first one is the worst case with low frame rate. The second one is the best case with high frame rate. You can see the difference in smoothness of these two animations.

Show training animations.

After each animation the slider will be positioned in the middle of the slider bar.

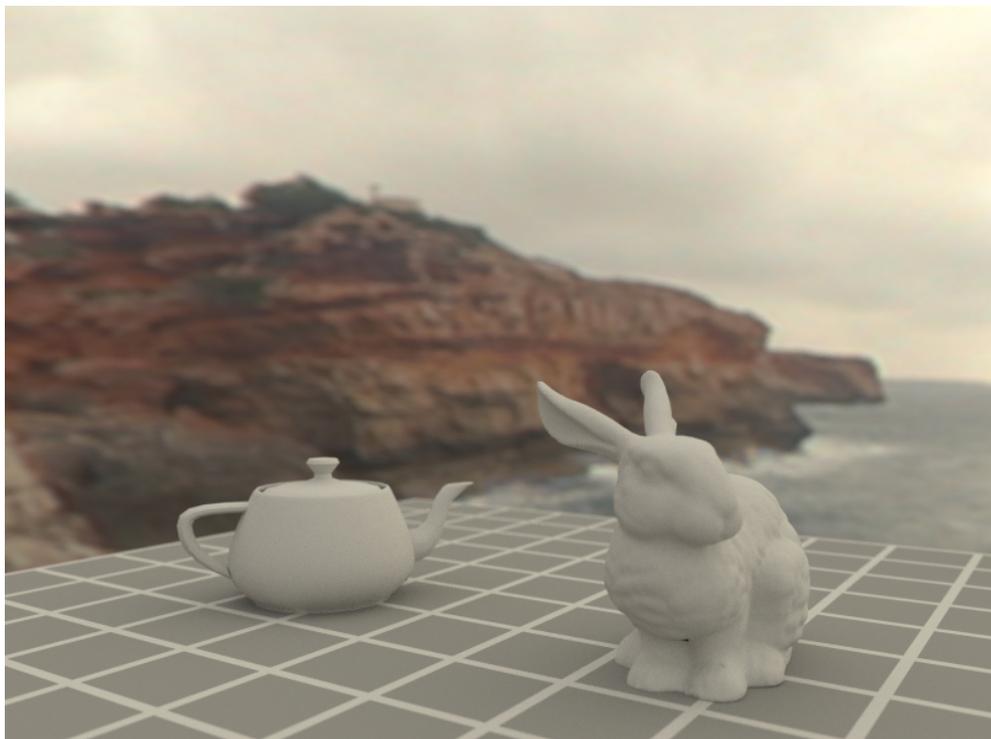
Some of the animations will be accompanied with sound, some of them not.

Could you please sign this consent form and fill in the questionnaire?

Could you please put on the headphones?

Once you are ready press the space button to start the experiment.

Visual stimuli







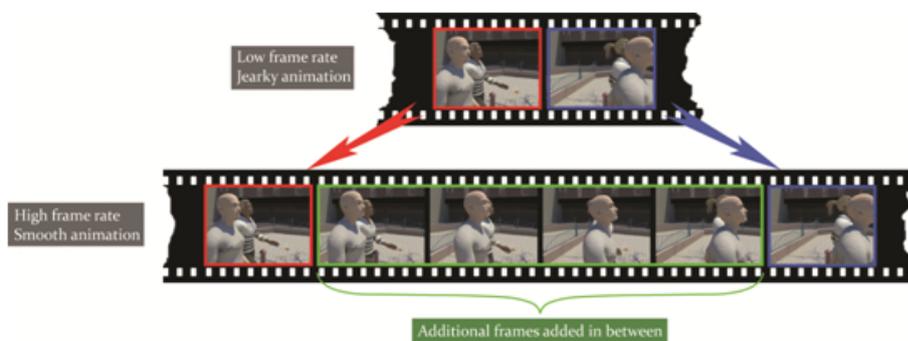
CHAPTER 11

Appendix C: Additional materials from
the study presented in Chapter 7

Instructions

Instructions

Every animation/video that we see in the cinema, on the TV or computer screens consists of series of consecutive images, called frames. If these images are delivered quickly enough the animation will appear as smooth, while if there are just few frames shown in a time unit the animation will look jerky. That frequency of image delivery in an animated content is called the frame rate. The more frames we show in a time unit the higher frame rate we get (smoother animation), while less frames shown means lower frame rate (jerkier animation).



In this experiment you will be asked to evaluate the smoothness of the presented animations.

In the experiment you will be shown 37 pairs of animations. In all the animations the different parts of the same walk-through will be presented. The speed (running or walking) and the frame rate (frequency) of the animations will vary.

Please watch the animations carefully.

Prior to each animation a grey box will be displayed. Then, animation will be presented. After each pair a screen with two boxes: A and B will be presented.

If you think the first video was of better quality, please click on A and if you think the second video was of better quality please click on B. The next pair of animations will start automatically after you click.

Some of the animations will be accompanied by sound and some of them will be silent.

Now you will be shown an example of the experiment with just two animation pairs. In both pairs the first animation is the worst case with the lowest frame rate (jerky). The second animation will represent the best case i.e. the highest frame rate (smooth).

Visual stimuli



