Author:
Mastoropoulou, Georgia

Title:
The effect of audio on the visual perception of high-fidelity animated 3D computer graphics

The Effect of Audio on the Visual Perception of High-Fidelity Animated 3D Computer Graphics

Georgia Mastoropoulou

A thesis submitted to the University of Bristol, UK in accordance with the requirements for the degree of Doctor of Philosophy in the Faculty of Engineering, Department of Computer Science.

2006

Declaration

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

Signed: Date: September 29, 2006

Georgia Mastoropoulou

Acknowledgements

First of all, my thanks must go to Professor Alan Chalmers, my dissertation supervisor, for his continuous encouragement, enthusiasm and support in my research endeavours; without his understanding and his 'optimism' my PhD would have never happened.

I also extend my gratitude to the Greek State Scholarships Foundation (IKY), from which I received funding for my PhD.

Next, I want to thank all the members of the graphics group at Bristol, both past and present, who made it such a friendly and stimulating place to work. In particular Kurt, who provided such a great help with his expertise in Radiance and graphics programming! Many thanks must go to Francisco and Matt, for being great 'neighbours' in the lab, and to Anna and Kurt, again, for offering their hospitality when I moved from Bristol. Thanks are also due to my colleagues and the people working in the Department of Computer Science in Bristol who have helped me along the way.

Thanks also to Professor Tom Troscianco and my friend Maria for their invaluable comments regarding statistical analysis and experimental design of psychophysical experiments.

Special thanks to my brother Christos for taking such a good care of me during his stay in Bristol. Many thanks to Hsiou for her friendship, her support and her humour (and also for her great cooking efforts!); to Panos and Ioulia for being buddies and sharing a flat with me for a year; to the people I met at the conferences, especially Greg Ward, for providing invaluable feedback and advice and making the conferences such enjoyable experiences.

Many thanks also to the hundreds of people who volunteered their time to participate in my experiments; without their help this research effort would have never reached any conclusions.

My friends in Greece remained a great support throughout this thesis, despite my always being late at replying emails and calling back. Apologies to everyone, and thanks for the great understanding.

My family, in its extended form now, has always supported me in everything I have done. There is no way I can put down on paper how thankful I feel and how much you all mean to me.

Finally, none of this would have been possible if it were not for the love of my George. This whole Ph.D. process has perhaps been harder on him than it has been for me. Thanks for sticking it out with me. I love you!

This work is dedicated to George.

ABSTRACT

Sound is often an integral part of interactive animated scenarios, such as VR applications, games and realistic simulations. Up to now, research has focussed on how visual stimuli can affect the user's perceived quality of the rendered graphics. Although there is a plethora of evidence coming from the psychology field about crossmodal interactions between visual and auditory stimuli, graphics researchers have not yet considered exploiting sound in order to affect the perceived quality of the 3D visual environment. Furthermore, although it is well known that sound is attention grabbing, researchers until now have not considered using sound to direct gaze to specific objects/areas in the visual environment, in order to allow for the selective rendering of the scene: only the sound emitting objects at high quality, while the reduced quality of the rest of the 3D scene goes unnoticed by the observers.

This gap we are trying to fill with our research in the influence of auditory stimuli on the perceived visual quality (rendering quality and frame rate) of computer generated animated imagery.

To gain a better understanding of the crossmodal interactions between the auditory and visual sensory modalities and identify whether such interactions could lead to a new generation of perceptually-adaptive graphics techniques, that would take into account not only the visual stimuli but also the auditory background of a 3D scene, 292 subjects participated in five experiments. Temporal and visual display quality perceptions were investigated by manipulating the frame rate and the rendering quality (number of rays shot per pixel of the image), separately, and by considering different auditory backgrounds.

Our experimental studies verified that we can affect the viewer's perception of delivered frame rate with the use of audio. Further results show that the viewers do fixate to sources of sound effects in a scene- even when engaged in a demanding visual task- allowing us to render the corresponding pixels to high quality and significantly drop the quality for

the rest of the scene, without any noticeable difference to the observer. In both cases, we save processing resources and/or significant computational time.

Contents

LIST OF TABLES xv

LIST OF FIGURES xxv

1 Introduction **1**

1.1 Motivation 1

1.2 The Research Problem and its Importance 3

1.3 Interdisciplinary Approach 5

1.4 Aims and Objectives of the Research 7

1.5 Thesis Outline 9

2 Psychoacoustics and Computer Graphics **12**

2.1 Introduction 12

2.2 The Psychology of Hearing 14

2.2.1 Perceptual grouping principles 14

2.2.2 The Spatial Attributes of Sound 15

2.2.3 Intersensory effects on Spatial Sound Perception 20

2.3 Moving to multimodality: More Auditory-Visual Sensory Interactions . . 23

2.3.1 Theories of intersensory interactions 24

2.3.2 How Sound affects the Visual Stimuli in the Temporal Domain . 25

2.4 Spatialised Sound for Multisensory Environments 28

2.5 Auditory Displays of Spatialised Sound 30

2.5.1 Headphone Simulation 31

2.5.2 Simulation Using Speakers 33

2.6 Summary 35

3 Perception, Attention and Computer Graphics **37**

3.1 Definitions of Attention 38

3.2 Perception Depends on Attention 40

3.2.1 Inattentional Blindness 41

3.3 Attention: Automatic or under endogenous control? 44

3.4 Attentional Selection between Spatially Defined Visual and Auditory Stimuli 46

3.4.1 Attentional Capture by distractors 48

3.5 Theories of Selective Attention 52

3.5.1 Early Theories of Attention: Bottleneck Theories 53

3.5.2 Limited Capacity Theories 54

3.6 Crossmodal Attention 56

3.6.1 Attending to one sensory modality versus another 58

3.6.2 Crossmodal links in spatial attention 59

3.7 Spatial Attention and Eye Movements 63

3.7.1 Types of Eye Movements 64

3.7.2 Eye Movement Control 66

3.7.3 How Shifts of Attention are related to Eye Movements 69

3.7.4 Tracking of Eye Movements 72

3.8 Human Perception and Computer Graphics 74

3.8.1 Rendering Quality/Fidelity Perception 75

3.8.2 Crossmodal Interactions on the Perception of Quality 80

3.8.3 Perceptually-Aware Rendering Techniques 84

3.9 Summary 94

4 Preliminary Investigation of Temporal Perception 99

4.1 Introduction 99

4.2 Experimental Methodology 101

4.2.1 Hypothesis 101

4.2.2 Participants 102

4.2.3 Design 102

4.2.4 Equipment and Materials 110

4.2.5 Procedure 111

4.3 Results 112

4.4 Discussion 120

5 Perceived Frame Rate under the Influence of Music and Sound Effects 122

5.1 Introduction 122

5.2 Experiment 1 - The Influence of Sound Effects and Music on the Perceived Smoothness of Rendered Animations 124

5.2.1 Participants 125

5.2.2 Design 126

5.2.3 Equipment and materials 127

5.2.4 Procedure 130

5.2.5 Results 132

5.2.6 Discussion 142

5.3 Experiment 2 on The Influence of Sound Effects on the Perceived Smoothness of Rendered Animations 142

5.3.1 Participants 143

5.3.2 Design 144

5.3.3 Equipment and materials 145

5.3.4 Procedure 148

5.3.5 Results 150

5.4 Summary 159

6 Experiments on the Perceived Rendering Quality under the influence of Sound Effects **160**

6.1 Introduction 160

6.2 Experiment on the Perceived Rendering Quality under the Influence of Sound Effects 163

6.2.1 The Selective Renderer 163

6.2.2 Participants 168

6.2.3 Design 169

6.2.4 Equipment and materials 169

6.2.5 Procedure 171
6.2.6 Results 174
6.2.7 Discussion 176
6.3 Eye tracking Experiment 177
6.3.1 Experimental Set Up 177
6.3.2 Participants 181
6.3.3 Design 182
6.3.4 Stimuli 184
6.3.5 Procedure 187
6.3.6 Results 188
6.4 Summary 197
7 Conclusions and Future Work **199**
7.1 Achievement of Goals 200
7.2 Thesis Contributions 203
7.3 Future Work 205
REFERENCES 208
APPENDICES 228

A Experiment on Temporal perception - Questionnaire 229

B Eye-tracking Experiment - Resulting Scan Paths 231

List of Tables

5.1 Frame Rate Experiment 1 - The "No Sound" condition results in percentages, separately for Familiar and Unfamiliar subjects, across the trial frame rate pairs (ordered according to frame rate difference within the pair of rates). 139

5.2 Frame Rate Experiment 1 - The "Sound Effects" condition results in percentages, separately for Familiar and Unfamiliar subjects, across the trial frame rate pairs (ordered according to frame rate difference within the pair of rates). 140

5.3 Frame Rate Experiment 1 - The "Music" condition results in percentages, separately for Familiar and Unfamiliar subjects, across the trial frame rate pairs (ordered according to frame rate difference within the pair of rates). 141

6.1 The results from our experiment summarised across the conditions. 174

List of Figures

2.1 The primary binaural cues of sound direction localisation. Image reproduced from [18]. 17

2.2 Acoustic Paths. Image courtesy of Prof. R.O. Duda. 19

2.3 Basic auralisation pipeline. 30

2.4 Illustrations of a 7.1 (left image) and a 6.1 (right image) Surround Sound System. 35

3.1 In a dramatic demonstration of inattentional blindness, half of the observers participating in Simons and Chabris' experiment [196] failed to notice a person wearing a gorilla suit (right image) and 35% failed to see a woman holding an open umbrella (image on the left) who walked into the middle of a basketball game, spending several seconds on screen. Images courtesy of Prof. Daniel J. Simons. 43

3.2 During visual perception and recognition, human eyes move and successively fixate at the most informative parts of the image (from [271]). . . . 67

3.3 "An Unexpected Visitor" by Repin. 67

3.4 Recordings of saccadic eye movements scanning "The unexpected Visitor" (from [271]). 68

3.5 According to Noton and Stark (1971) [156], each object is memorised and stored in memory as an alternating sequence of object features and eye movements required to reach the next feature. 69

3.6 Comparing real and synthetic scenes using human judgements of lightness perception. Image courtesy of [136]. 77

3.7 The initial bottom-up model of attention introduced by Itti, Koch and Niebur [91, 92, 93]. The input image is separated into three parallel feature channels (color, intensity, and orientation) and sampled at a series of spatial scales. Feature activity is propagated to the next level and reorganised into a center-surround fashion. Activity is normalised within each feature channel and linearly summed to form the salience map. Attentional focus is determined through a winner-take-all network. Once attended to, the current location is transiently inhibited in the saliency map by an inhibition-of-return (IOR) mechanism. Image courtesy of [93]. 88

3.8 An example of multi-resolutional images used in GCMRDs, where high resolution information is put only where the user is looking at each moment, and lower resolution everywhere. Image courtesy of [119]. 93

4.1 Less number of frames per second may be displayed when musical 'distractors' applied, without any noticeable difference to the observer. 100

4.2 Experiment on Temporal Perception - The conditions tested. 103

4.3 Example frames from the animated sequence used for the first pilot study. 105

4.4 The results of the first pilot study on Temporal Rate Perception. The figures represent numbers of participants who gave the corresponding answers. 107

4.5 Example frames from the Base1 and Base2 animated sequences used for the second pilot study and the main experiment. 108

4.6 Results of the second pilot study on Temporal Rate Perception. The figures represent numbers of participants who gave the corresponding answers. 109

4.7 Experiment on Temporal Perception - Results for the relative perceived duration across conditions. The figures represent numbers of participants who gave the corresponding answers. 113

4.8 Experiment on Temporal Perception - Results for the relative perceived motion velocity across conditions. The figures represent numbers of participants who gave the corresponding answers. 114

4.9 Experiment on Temporal Perception - t-test results for the perceived duration regarding the "Exciting music" groups. 114

4.10 Experiment on Temporal Perception - t-test results for the perceived duration regarding the "Relaxing music" groups. 115

4.11 Experiment on Temporal Perception - t-test results for the perceived motion velocity regarding the "Exciting music" groups. 115

4.12 Experiment on Temporal Perception - t-test results for the perceived motion velocity regarding the "Relaxing music" groups. 115

4.13 Experiment on Temporal Perception - Results in percentages for the '12 vs. 16 fps' (top) and the '16 vs. 20 fps' (bottom) conditions. 117

4.14 Experiment on Temporal Perception - Graphs of the Perceived Duration across conditions 118

4.15 Experiment on Temporal Perception - Graphs of the Perceived Camera Motion across conditions 119

5.1 With the use of auditory 'distractors', such as sound effects, fewer frames may be displayed per second compared to a silent animation, without any noticeable difference in the motion smoothness. 123

5.2 Frame Rate Experiment 1 on the influence of sound effects and music on the perceived smoothness of rendered animations - The Conditions tested. 127

5.3 Frame Rate Experiment 1 - Example frames from the animated sequence used for the experiment. 128

5.4 Frame Rate Experiment 1 - Timeline depicting the period during the 8-second animation that each sound effect was audible. 129

5.5 Frame Rate Experiment 1 - The results for the Familiar subjects across the 3 conditions, given as the number of correct answers in the total of 49 rate pairs. 132

5.6 Frame Rate Experiment 1 - The results for the Unfamiliar subjects across the 3 conditions, given as the number of correct answers in the total of 49 rate pairs. 132

5.7 Frame Rate Experiment 1 - The performance of Familiar vs. Unfamiliar Subjects for the Control ("No Sound") condition across the trial frame rate pairs, which are ordered in our graph according to frame rate difference within the pair of rates. 133

5.8 Frame Rate Experiment 1 - The performance of Familiar vs. Unfamiliar Subjects for "Sound Effects" condition across the trial frame rate pairs (ordered according to frame rate difference within the pair of rates). . . . 134

5.9 Frame Rate Experiment 1 - The performance of Familiar vs. Unfamiliar Subjects for "Music" condition across the trial frame rate pairs (ordered according to frame rate difference within the pair of rates). 134

5.10 Frame Rate Experiment 1 - The performance of Familiar Subjects across all conditions. 135

5.11 Frame Rate Experiment 1 - The performance of Unfamiliar Subjects across all conditions. 136

5.12 Frame Rate Experiment 1 - The results of the 3 × 2 ANOVA, which examined whether there was a significant between-subjects performance difference as a combined function of the auditory background difference and the level of familiarity with animated computer graphics. 137

5.13 Frame Rate Experiment 2 - The Conditions tested 145

5.14 Frame Rate Experiment 2 - Example frames from the animated sequences used for the experiment. 146

5.15 Frame Rate Experiment 2 - Visual signals indicating the beginning of the first (left image) and the second clip (right image) within each test pair. . 149

5.16 Frame Rate Experiment 2 - The performance of Familiar vs. Unfamiliar Subjects for the control ("No Sound") condition across the test frame rate pairs. 151

5.17 Frame Rate Experiment 2 - The performance of Familiar vs. Unfamiliar Subjects for the "Sound Effect" condition across the test frame rate pairs. 151

5.18 Frame Rate Experiment 2 - The performance of all Subjects across the "No Sound" and "Sound Effect" conditions, separately for each frame rate combination. 153

5.19 Frame Rate Experiment 2 - How errors (i.e. incorrect detections of the higher frame rate within each pair of animations) were distributed between Familiar and Unfamiliar subjects in the "Sound Effect" condition. . 153

5.20 Frame Rate Experiment 2 - The results of the two-way ANOVA, which examined whether the performance in the experimental task was jointly influenced by the auditory background and the familiarity of the subjects with animated computer graphics (i.e. whether there is an interaction between these two independent variables). 154

5.21 Frame Rate Experiment 2 - The results of the unpaired t-test between the means of the independent "No Sound" and "Sound Effect" conditions, separately for Unfamiliar subjects (left) and Familiar subjects (right). . . . 155

5.22 Frame Rate Experiment 2 - The performance of all subjects across the 2 types of camera motion (translation and rotation). 156

5.23 Frame Rate Experiment 2 - The results of the two-way ANOVA, which examined whether the performance in the experimental task was jointly influenced by the auditory background and the type of camera motion in the 3D scene. 157

5.24 Frame Rate Experiment 2 - The results in percentages for the two auditory background conditions, separately for Familiar and Unfamiliar subjects, across the trial frame rate pairs (ordered according to frame rate difference within the pair of rates). 158

6.1 Our Selective Rendering approach. It renders at higher quality the sound emitting objects (SEOs) and the surrounding pixels, while reducing the rendering quality of the rest of the pixels, by shooting to each a lower number of rays (what we call reduced rendering quality, RQL). When there are no SEOs on screen all the pixels are rendered at the predefined high quality. 162

6.2 Our Selective Rendering Pipeline. 165

6.3 Example visualisations of the q-buffer, as maps depicting the quality gradient around the sound emitting object (phone), for a frame selectively rendered using high detail insets of size 150 × 150 (left) and 350 × 350 (right). In the maps, white represents the highest quality and black the base (reduced) quality, 1 ray per pixel. 167

6.4 Rendering Quality Experiment - Single frame from the walkthrough in the 3D scene which represented the visual stimuli for our experiment. . . 171

6.5 The area, 350x350, rendered at higher quality in one of the selectively rendered frames. In this area, the sound emitting object is rendered at the highest quality, 16 rays per pixel, and the quality of the surrounding pixels is gradually reduced as their distance from the boundary of the object increases. 172

6.6 Rendering Quality Experiment - Close up of scene details rendered at high quality (16 rays/pixel) and reduced quality (1 ray/pixel), respectively. 172

6.7 Saliency map of the example frame depicted in Figure 6.4. The brightest areas in the map represent the areas of greatest saliency. 173

6.8 Preliminary eye tracking tests: Example saccades and fixations on the phone while the ringing sound was audible. 177

6.9 The Tobii x50 desk mounted eye tracker. 178

6.10 The overall hardware experimental setup for the eye tracking experiment. 178

6.11 Tobii x50 and a monitor, TV or projection screen setup. Image courtesy of *Tobii Technology* (http://www.tobii.com) 179

6.12 Snapshots of the eye tracker calibration procedure. The subject is asked to fix his gaze on the dot, which is presented successively and randomly at different positions on the monitor screen, every time it appears. The bottom picture is an example of a very good calibration quality. The left panel includes the calibration points for the left eye and the right panel the calibration points for the right eye. Where the calibration quality is very good, the green and red marks coincide. 181

6.13 The Conditions tested in the Eye tracking Experiment. 182

6.14 The task objects combinations used for the "Task" condition in our experiment. 183

6.15 Example frames from the walkthrough in the 3D scene which represented the visual stimuli for our eye tracking experiment. 186

6.16 A participant seated at the eye tracker during the experimental session. . . 188

6.17 Eye tracking experiment - Example saccades and fixations from the "Free-view" condition. 189

6.18 Eye tracking experiment - Example saccades and fixations from the "Task" condition. 190

6.19 Top: Combined scan paths for the "Freeview-Sound Effect" group, regarding the 3-second period that the sound effect of the ringing phone was audible. Bottom: Fixation hotspots for the same time interval. The deeper the colour of the hotspot is, the more fixations the corresponding area has attracted. 191

6.20 Top: Combined scan paths for the "Freeview-No Sound" group. Bottom: Corresponding fixation hotspots. 192

6.21 Top: Combined scan paths for the "Task-Sound Effect" group. Bottom: Corresponding fixation hotspots. 193

6.22 Top: Combined scan paths for the "Task-No Sound" group. Bottom: Corresponding fixation hotspots. 194

6.23 Example saccades and fixations on the phone while the ringing sound was audible. 195

6.24 Closeup of the fixation hotspots for the "Freeview" (left image) and "Task" (right) conditions, around the time that the ringing sound was audible. The deeper the colour of the hotspot is, the more fixations the corresponding area has attracted 195

6.25 Eye tracking experiment - Performance (accuracy) in the search and memory task (percentage of correct counts of the target objects across participants) as a function of the auditory background. 197

6.26 Eye tracking experiment - Performance (accuracy) in the search and memory task given separately for the two combinations of task objects. *TaskObjs_1* represents the "Orange cylinder-red ball" target objects, while *TaskObjs_2* are the "Red cylinder-blue ball" target object combination. 197

7.1 Our bottom-up visual attention model for arbitrary animations which include sound emitting objects. 207

B.1 Scan paths of the participants in the "Freeview-Sound Effect" group, corresponding to the 3-second period that the sound effect of the ringing phone was audible. 232

B.2 Scan paths of the participants in the "Freeview-No Sound Effect" group (corresponding to the 3-second period that the sound effect was audible in the audiovisual animation). 233

B.3 Scan paths of the participants in the "Task-Sound Effect" group, corresponding to the 3-second period that the sound effect of the ringing phone was audible. 234

B.4 Scan paths of the participants in the "Task-No Sound Effect" group (corresponding to the 3-second period that the sound effect was audible in the audiovisual animation). 235

Chapter 1

Introduction

1.1 Motivation

Despite the huge progress in the performance of graphics-related hardware and algorithms during the past few years, high fidelity renderings of complex scenes may still take several minutes to render on a single computer. For real time interactive applications, such as VR environments, simulations and games, where a minimum frame rate needs to be maintained and therefore the time for rendering and displaying the generated images is a fundamental issue, speed-fidelity trade-offs need to be made, while trying to keep the perceptibility of any resulting anomalies to the minimum. At the same time, users constantly demand graphics environments more aesthetically pleasing and of higher visual fidelity, in order to enhance their sense of immersion.

Research on visual perception has shown that the perceived quality of the rendered graphics depends not only on the fidelity of the generated imagery, but also on mechanisms of visual attention and on the characteristics of the human visual system. Perception findings suggest that a viewer faces similar constraints to identify scene elements as the renderer

does in displaying them and therefore elements that are likely to be outside the focus of the viewer's attention may be rendered at a degraded quality, without any noticeable difference to the observer.

These findings promoted the development of adaptive techniques in Computer Graphics (CG) which employ perceptually-based criteria to reduce the computational complexity of the rendering solution, by focussing only on those features that are readily perceivable under certain viewing conditions. For a recent review of perceptually-based rendering techniques see also [160].

Despite the recent growth in the development of perceptually adaptive graphics techniques, researchers in the rendering field have restricted their focus on visual stimuli and have not as yet taken into consideration crossmodal interactions, although it is well known that stimuli reaching the various senses are, in general, not processed independently. For instance, the interpretation of human speech relies on both seeing the lip movements and hearing the sounds. In the real world there is an intimate linkage between sound and visual stimuli and in the recent past a number of researchers investigated such auditory-visual interactions.

The research findings in intersensory phenomena strongly suggest that the fidelity and perceptual 'realism' of 3D graphics environments must be based on multimodal criteria comprising all of our senses, as opposed to the current use of unimodal criteria. However, insufficient experimental data exists to make informed multimodal design decisions for the rendering and delivery of animated 3D scenarios, and computer graphics would benefit from further research in this topic. Since most current 3D graphics environments include mainly visual and auditory stimuli, it seems rational to focus research efforts on vision and audition. Sound is used anyway in VR and gaming applications, therefore, it is very interesting to seek ways for graphics developers to take advantage of auditory stimuli, in order to reduce the computational load for the rendering and the delivery of 3D graphics.

1.2 The Research Problem and its Importance

Our research relates to the popular demand over the past 20 years for 3D computer displays that have a high visual fidelity and deliver graphics at fast frame rates, to enable a greater perceptual experience for the viewer. However, these needs are hard to meet due to computational complexity and existing processing limitations. Optimisation of graphics rendering under varying resource constraints poses challenges at the interface between computation and cognition. It is well known in the human perception community that many factors, including auditory stimuli, may influence a human's cognitive resources available to perform a visual task. In addition to reducing the available perceptual processing resources available to process the stimuli from the visual domain, sound attracts spatial attention to specific parts of the scene (the perceptual origin of the sound, which receives the "spotlight" of attention), leaving the rest of the scene practically unattended. From the research findings in multimodal perception we can infer that the redirection of attention and the allocation of cognitive resources to the processing of auditory stimuli while watching rendered animations, may potentially reduce a viewer's ability to perceive artifacts resulting from a reduced quality of the visuals.

Our goal is to create images of the highest possible quality by minimising the expected perceptual cost of frames, subject to the constraints on available resources. For graphics architectures that commit to a constant frame rate, the time for rendering a scene to the 'gold standard' may be less than the time available. In such cases, degraded approximations of the rendering solution must be considered, while trying to keep the perceptibility of the resulting artifacts to the minimum. On the other hand, in systems that allow for variable frame rates, we must consider the cost of diminishing the frame rate, along with other kinds of degradations that could be performed (although we must bear in mind that

a general investigation of all feasible degradation actions of a scene is intractable).

The manipulation of the perception of frame rate and/or the perception of the fidelity of the visuals in the presence of auditory distractors in animated scenarios, could help towards more realistic graphics at interactive (or almost interactive) rates. If, with the help of audio, we can affect the user's perception of delivered frame rate so that he/she does not notice anomalies resulting from reduced frame rates, such as motion jerkiness, then we may render less frames per second and thus pursue a higher image quality per frame, without any cost in the user's perception of the overall scene quality.

In addition, video compression algorithms, which have started to take into account perceptual issues for low bit-rate applications, would also benefit from our findings. The spatial and temporal compression artifacts of coded video have recently been studied intensively. It is difficult to support both good spatial and temporal quality at very low bit rates. Up until now, developers have mainly opted for the degradation of the spatial quality under a fixed frame rate. Apteker et. al (1995) examined the effects that degrading frame rates has on user perception of a video application and showed that the perceived differences depend on the nature of the application [6]. According to Song et al. (2001), "more flexible and robust rate control is needed under time-varying communication channels, such as the Internet, and under these environments, variable-encoding frame-rate control can provide a satisfactory solution" [199]. They developed a variable-encoding frame rate control scheme which pursues an efficient tradeoff between spatial and temporal qualities. In our case, by encoding frame rate control, the sudden frame skipping which results from existing techniques and degrades motion smoothness significantly, could be reduced.

Also, by identifying the area(s) where the user will probably fixate, determined by the presence of sources of sound effects, we may render only the corresponding pixels to high quality (while significantly dropping the quality for the rest of the scene) and greatly reduce computation time.

We seek to investigate how degradations along different dimensions of quality approximation (i.e. frame rate and rendering quality) under the influence of auditory stimuli can influence the perception of the visual quality of a 3D scene. In addition, we wish to understand how a user's attention to different temporal and spatial components of an animated scenario and also to different sensory stimuli (auditory, visual) can change impressions of quality. After all, our goal is to provide content that is visually satisfying to people. This path of research incorporates multiple fields of study as explained in the following section.

1.3 Interdisciplinary Approach

Satisfactory answers to the problems of limited available transmission bandwidth and processing limitations in three-dimensional displays can only be gained through a combined understanding and employment of relevant research, algorithms, and experimental findings from the fields of Computer Graphics and Psychology.

Perception, and in particular visual perception, is becoming increasingly important in computer graphics. The use of modern graphics hardware is combined with special rendering techniques in order to trade off between visual quality and rendering time. These techniques exploit limitations of the human visual perception, in order to produce imagery of quality lower to the 'golden standard', but with as little perceptual difference to the observer as possible and with a significant saving in computational time. Knowledge of the human visual system has been used to improve the quality of the displayed image, for example [175, 162, 145]. Other research has shown how images can be selectively rendered without perceptual difference to the user, using level-of-detail, peripheral vision, saliency and visual tasks, for example [120, 248, 273, 31]. A review on the latest advances in perceptually adaptive computer graphics can be found in O'Sullivan et al. [160].

The human visual system is physically incapable of capturing a scene in full-detail. Attention determines which portions of the available visual information will be noticed and further mentally processed and which will be ignored [26, 30]. Also, according to the established phenomenon of "inattentional blindness", there is no conscious perception without attention [122]. On the basis of the limits on attention, Rensink (2000) claims that "there is usually only one object in play at any one time ... tasks involving more than a few objects can be handled by rapidly switching attention between the objects" [178] (see also [196, 195]). From personal experience, when viewers are engaged in a highly interactive application, such as playing a racing game, their attention is usually limited to very small portions of a computer screen. Furthermore, the faster the interaction, the more dramatic the effect is.

Our attention is directed to things we may be interested in, objects that are salient, or perhaps unusual. Although attention may be voluntarily allocated, sometimes an intense novel stimulus may capture it [166]. Capture relates to an involuntary access to (unexpected) events or objects that deserve attention due to salient features. What is relevant for our discussion is that attentional capture is a crossmodal phenomenon. For instance, an auditory event also 'attracts' visual attention and facilitates the detection of events/objects within the same region of space [48], regardless of the visual perceptual load and of any task the viewer may be engaged to [219]. The general conclusion from the research in auditory-visual perception, is that sound influences visual perception and vice-versa, both in the spatial and temporal domain, see for example, [9, 210, 245, 261].

Intersensory phenomena have been studied for many years by researchers in numerous disciplines such as: Psychoacoustics, Psychology, Physiology, Neurology, Philosophy and Computer-Human Interaction. Thus, there is a large amount of intersensory research, but there is little cross-disciplinary transfer of intersensory knowledge. Computer graphics in particular is severely lacking in its understanding and use of intersensory phenomena; The results currently available in the field of graphics rendering regarding crossmodal

interactions, are too few and computer graphics practitioners would benefit a lot from further study in this topic. The main goal of our research effort is to aid the development of high fidelity interactive 3D scenarios, such as virtual worlds, by taking advantage of auditory-visual crossmodal perception phenomena.

1.4 Aims and Objectives of the Research

In this section the main aims of the thesis will be summarised. The primary goal of this research is to gain a better understanding of the crossmodal interactions between the auditory and visual sensory modalities and investigate whether such interactions could lead to a new generation of perceptually-adaptive graphics techniques, that would take into account not only the visual stimuli, but also the auditory background of a 3D scene in order to reduce the computational or transmission load, without compromising the user's perception of experienced visual fidelity.

The following are the objectives of this thesis:

- Investigate whether auditory stimuli combined with various dimensions of visual quality degradation, including diminishment of spatial and temporal resolution, such as diminished rendering quality and reduced frame rate, individually and in combination, can influence impressions of the quality of a scene. More specifically:

  - Investigate the influence of sound on the human experience in time passing and perception of temporal rate during the display of 3D animated scenarios (effect of sound in the temporal domain of 3D graphics), i.e whether sound can serve as a 'driving' distractor of the presentation rate of the visual information.

  - Investigate whether auditory stimuli affect a viewer's perceived quality of

rendered images while watching computer generated walkthroughs (effect of sound in the 'spatial' domain of rendered graphics).

– Investigate the influence auditory stimuli, such as music and sound effects, have on the perception of frame rate and more specifically on the perceived smoothness of rendered animations (influence of audio in the temporal domain of rendered graphics). That is, examine whether viewers, in the presence of audio stimuli, fail to notice variations in the motion smoothness between walkthrough animations displayed at different rates, which are apparent in the absence of sound.

- Compare the effect of audio on the perception of visual quality parameters (rendering quality and frame rate) in passive animated scenarios and in scenarios involving visual and/or memory tasks.

- Investigate whether sound-emitting objects attract visual attention and result in gaze shifts in 3D scenes.

- Explore the selective rendering of a 3D scene by assigning high resolution to the sound sources, which may potentially capture visual attention and become the centre of gaze, and lowering the quality elsewhere in the image. The finding that attentive vision is a serial, resource-constrained process suggests that it is likely for the perceptual losses to be minimised when the viewer is attending to portions of an image that are not degraded. Will the observer, under the influence of auditory stimuli, notice the difference in quality between various parts of the image? If not, selective rendering could greatly reduce computation time.

- Examine whether familiarity/unfamiliarity with computer graphics concepts, such as rendering quality, aliasing, motion artifacts etc., affect a viewer's ability to notice artifacts resulting from reduced frame rates or degraded rendering quality.

- Examine whether habituation to an auditory stimulus reduces its 'distractive' ability.

- Future work - Because this thesis is one of the first to introduce concepts of perceptually-adaptive graphics based on auditory-visual crossmodal interactions, it should lay the groundwork and direction for future research in this area.

1.5 Thesis Outline

The remainder of this thesis is organised as follows:

Chapters 2-3 discuss relevant background material, including research findings on Audition, Vision, Perception and Attention. The intent of these chapters is to give the computer scientist a high-level overview of some of the basic background knowledge which is required in order to understand this multi-disciplinary research. Because of the wide variety of topics covered, the reader will hopefully gain a better appreciation for the interdisciplinary nature and breadth of knowledge required for our research effort.

Chapter 2 describes basic perceptual abilities of the auditory system, with an emphasis on spatial hearing and provides a detailed discussion of auditory psychophysics related to auditory displays. It then moves on to multimodal environments and discusses the relevant theories regarding crossmodal auditory-visual interactions in such environments, as well as auditory-visual phenomena of intersensory bias, where input to one modality can influence perception in another. It only covers phenomena that take place in the temporal domain, where audition is the dominant' sense, since these phenomena are most relevant to our research. Once this groundwork has been completed, design considerations for auditory interfaces in multimodal worlds and types of auditory displays are covered. On the background material included in this chapter we based the 'temporal' aspect of our research, discussed in Chapters 4-5.

Chapter 3 presents background information on Attention, Visual Perception and Com-

puter Graphics. More specifically, it discusses the effects of attention on perceptual processing, the selectivity of attention in vision and relevant existing theories, the effects of perceptual and cognitive load on attentional allocation, the phenomena of attentional capture by auditory and visual distractors, the relation between shifts of attention and eye movements.

The chapter continues with a discussion of visual perception principles relevant to 'realistic' computer graphics, followed by an overview of research findings regarding perceptually-based metrics employed in the assessment of visual quality of static images, animations and interactive computer graphics scenes, such as virtual reality worlds and realistic 3D simulations. Auditory-visual interactions in assessing complimentary audiovisual quality are also discussed. The chapter concludes with a presentation of the existing perceptually-based rendering techniques, with a focus on the interactive rendering of 3D scenes. On the research findings presented in this chapter we based the 'spatial' aspect of our research efforts, presented in detail in Chapter 6.

The research presented in chapters 2-3 brought about my interest in this research topic and helped me gain a better understanding of the methodology to be employed.

The next chapters, Chapters 4-6, describe the experiments that were conducted and an analysis and discussion of the results obtained follows.

Finally, Chapter 7 concludes the thesis with the work accomplished and future research defined. More specifically, this chapter discusses achievement of goals specified in this introductory chapter of the thesis, enumerates contributions of the thesis to the field and discusses ideas for future research on this topic.

Appendix A presents the questionnaire given to the participants of the experiment regarding "the Perception of Temporal Rate and Duration".

Appendix B depicts the resulting scan paths of the Eye tracking Experiment participants,

corresponding to the period that the sound effect was audible, given separately for each participant of the four experimental conditions.

Chapter 2

Psychoacoustics and Computer Graphics

2.1 Introduction

In the chapter we will discuss basic perceptual abilities and characteristics of the auditory system, with an emphasis on spatial hearing, the importance of rendering auditory cues in computer graphics environments and the existing types of auralisation configurations employed by sound engineers.

More specifically, in section 2.2, after a very brief introduction to psychoacoustics, we will present the basic principles according to which humans perceptually group sounds as coming from the same or different sources and will analyse the cues humans utilise in order to locate the origin of a sound, both its direction and its distance from a listener. The section will also cover well documented research findings regarding intersensory auditory-visual interactions on spatial sound perception. These findings were very crucial for our experimental work, since we needed to manipulate the perceived location of the origin of a sound without going into rendering proper 3D sound cues, which would be unnecessarily complicated and outside the scope of our work.

Once this groundwork has been completed, we will move to multimodal environments and discuss the relevant theories which have been proposed regarding the interactions between the various senses which take place in such environments (section 2.3). The examination of audio-visual interactions is critical because auditory and visual channels 'cooperate' on both cognitive and sensory levels of human perception [265].

We will also describe intersensory auditory-visual phenomena and more specifically we will portray instances of intersensory bias, where input to one sensory system can influence another. The discussion here will be restricted to phenomena that take place not in the spatial but in the temporal domain and where audition is the 'dominant' sense, since these phenomena are most relevant to the 'temporal' path our research. Findings regarding other auditory-visual interactions which take place in the spatial domain and have to do with crossmodal attentional allocation and visual attention orienting, attentional capture and auditory-visual interactions in the perception of image quality are presented in Chapter 3 (in section 3.6 and its subsections and also in 3.8.2), since these are more directly linked to the 'spatial' aspect of our research. We decided to separate the presentation of the crossmodal auditory-visual phenomena into two chapters, in order to highlight better the two different paths of our experimentation, the temporal and the spatial.

Next, section 2.4 will highlight the importance of spatialised auditory interfaces for virtual multisensory environments and will also briefly discuss design considerations for such auditory interfaces. Finally, section 2.5 will cover the differences in auditory simulations using headphones and speakers, in order for the reader o understand the sound design decisions we made for our experimental setups.

It is impossible to include everything that needs to be known about three dimensional sound and designing auditory interfaces in a single chapter. Instead of trying to review all perceptual and technical issues relevant to creating 3D auditory displays, this chapter unapologetically focusses on issues of spatial auditory perception and the generation of spatial auditory cues, since this area has undergone rapid development with the advent of

virtual environments and high-fidelity 3D simulations in general. Other important aspects of auditory perception are ignored or given relatively little consideration. Similarly, techniques for the actual rendering of spatialised auditory stimuli are outside the scope of this thesis and are not discussed in this chapter.

2.2 The Psychology of Hearing

In the past, research in auditory perception has concentrated on low-level factors, such as thresholds for pitch and loudness. Recently, however, the importance of higher-level, cognitive factors has become increasingly evident, and there is growing recognition that the auditory system of the brain contains some remarkably ingenious circuitry, perhaps the most ingenious of all the sensory modalities [45]. The *Psychology of Hearing* examines the physiology and psychophysics of hearing, in order to understand how the auditory system processes information to create perceptions of acoustic events. *Psychoacoustics* is the term applied to the contribution of the mental aspects of sound interpretation.

The auditory interface designer must be aware of the effects of psychoacoustics when designing sounds for the interface. As Frysinger [65] notes: "The characterisation of human hearing is essential to auditory data representation because it defines the limits within which auditory display designs must operate if they are to be effective".

2.2.1 Perceptual grouping principles

In everyday life, we are continuously bombarded with mixtures of sounds emanating from many different sources. A major task for our auditory system is to sort out the components of such mixtures so as to reconstruct the original sound events. Somehow it groups those components that have come from the same source and separates out those that have emanated from different sources.

Early in the century, Wertheimer [253], cited in [45], argued that we link elements of perceptual arrays according to a number of simple principles, such as:

- **Proximity,** according to which we form connections between elements that are closer together, either in time or pitch, rather than between those that are distant from each other along such dimensions.
- **Similarity,** according to which we combine elements that are similar to each other in some way.

The human perceptual system has probably evolved to form groupings in accordance with such principles because they enable us to interpret our environment most effectively [45].

2.2.2 The Spatial Attributes of Sound

A key element of the human ability to combine the streams of sound they receive into auditory objects and scenes, is the ability to localise sound sources. Auditory localisation, is a complex phenomenon affected by physiology, expectation, and even the visual interface. Most of the research papers on spatial hearing have dealt with the localisation of sound sources' direction, with only a few studies addressing the perception of distance [114].

While some spatial tasks can be accomplished solely on the basis of direction, for example localising the direction of a breaking car so that the head may be turned to locate it, distance is often crucial- for example, perceiving whether the car is so close as to require an immediate action. The emerging technology of virtual acoustics is generating increased interest in auditory distance perception, not only because of those tasks for which distance perception is essential, but also because in the entertainment industry, e.g. computer games, the impression of sounds varying in distance has been found to add immensely to the aesthetic impact [114].

This section provides a brief review of how acoustic attributes convey spatial information

to a listener and how the perceived position and distance of a sound source is computed in the brain.

Although much is still unknown about how we localise sounds, it has been discovered that the following physical cues play a major role: binaural cues, pinna response, shoulder echo, head motion, reverberation, and intensensory interactions (especially between vision and audition). Other cues include atmospheric absorption, bone conduction, and a listener's prior knowledge of the sound source [18].

Sound Direction Localisation

Sound transmitted from a source to the listener's head travels along direct and indirect paths. When the sound arrives at the listener, it is modified by the head, shoulders, and pinnae (the visible outer structures of the ears). One of the key differences between vision and audition is that sounds can be heard as coming from all around the listener. The stimulus cues for localising sound in direction, azimuth and elevation, are now well understood and they include the following:

Binaural Cues Lord Rayleigh [J. W. Strutt] (1907) [115], cited in [100], introduced one of the earliest theories of human sound localisation, the *Duplex Theory*, which explains audio localisation as a function of differences in intensity and arrival time between sounds reaching the ears. *Interaural intensity difference* (IID) refers to the difference in intensity of a sound heard by each ear, see Figure 2.1. This cue is generally considered ineffective for frequencies below 1500 Hz, as at these low frequencies sound waves wrap around the head and intensity differences are minimised, practically eliminated. At frequencies above 3000 Hz, intensity differences are significant enough to act as cues for a sound source's position.

The difference in arrival time of sound waves reaching each ear is designated as the *interaural time difference* (ITD), see Figure 2.1. Unlike the IID, this cue is effective for

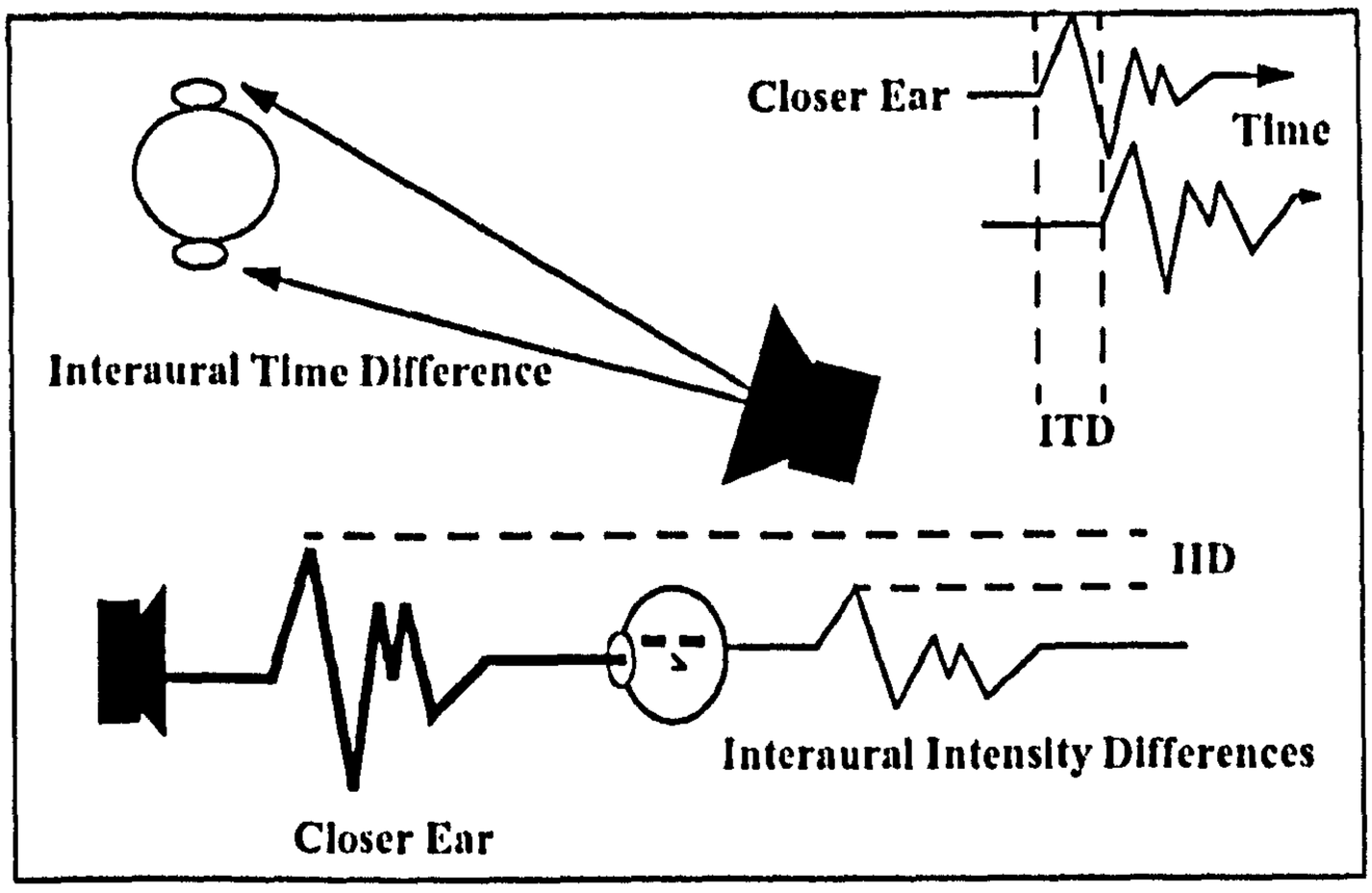


Figure 2.1: The primary binaural cues of sound direction localisation. Image reproduced from [18].

low frequency signals. Identical IID and ITD cues can be generated for multiple points in space and therefore it is necessary for humans to rotate their heads in order to accurately localise the origin of a sound.

If a sound source is located to one side of the head, then the sound will reach the further ear later and its intensity (IID) will be reduced compared to the other ear [13]. These two factors are key in allowing a listener to localise a sound in an environment. Humans are able to discriminate between the positions of sound sources which are very proximal in space. The minimum auditory angle (MAA) is the smallest separation between two sources that can be detected. Strybel, Manligas and Perrott (1992) [213] reported that in the median plane sound sources only 1^o apart can be detected, while at 90^o azimuth (directly opposite one ear) sources must be around 40^o apart to be reliably separated.

Another important aspect of human audio perception is exhibited when the same sounds (i.e. identical soundwaves of the same intensity) emanate simultaneously from two sources . The sound is perceived as coming from a location in-between the two sources. However,

if the sound from the right source is delayed (relatively to sound from the other source) by a time interval between 1 and 30 milliseconds, then the sound will be perceived as coming from the left source only. This psychoacoustic phenomenon is known as the *Haas* or *precedence effect* or *law of the first wave front.*

Pinnae Response The outer ear, or pinnae, also plays a crucial role in localising sound. The term "pinnae response" describes the shape of the ears and their role in externalising sounds. When sound arrives at the pinnae, its frequency characteristics are modified. These modifications vary depending on the position of the sound origin, thus providing another important directional cue, which helps compensate for the shortcomings of the ITD and IID cues.

The IID, ITD, and pinnae localisation cues can be mathematically expressed by the *head related transfer function* (HRTF), which represents the complex variations in IID and ITD, depending on the frequency of sound and the azimuth and elevation of its source. Listeners are most accurate in localising the azimuth and elevation of a source when they listen to sounds with their own HRTFs [114]. 3D sound systems apply HRTFs to digital audio files to generate the illusion of a sound originating from a point in 3D space.

Shoulder Echo also contributes to sound localisation. Echoed sound waves are reflected off a listener's shoulders and come in contact with the ears at different angles/times, compared to the sound waves travelling directly from the sound source to the ears. Other echoes are present as well. Any object that reflects sound produces an echo which also strikes both ears. The difference in arrival times and intensities of these echoes contribute to sound localisation [18]. Figure 2.2 shows examples of different echo sources.

Head Motion. Although front/back and up/down confusions are fairly common when a listener's head is stationary, these confusions are practically eliminated when the sound lasts long enough for the listener to derive additional information by moving his/her head.

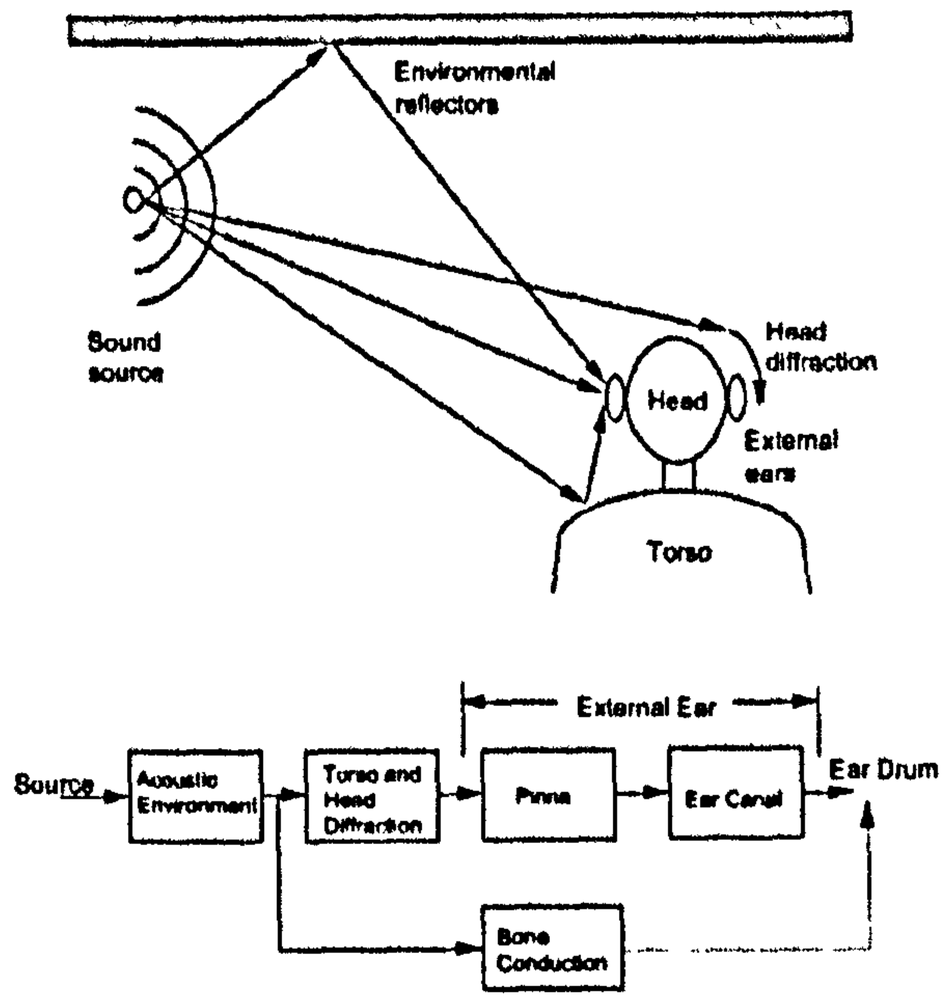


Figure 2.2: Acoustic Paths. Image courtesy of Prof. R.O. Duda.

Humans have the natural tendency to orient their head toward the perceived direction of a sound. As the head rotates, the localisation cues shift as well. The analysis of directional localisation during head motion reveals that the perceived sound direction depends on the sensed change in lateral angle relatively to the sensed rotation of the head [114].

Perception of the Distance of a Sound Source

The intensity of a sound reaching directly a listener decreases with the square of the distance between the listener and the sound source. As a sound travels in space, energy in high, audible frequencies is absorbed by the atmosphere, causing small changes in the spectrum of the received signal, depending on the source distance. Therefore, if a source is unfamiliar, the intensity and spectrum of the direct sound reaching a listener are not robust cues for distance, since they may be different from the corresponding cues of the signal emitted from the source. Sound level serves as an absolute distance cue if the observer has

independent knowledge about the source intensity; otherwise, it gives information only about the changing distance of a source [114]. Mershon and King (1975) and Zahorik (1998), among others, conducted experiments which demonstrated that sound level does act as a relative distance cue and influences the observer's judgement of distance over multiple presentations of the same source [114].

A cue that helps a listener determine absolute distance, even on the first presentation of a stimulus, is *reverberation*, which refers to the acoustic energy reaching the listener via indirect paths, such as walls, floors, etc., see for example [12]. At least grossly, the intensity of reflected energy arriving at the ears does not depend on the position of the sound source relative to the listener, but can vary dramatically from one room to another. Hence ,for a given environment, an absolute measure of source distance can be expressed as a function of the relative amount of reverberant or reflected sound to the amount of direct sound received by the ear [12]. This quantity is expressed as the *reverberant/direct* (R/D) ratio.

Reverberation is not only a robust cue for source distance, but it also gives information regarding the size, 'spaciousness' and configuration of the listening environment [12]. Despite the fact that reverberation is present, in varying degrees, in virtually all normal listening conditions, many psychophysical studies of sound localisation are performed in anechoic, or simulated anechoic, environments, which seem subjectively 'unnatural' and 'strange' to naive listeners.

2.2.3 Intersensory effects on Spatial Sound Perception

Information from modalities other than audition can have a pronounced effect on spatial perception of auditory and multisensory events. In particular, auditory spatial information is combined with visual and/or proprioceptive spatial information to form the percept of a single, multisensory event, especially when the inputs to the different modalities are

correlated in time [252]. When this occurs, visual spatial information is much more dominant than that of auditory information, so that the perceived location of the multisensory event is primarily determined by the visual spatial information [174] ("visual capture").

Perhaps the most familiar multisensory phenomenon of visual spatial capture is the *Ventriloquism Effect* [14, 16, 173], which creates the illusion during the synchronous presentation of auditory and visual events in somewhat separate locations, that the location of the sound is shifted towards the location of the visual stimulus. A standard explanation of ventriloquism is that when auditory and visual stimuli occur in close temporal and spatial proximity, the perceptual system assumes that they represent a single event. The perceptual system then tries to reduce the conflict between the location of the visual and auditory data because there is an a priori constraint that an object or event can have only one location [11]. Shifting the auditory location in the direction of the visual event rather than the other way around would seem to be more 'natural', because spatial resolution in the visual modality is better than in the auditory one.

A relevant phenomenon is the *spatial magnetisation of sound*, i.e. the mental relocation of the sound emitter to an object on screen [35]. We regularly experience this effect when watching television and movies, where the voices seem to emanate from the actors' lips rather than from the auditory display devices. According to Chion [35], this is due to the 'audiovisual contract'- the agreement to accept the pseudosound emitters on screen as real:

"the audiovisual relationship is not natural but rather a sort of symbolic pact to which the audio-spectator agrees to forget that sound is coming from loudspeakers and picture from the screen. The audio-spectator considers the elements of sound and image to be participating in the same entity or world. The result of the audio-visual contract is that one perception influences the other and transforms it so that we never see the same thing when we also hear and we don't hear the same thing when we see as well" [35].

Vroomen et al. (1998) investigated whether it is possible to train subjects to ignore the visual 'distractor' [234]. Subjects were trained to discriminate among sequences of tones that emanated either from a central location or from alternating locations, in which case two speakers located next to a computer screen emitted the tones. Instructions and feedback provided during each trial could not overcome the effect of the visual distractor on sound localisation, which indicates that the ventriloquism effect is indeed very robust.

Such an effect does not have to do with the subjects' cognitive strategies. When they are asked to point to the location of auditory stimuli while ignoring spatially discrepant visual distractors, subjects might be aware of the spatial discrepancy and adjust their response accordingly. In this case the visual bias would reflect post-perceptual decisions rather than genuine perceptual effects. Nevertheless, several researchers have argued that the ventriloquism phenomenon reflects an automatic perceptual process [14, 15, 28, 46] rather than a conscious response strategy.

A relevant question is whether the ventriloquism effect is influenced by attention. It was shown that the influence of ventriloquism is not altered by where *endogenous* visual attention is focussed [17]. When subjects had to localise auditory targets and ignore bright visual distractors presented synchronously to the left or the right of the sound sources, it did not matter whether they were focussing on the distractor being at the periphery rather than at a central location. Equal amounts of ventriloquism were demonstrated in the two cases. In another set of experiments, it was also shown that ventriloquism is not influenced by whether or not the visual distractor receives *exogenous* visual attention [235].

We exploited the Ventriloquism Effect in our experiments, described in later chapters, in order to avoid computationally expensive spatial 3D sound rendering.

2.3 Moving to multimodality: More Auditory-Visual Sensory Interactions

"It is a commonly observed fact that most objects of our everyday lives are perceived by means of two or more sensory modalities working in cooperation".

Ryan (1940) [185]

Psychophysics of vision, sound, and touch will change when the environment is multimodal. Humans monitor the environment through different sensory channels, which receive correlated input about the same external object or event. Our perceptual system then combines this input to yield a multimodally determined percept. For example, seeing a speaker not only provides auditory information about what is said, but also visual information about movements of the lips, face, and body, as well as visual cues about the origin of the sound.

Furthermore, the different sensory modalities interact even at the early stages of neural information processing in the brain stem. Stein and Meredith suggest that this is a fundamental basis for perception:

"...the world is not perceived as a series of independent sensory experiences in which the integrity of each modality's 'snapshot' view is preserved intact in its own location in the brain; rather, there is an interweaving of different sensory impressions through which sensory components are subtly altered by, and integrated with, one another. The product of these integrative processes is perception." [209]

Stein et al. [209] also investigated visual-auditory interactions in perceived intensity and they reported evidence of crossmodal enhancement. They suggested that "combinations of, for example, visual and auditory cues can enhance one another and can also eliminate any ambiguity that might occur when cues from only one modality are available" [209].

All these results indicate that multisensory cues may play an important role in perception. Therefore, it is important to retain these natural relationships when developing systems for multisensory perception or display. In practice, though, comparatively little attention is paid to this multimodal state of affairs and the different senses (e.g., seeing, hearing, touch) are often treated as distinct modules with little or no interaction.

2.3.1 Theories of intersensory interactions

As we mentioned above, from studies in perception we know that observation in one modality, or sensory system, can influence another and even that a particular system can substitute for another system [209]. This phenomenon is known as "intersensory bias".

Visual superiority in situations of intersensory conflict has inspired some theoretical explanations about how crossmodal discrepancies are perceptually resolved, for example [168, 182]. According to the early theories, when there are no great differences in the intensities of the stimuli, the effect of the visual stimuli on the stimuli presented in other modalities is greater than vice versa [209]. Consistent with this, Posner et al. (1976) [168] proposed the *Visual Dominance* account, according to which humans have the tendency to attend more to visual information than to other modalities, see also [37], in order to compensate for the "low alerting capability of visual signals"; hence, the result is that visual input tends to dominate in intersensory conflict situations. Otherwise, visual input would be disadvantaged by a lag which is greater for the visual modality than for other modalities.

While the visual dominance account may be consistent with most findings in intersensory spatial biases, such as ventriloquism, it overlooks many other cases of crossmodal interactions in which visual perception is systematically biased by sounds or tactile stimulation. For instance, ambiguous visual stimuli allow for auditory (or tactile) information to influence the way in which visual motion is perceived [137, 189, 192]. Many other instances of auditory or even tactile dominance over non-ambiguous visual information have also

been reported outside the domain of motion perception, for example, in the perception of temporal order, see [200] for a review.

Another class of explanations is based on the adequacy and/or the relative precision of each modality involved in a particular situation. One popular explanation of this kind is the *Modality appropriateness* hypothesis, which was defined by Welch and Warren [251, 252]. According to this explanation, when a modality is better suitable for a certain task, it dominates over other modalities. In support for this hypothesis, vision typically dominates in spatial perception tasks in which its accuracy is superior to audition, as in the case of ventriloquism, whereas audition influences vision in temporal tasks, where it is superior to vision [176].

Massaro [126] introduced a 'fuzzy logic' model of perception which encompasses many different instances of crossmodal integration within the same general framework of optimal information combination. One modality plays a more important perceptual role than another in this combination, provided that it reduces the overall uncertainty/ambiquity.

More information about some of the most studied phenomena of auditory-visual cross-modal interaction will be given in the following section. Having already discussed in a previous section how visual stimuli affect the spatial dimensions of a sound, we will analyse how sound affects the visuals in the temporal domain. The findings that we will be reported largely influenced the hypotheses we made for the 'temporal' path of our research.

2.3.2 How Sound affects the Visual Stimuli in the Temporal Domain

As we mentioned above, there are several examples in the literature where sound captures vision, particularly in the temporal domain.

The most well-known phenomenon of auditory dominance over vision, is what is termed

the **Auditory Driving Effect** [68, 140, 194, 250]. Auditory driving represents a case of the auditory temporal capture of vision. If observers are asked to judge the rate at which a light is flickering when that light is presented simultaneously with a repeating ("fluttering") sound, increasing or decreasing the flutter rate can cause the apparent flicker rate to increase or decrease respectively, although the flicker rate does not change. Gebhard and Mowbray (1959) [68] found no indication of the reverse phenomenon- that is, varying flicker rates did not change the perception of concurrent flutter. Shipley (1964) [194] attempted to determine the capture range of auditory driving by varying flicker rates in the presence of fixed flutter until observers reported that they were clearly different. He found that a 10-Hz flutter could perceptually shift flicker rates from 7 Hz up to 22 Hz. Welch, Duttonhurt, and Warren (1986) [250] attempted to measure the strength of auditory driving using magnitude estimation. When flicker and flutter rates were discrepant, reported flicker rates shifted toward flutter rates so as to eliminate an average of 52% of the discrepancy. Flutter rates also shifted toward flicker rates but to a much smaller extent, eliminating an average of 13% of flicker-flutter discrepancy. Recanzone (2003) [176] replicated previous findings about auditory driving and concluded that this phenomenon provides support for the *Modality Appropriateness* hypothesis. It has been argued that auditory driving is very likely to depend on low level sensory processes, as there is no evident reason for observers to assume that the flashing light and the fluttering sound occur at similar rates.

It should be pointed out that auditory driving does not necessarily imply that the auditory flutter breaks a single flash into two or more flashes resulting in a perceived higher flicker frequency. An alternative and perhaps more plausible explanation for this phenomenon is that the perceived duration of each flash or the gap between two successive flashes is altered by accompanying flutter. Indeed, such alteration of duration and gap of flashes by sound has been shown in other studies [240].

Repp and Penel (2002) [179] also reported a case of *temporal ventriloquism* during a sen-

sorimotor task. Their participants were asked to synchronise finger taps with a sequence of audiovisual events which were presented at a rate of 2 Hz. They observed that, despite their instructions to the subjects to synchronise with the visual stimuli, the variability of the finger-tap asynchronies was controlled by the simultaneously presented auditory tones.

Morein-Zamir et al. (2003) [139] observed that sounds presented before and after two flashes improved visual temporal-order judgements, whereas sounds intervening between the two lights impeded performance, as if the sounds attracted the lights in the temporal dimension.

Vroomen and de Gelder (2004) [237] showed that a single sound presented in close temporal proximity to a flash attracts the temporal dimensions of the flash. Their results can be summarised as follows: When a sound was presented simultaneously with a flash, it "sharpened the temporal boundaries of the flash" and made the flash appear earlier (approx. 5 ms). Moreover, when the sound was presented before or after the flash with their discrepancy being in a range of approx. 100 ms, the sound attracted temporally the flash by approximately 5.2%. These findings demonstrate that not only rhythmic sequences of sounds but also a single sound can affect the perceived temporal dimensions of visual information in a task in which visual space and not time or rate, is the relevant dimension.

Vroomen and de Gelder (2004) [236] described another case where sound affects vision. The basic phenomenon is that when subjects are shown a rapidly changing visual display, an abrupt sound may 'freeze' the display with which the sound is synchronised. Subjects perceived the display as either brighter or shown for a longer time. The results from several experiments they conducted suggest that this 'freezing' phenomenon is a perceptually genuine effect and not an instance of crossmodal attentional cuing.

2.4 Spatialised Sound for Multisensory Environments

Audio is often given minimal attention by developers of virtual environments or simulations, despite the fact that auditory cues play a key role in everyday life; they "increase awareness of surroundings, cue visual attention, and convey a variety of complex information without taxing the visual system" [191].

The importance of multimodal interactions involving the auditory system cannot be ignored. It has been shown that using medium and high quality auditory displays can enhance the perception of quality in visual displays. Inversely, low quality auditory displays result in the reduction of the perceived quality of visual displays [210]. Despite the fact that the importance of sound to create ambience and emotion has been recognised for a long time in the entertainment industry, in the past the developers of computer graphics environments, such as VR worlds and simulations, rarely incorporate these aspects in their work. Regardless of considerable evidence on its immersive potential, proper 3D audio was often excluded from virtual reality applications, partly due to technical resource limitations of computer systems. Since the visuals were given the highest priority, audio quality was sacrificed for graphics performance. However, these restrictions no longer exist and it is now possible to implement high-fidelity, spatialised audio in visually complex 3D graphics applications, without impairing performance or interactivity [191].

Nowadays, in the computer graphics community, spatialised sound is most commonly found in immersive virtual environments. Such systems allow a user to 'explore' a virtual world and/or interact with other users by rendering images and sounds of the environment in real-time while the user 'moves' around the virtual space. Example virtual environment systems include entertainment, electronic commerce, tele-education, medicine, distributed training, computer-aided etc. design applications.

Spatialised sound effects are important in such applications because, in combination with

visual cues, they help the users locate objects, discriminate between concurrent sound signals, and form spatial impressions of an environment [66]. For instance, binaural auditory cues are essential for the localisation of objects objects outside a user's field of view, such as when an enemy is coming around a blind corner in an adventure game. They also help users separate concurrent audio streams, such as when we follow a conversation among several others at a party. Finally, sound reverberation can enhance and reinforce the visual comprehension of an environment, such as when a 3D game player moves between spaces of different sizes and constructed with different materials (wood, marble, stone etc.). Moreover, experiments have shown that more accurate acoustic modeling results in a stronger sense of presence in virtual environments [66].

A difficult challenge for non-trivial virtual environment systems is to provide accurate (or at least plausible) spatialised sound. Sound waves emanating from a source and arriving at a receiver travel along a very large number of propagation paths representing different sequences of transmissions, reflections and diffractions at the surfaces of the environment, which result in the addition reverberation, for example echoes, to the original source signal as it reaches the receiver. To auralise a sound for a particular source, receiver, and environment, the sound engineer needs to apply one or more filters to the initial audio signal, in order to mimic the acoustical effects of sound propagation through the environment [66].

Figure 2.3 shows a basic processing pipeline for the auralisation of spatialised sound. The input to the system is a description of the geometry of a virtual environment, an audio source location, an audio receiver location, and an input audio signal. The auralisation system computes a model for the propagation of sound waves through the environment and constructs digital filter(s) that encode the delays and attenuations of sound propagating along different paths. Convolution of the input audio signal with the filter(s) results in a spatialised sound signal for output with an auditory display device. The final stage of the auralisation pipeline is to reproduce a 3D sound field for the ears of the listener,

by using appropriate 3D auditory display devices [66]. For virtual environments, it is especially important that the auditory display device at least produces directional sound waves that provide 3D localisation cues. In the next section, we provide an overview of common 3D auditory display techniques, comparing them in terms of setup, directional accuracy, robustness of imaging and complexity.

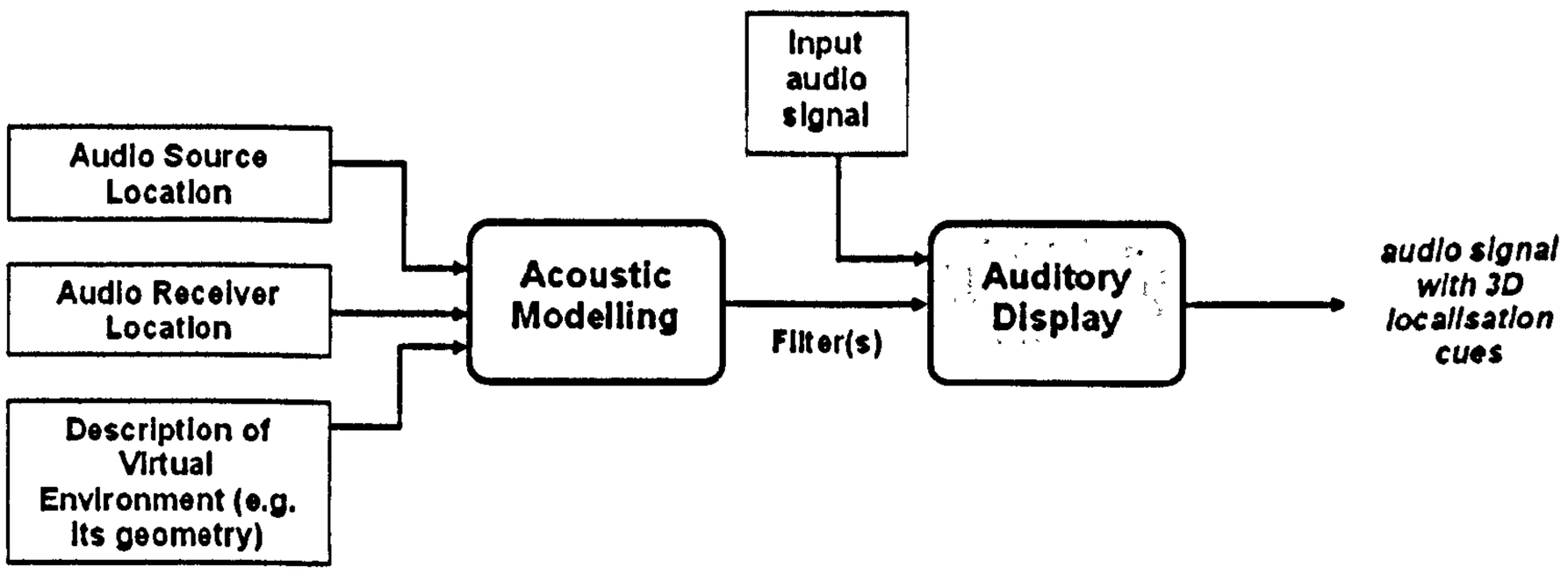


Figure 2.3: Basic auralisation pipeline.

2.5 Auditory Displays of Spatialised Sound

Spatial auditory cues can be simulated using headphone displays or loudspeakers. Headphone displays generally allow more precise control of the spatial cues reaching the listener, both because the signals reaching the two ears can be controlled independently and because there is no indirect sound reaching the listeners (i.e., no echoes or reverberation) [100]. Spatialised audio using headphones is the only audio technique that is truly 'virtual' since it reproduces azimuth, elevation, and distance and offers the sound engineer the greatest amount of control over the auditory experience of the listener.

On the other hand, headphone displays are generally more expensive than loudspeaker configurations and may be impractical in the cases that the listener does not want to wear a device on the head. Moreover, loudspeaker-based simulations are relatively simpler

and less expensive to implement and do not physically interfere with the user. Properly designed speaker systems incorporating subwoofers may contribute to emotional context [191], a benefit you cannot derive from using a headphone display.

Simulations using either headphones or speakers can vary in complexity from providing no spatial information to providing nearly all naturally-occurring spatial cues. The following paragraphs briefly review both headphone- and speaker-based approaches for the creation of spatial auditory cues.

2.5.1 Headphone Simulation

As we mentioned above, headphone-based displays allow the greatest degree of control over the perceived location of a sound source, and are very important to applications where performance of a subject in a specific task is involved. Moreover, they allow the influence of background noise in the listening room to be practically eliminated. In these regards, headphone playback is considered an "optimal" condition for the reproduction of 3D sound.

Nevertheless, the reproduction of 3D sound over headphones can cause one or more of the perceptual errors, i.e. mismatches between the intentions of the sound designer and the resulting percept of the listener, such as localisation blur, reversals, and problems with externalising the stimuli.

A significant problem for the implementation of 3D sound systems is the fact that the spectral features of HRTFs differ between individuals, and therefore localisation errors increase when listening with what are termed "non-individualised HRTFs", particularly in the perception of up/down and front/back position [13]. In addition, HRTFs should be ideally sampled in both distance and direction at a high spatial density dictated by human sensitivity. Most current auditory display systems use generic HRTFs, sampled coarsely

in direction and at only one distance. However, a 'realistic' auditory display system, tied to a particular application which requires the user to extract 3D spatial information from the auditory display, would use HRTFs tailored to the listener to preserve directional information and would probably include reverberation cues in order to encode source distance. Reverberation has been shown to dramatically increase the externalisation of stimuli relative to non-reverberated stimuli, in one case, from 2% to 90% [13].

Another factor frequently cited as a means for improving localisation within a 3D audio headphone-based display system is head motion cues [13]. Listeners somehow combine the changes in ITD, IID and spectral cues resulting from head motion over time, and subsequently use this information to resolve ambiguities regarding the position of a sound source.

On the other hand, if a particular application only makes use of one spatial dimension, a rough simulation of ITD and IID cues, even without detailed HRTF simulation, is probably sufficient [191]. While normal interaural cues vary with frequency in complex ways, simple frequency-independent ITDs and IIDs affect the perceived lateral position of a sound source. Stereo signals that only contain a constant ITD and/or IID are referred to as "dichotic" signals. Generation of a constant ITD or IID is very simple over headphones since it only requires the original signal to be delayed or scaled, respectively, at one ear. Dichotic signals result in sources that appear to be located on an imaginary line inside the head, connecting the two ears. Varying the ITD or IID causes the lateral position of the perceived source to move toward the ear receiving the louder and/or earlier-arriving signal. For this reason, such sources are usually referred to as "lateralised" rather than "localised".

Dichotic headphone displays are simple to implement, but they can only indicate whether a sound source is located to the left or right of the listener. Some binaural unmasking can be obtained when multiple sources are lateralised at different locations (using different ITD and/or IID values).

Although the above techniques may allow localisation improvement, it is important to recognise that the most accurate 3D sound localisation seems to require active, attentive listening, unaffected- where possible- by external distractions from undesired visual, auditory and tactile stimuli. The influence of cognitive cues, memory, and associations must also be a controlled factor.

2.5.2 Simulation Using Speakers

The total acoustic signal reaching a listener's ear is the sum of the signals reaching that ear from each auditory source in an environment. Therefore, it is possible to vary spatial auditory cues by manipulating the signals played from multiple speakers arranged suitably around the listener. In contrast with headphone simulations, the signals reaching the two ears cannot be independently manipulated, as the changes to the signal emitted by any of the speakers affects the signals reaching both ears [191]. As a result, it is difficult to precisely control the interaural differences (ITD and IID) and spectral cues of an auditory signal reaching the listener to mimic the signals corresponding to a real sound source. However, loudspeaker simulations often achieve reasonable results for their cost and additionally there are various methods for specifying the signals played from each loudspeaker in order to simulate spatial auditory cues.

Stereo Display

Stereo displays present signals via two speakers simultaneously, in order to control the perceived laterality of a 'phantom' source of sound [191]. For instance, simply by varying the intensity of otherwise identical signals emitted from a pair of speakers can alter the perceived laterality of a phantom source. Most commercial stereo recordings are based on variations of this approach. This simple technique, referred to as "panning", produces a robust perception of a source at different lateral locations. However, it is nearly impossible to precisely control its exact location, as the perceived direction of a sound source

depends upon the location of the listener with respect to the two loudspeakers. As the listener moves outside a restricted area (the "sweet spot"), the simulation degrades rather dramatically. In addition, reverberation can distort the interaural cues, causing biases in the resulting simulation [191].

Surround Sound

The ability to convey accurate spatialised sounds using loudspeakers increases dramatically as the number of speakers used in the audio configuration increases.

True surround sound formats rely on dedicated speakers that literally and physically surround the audience. Currently, the most common Surround Sound format is the 5.1 speaker system, which includes the following speakers. One central speaker, which carries most of the dialog and part of the soundtrack. Two left and right front speakers which carry most of the soundtrack (music and sound effects) and may carry parts of the dialog as well. There is also a pair of surround sound speakers that is placed to the side (and slightly above) of the audience to provide the surround sound and ambient effects. Finally, a subwoofer, the so-called ".1" speaker, can be used to reproduce the low and very low frequency effects (LFE). The ".1" signifies that the sixth channel is not full frequency, but contains only deep bass frequencies (3 Hz to 120 Hz).

Typical 5.1 Surround sound formats are Dolby Digital Surround and Digital Theater Systems (DTS). Dolby Digital is the successor to Dolby Surround Pro-Logic which emerged in home theatre systems in the early 1990's and became the surround sound standard for Hi-Fi VHS. The Dolby Digital surround sound format provides up to five independent channels of full frequency (from 20 Hz to 20 KHz) and an optional sixth channel dedicated to low frequencies (LFE). This sixth channel is commonly reserved for the subwoofer speaker. An alternative format to Dolby Digital is DTS Digital Surround (DTS), which retains much of the initial audio signal information intact by applying less compression to the higher frequencies and delivers 6-channel audio and 24-bit audio support.

Newer surround sound formats include the Dolby Digital Surround EX (Extended Surround) which is a 6.1 speaker system. It adds a central speaker behind the listener, see Figure 2.4 (right). This allows certain soundtrack effects to be presented behind the audience, thereby achieving more enveloping and complete surround sound. While the Extended Surround sound format requires one surround back channel, two surround back speakers are generally recommended for better envelopment. Acknowledging this widely accepted industry position, some audio receiver manufacturers have introduced "7.1-channel" capable receivers, see Figure 2.4 (left), which include decoding and sometimes amplification features for the two extra surround back channels.

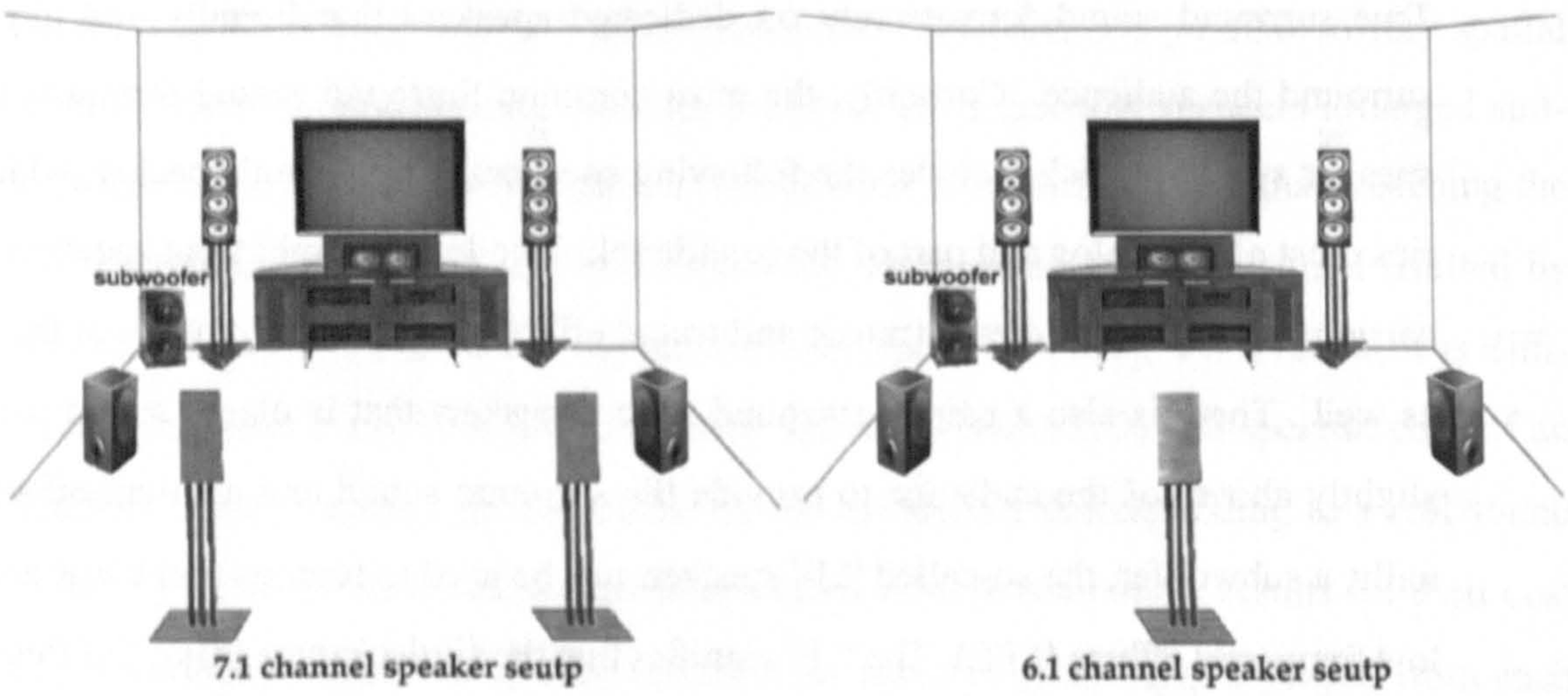


Figure 2.4: Illustrations of a 7.1 (left image) and a 6.1 (right image) Surround Sound System.

2.6 Summary

This chapter has given an introduction to the study of the spatial perception of sound, including crossmodal interactions on spatial hearing. More specifically, after a very brief introduction to psychoacoustics, we presented the basic principles according to which humans perceptually group sounds as coming from the same or different sources and analysed the cues humans utilise in order to locate the origin of a sound, both its direction

and its distance from a listener. In our discussion, we also described cases where visual stimuli seem to affect the human perception of the sound source position. Of particular interest to our research are cases of visual 'dominance' over discrepant auditory cues, for example the *ventriloquism effect* [36]. The relevant findings, presented in section 2.2.3, were exploited during the design and implementation of our experiments, in order to avoid the computationally expensive rendering of proper 3D sound.

Next, we moved to multimodality and gave an overview of some of the most well known research findings in intersensory auditory-visual phenomena and also presented the most important theories which have been proposed by researchers in order to explain these phenomena. During the presentation of the most well documented phenomena of auditory-visual interaction, we focussed on instances of intersensory bias, where input to one sensory system can influence another. The findings from the studies described attest that multisensory cues play an important role in perception.

Phenomena where sound captures vision, particularly in the temporal domain, such as the *auditory driving* effect, were also portrayed. These effects inspired our experiments on the influence of audio on the perception of the temporal characteristics of rendered animations, presented in Chapters 4 and 5.

Section 2.4 provided the reasoning for the need of spatialised auditory interfaces for virtual multisensory environments and also discussed design considerations for such auditory interfaces. Finally, section 2.5 discussed the differences in auditory simulations using headphones and speakers.

Chapter 3

Perception, Attention and Computer Graphics

The human information processing system is a limited resource system [4]. In terms of information, our environment contains a great deal more than we could possibly hope to process. Nonetheless, we seem to manage, despite the limits of our cognitive resources. To deal with the constant barrage of information coming in through our sensory systems, our brain parses incoming stimuli into pieces, then tackles each portion one by one. A key factor in this process is our ability to deploy our attention to likely areas of important activity.

The role of visual attention is a very intensely studied topic (e.g., [172, 188]), namely because vision informs about the spatial layout of objects and scenes in a way that is valuable to identifying things in the world. Although vision maybe the 'primary' sense for recovering the spatial layout of objects, there is no justification to conceive of this sense as the exclusive system for identifying objects. Perceivers can scrutinise, recognise and perhaps identify objects on the basis of a non-visual sensory system such as audition (e.g., [76, 132, 258] or haptic perception.

Many researchers ([63, 98, 221, 268, 269, 270] etc.) explored extensively the distinction

between stimulus and cognitive control over the orienting of attention [97].

In this chapter we will first examine the phenomenon we call attention, try to understand how the limits of human information processing are related to it and look at the theoretical attempts to explain it, including the evidence upon which they are based. In general, with the help of pre-attentive processes, we decide what to pay attention to and what to filter out and ignore. Attention filters and feeds information about the world around us into our minds. This filtering implies that objects or features that are outside the scope of attention are not noticed by the viewer. We will continue with the role of emotions in cognitive processing and crossmodal auditory-visual interactions on attention, together with their practical implications on this research.

Related interdisciplinary research in the fields of computer graphics and visual perception will also be presented. We will next discuss the latest advances in *Perceptually-adaptive* graphics techniques for interactive realistic rendering. Relevant research incorporates principles of human visual perception to improve computer graphics rendering and develops perceptually-based metrics that assess the quality of computer graphics scenes.

3.1 Definitions of Attention

Although systematic research on attention has only been carried out over the past fifty years or so, William James gave the first definition of the concept over a hundred years ago:

> *"Everyone 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."* [95, pages 403-404]

Much later, Ellis and Hunt (1989) state that attention is the "process of focusing selectively on some part of the environment while ignoring other aspects" [56, page 301].

This definition is a starting point from which to consider a broad class of phenomena. For example, attention, by this definition, may be conscious or unconscious (e.g., not paying attention to the person talking to you, when something interesting is happening in your field of vision). It can be focussed on external stimuli or on internal events, for example, thinking about someone who is not present. Possibly most importantly, attention is defined as a process, that is, a dynamic phenomenon in which cognitive resources are constantly expended. Finally, implicit in this definition is the idea that attentional resources are limited, and thus some amount of filtering or selection must take place in the process of attending to stimuli.

The concept of **selective attention** may thus be clarified using Woods' (1990) definition as "the preferential detection, identification, and recognition of selected stimuli in an environment containing multiple sources of stimulation" [263]. In this way, selective attention will be differentiated from attention by saying that selective attention involves a possibly temporary predisposition towards attending to some stimuli while ignoring others which would, under usual circumstances, be equally, or more noticeable.

Biased attention is defined as occurring when selective attention towards a particular type of stimulus lasts for a prolonged period. Biased attention may involve systematically attending to one type of stimulus at the expense of attending to another stimulus which is more relevant to doing a particular task.

The above definitions are still somewhat unclear but, still, the notions of selective attention and biased attention as directed processes which use cognitive resources and which have some interplay with both perception and interpretation, will be further analysed in following sections, as they are two of the cornerstones of our research.

3.2 Perception Depends on Attention

As we mentioned above, the concept of attention rests on the idea that humans have limited mental resources. There is simply too much information in the world for humans to handle, and we need some way of limiting the amount of information that we must process at any time.

We have two ways of dealing with this problem. One is attention. Attention is the cognitive mechanism by which salient or behaviourally relevant sensory information is selected for perception and awareness [42]. Our second trick for conserving mental resources is *automaticity*. We learn some tasks so well, that we can perform them, apparently, without attention. For example, when learning a new piece, a pianist might have to think about every note. With practice, the pianist can play the piece while having a conversation, thinking about other things, etc. It is as if playing becomes 'automatic' and requires no mental resource [187, 229].

Attention acts as a filter which only allows important information through and causes us to ignore the rest. Posner, Snyder, and Davidson (1980) [169] suggested that once oriented, visual attention serves to enhance perceptual processing at the attended location in a manner analogous to the illuminating aspects of a spotlight, see also [59]. Attention also facilitates visual information processing by increasing sensitivity and resolution [77, 81]. This way attended visual events are perceived more rapidly and accurately than are ignored ones. In addition, evidence from functional neuroimaging and from neurophysiology has revealed that neural activity is greater and neural spiking more synchronous in sensory cortical regions responding to attended visual events than to ignored ones, for example [42].

We are only very dimly, if at all, aware of information which falls outside of attention. The image of an object may fall on the eye, yet we may not see it if we have no attentional

capacity available to process it. We can 'see' an object, its highly visible image falls on the retina, yet not perceive it - we are not aware that it is there. Studies have shown that under some circumstances, people will not consciously perceive an object even if it is plainly visible [122]. As noted, this is a common occurrence - in many accidents, the driver claims that she never saw the pedestrian, bicycle or other car- an example of inattentional blindness. The above findings have proved to be invaluable to the 'spatial' aspect of our research (refer to Chapter 6).

We generally attend where we are looking, but attention is shared among stimuli from the various sensory modalities and cognition (thinking and memory). That is, focussing attention on auditory input (e.g., a cellular telephone conversation) can impair the detection of important visual events while, for example, driving a car [211, 212]). Similarly, thinking also uses up mental resources, so a driver lost in his thoughts will also have less attention available for seeing. According to Strayer et al. (2003) [211], attention to one sensory modality can impair the perception of otherwise salient events in another. They found that when attention must be directed to audition, the strength of early cortical representations in the visual system are compromised (and vice- versa), leading to potentially significant behavioural impairments. We exploited this evidence in the 'temporal' path of our work discussed in Chapters 4 and 5.

3.2.1 Inattentional Blindness

The term "inattentional blindness" was introduced in 1998 by psychologists Mack and Rock, when they published the book "Inattentional Blindness" [122], describing a series of experiments on the phenomenon.

In their standard experimental procedure, they briefly presented a small cross on a computer screen for each of several experimental trials and asked participants to judge which arm of the cross was longer. After several trials, an unexpected object, such as a brightly

coloured rectangle, appeared on the screen along with the cross. Mack and Rock reported that participants, busy paying attention to the cross, often failed to notice the unexpected object, even when it had appeared in the centre of their visual field. When participants' attention was not diverted by the cross, they easily noticed such objects [122].

Following these initial findings, Mack and Rock found that objects or events that are personally meaningful are most likely to capture people's attention [122]. Finally, the two researchers discovered that even though participants did not detect the presence of unattended words that were presented on a computer screen, such stimuli nonetheless exerted an implicit influence on participants' later performance on a word-completion task [122]. They concluded that "there is no conscious perception without attention" [122].

Mack and Rock's findings soon attracted other researchers' interest, and research on inattentional blindness has proliferated quickly. In 1999, Simons and Chabris extended Mack and Rock's results using a *selective looking* procedure introduced in the 1970s by Neisser [196].

In a replication of Neisser's study, Simons and Chabris showed participants a film of two basketball teams, one wearing black shirts and the other wearing white. These displays were created such that all of the actors were partially transparent and thus could simultaneously occupy the same locations. The researchers instructed participants to count how many times a basketball passed between members of one team, ignoring the other team. Just as Neisser had found two decades earlier, many participants did not notice a woman who walked through the scene carrying an open umbrella, even though the woman was present for several seconds [196].

Furthermore, Simons and Chabris extended the original findings by showing that inattentional blindness also occurs in more natural displays, in which all of the actors are fully visible and opaque. They repeated the previous experiment with the two teams of basketball players, dressed in black or white, each team passing a basketball among them-

selves [196]. The observers were again instructed to count the number of passes made by either the white team or the black team. Partway through this task, either a woman with an umbrella or a person in a gorilla costume unexpectedly walked through the center of the action, remaining clearly visible for about five seconds before exiting the display, see Figure 3.1 [196]. The observers were then asked if they had seen the unexpected object. Thirty-five percent of the observers failed to notice the woman with the umbrella, even though her presence was obvious to anyone not engaged in the counting task. Perhaps more surprisingly, given its more unusual nature, even more people (56%) failed to notice the gorilla who stopped to face the camera and thumped its chest before walking off the screen. In both cases the unexpected figure moved through the same spatial locations that were being occupied by the attended basketball players.

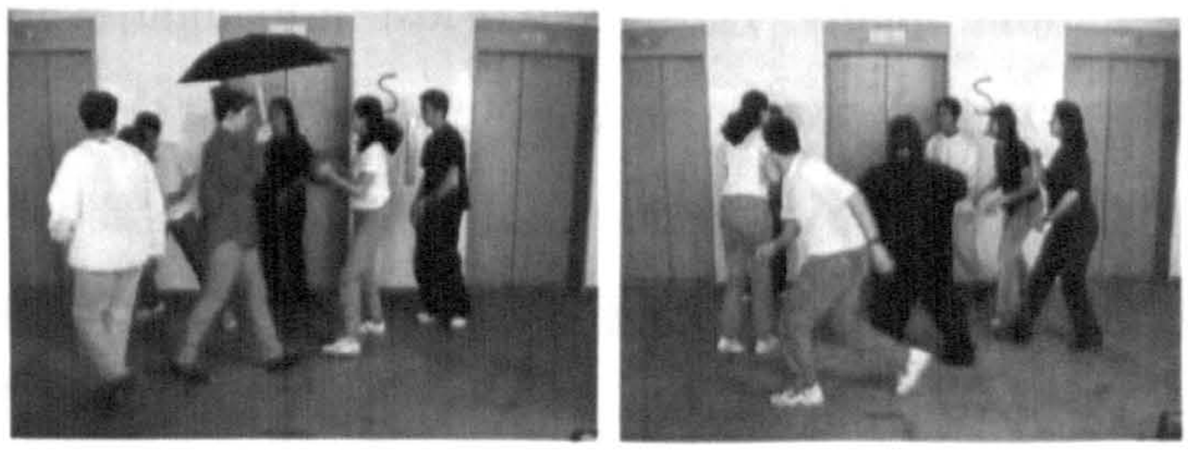


Figure 3.1: In a dramatic demonstration of inattentional blindness, half of the observers participating in Simons and Chabris' experiment [196] failed to notice a person wearing a gorilla suit (right image) and 35% failed to see a woman holding an open umbrella (image on the left) who walked into the middle of a basketball game, spending several seconds on screen. Images courtesy of Prof. Daniel J. Simons.

Their results also suggest that inattentional blindness may depend on the similarity between unexpected and attended objects [196]. Participants were more likely to notice a gorilla, whose fur was black, when they were attending to the basketball team in black shirts than when they were attending to the team in white shirts. Interestingly, spatial proximity of the critical object to attended locations does not appear to affect detection, suggesting that observers attend to objects and events, not spatial positions [196].

The phenomenon of inattentional blindness together with findings regarding auditory-visual interactions in visual orienting, which will be discussed in subsequent sections,

formed the basis of our perceptually- adaptive selective rendering approach, described in sections 6.2 and 7.3

3.3 Attention: Automatic or under endogenous control?

We have mentioned that attention lets important information through and filters the rest. This is not quite true. Certain information has the ability to 'break through' the attentional barrier, refer for example to [222]. It has been demonstrated that irrelevant stimuli can control where we attend, for example [222, 223, 267]. For instance, objects which are large and/or move are more likely to capture attention. In addition, we are more likely to notice very familiar inputs. Thus, even if lost in thought, we will notice when someone mentions our name. However, in the literature about the so called attentional capture phenomenon, there has been a great deal of discussion about the extent to which such effects are automatic or rather modulated by endogenous factors. Here we review data and theory related to this debate.

It is well known that there are at least two ways of allocating spatial attention. Observers can **voluntarily** allocate attention to the spatial location(s) that may contain information relevant to their task and focus their resources in order to accomplish that task. Hence, voluntary attention is also termed top-down, goal-directed attention, or endogenous attention. In this case, the intentions and strategies of the observer are in control of the allocation of attention, which in turn enhances the processing of selected events or locations [184, 221]. For example, if you are having a conversation with a friend in a crowded and noisy environment, then you pay attention to his voice, and this attentional process filters out the irrelevant noise.

On the other hand, stimulus events can **involuntarily** capture or demand our attention, even when the stimulus event is unrelated to the current goal-directed activity. Hence, involuntary attention is also called bottom-up, stimulus-driven capture or exogenous at-

tention. Exogenous orienting of attention, has often been assumed to be mediated by reflexive, low-level perceptual processes that automatically capture attention, independently of someone's intentions. For instance, a stimulus that suddenly appears in the periphery of the visual field can attract exogenous attention [98, 221, 269]. Or, if a car crashes behind us, we will certainly turn abruptly in the direction of the crashing sound.

Nakayama and Mackeben (1989) suggested that these two forms of attention have different time courses: endogenous attention rises slowly and can be sustained whereas exogenous attention shows a very rapid onset and cannot be sustained [146].

Behavioural, neurophysiological, and neuropsychological evidence suggest that the two forms of attentional orienting may be controlled by different neural substrates (e.g., [25, 201]).

Jonides (1981) [97] concluded that memory load affects endogenous but not exogenous attention orienting and concluded that top-down orienting is resource-limited, is easily suppressed, is affected by a subject's expectancies and by concurrent memory load and requires conscious awareness. On the other hand, bottom-up orienting is resource-free, cannot be suppressed, is unaffected by a subject's expectancies or by concurrent memory load and does not require conscious awareness [97]. McCormick's (1997) experimental results provided further support for the notion that exogenous orienting is an automatic process which occurs before detection and does not require conscious awareness, while endogenous orienting is a controlled and strategic process [133]. Moreover, when the task participants have to perform is very demanding and requires a focussed attentional state, it seems that attentional capture by irrelevant visual information is less likely to occur [105, 110]. As the attentional state becomes less focussed, it is more likely that irrelevant information will affect ongoing processing.

Thus, the general conclusion to be drawn is that attentional capture by visual distractors may be automatic by default, but can be either suppressed or enhanced by endogenous at-

tention processes (see, [184] for a detailed review). Moreover, human attentive behaviour appears to be composed of a complex mixture of both bottom-up and top-down attention processes [89].

Most of the studies above address differences in evoking voluntary and involuntary attention, but they do not address possible differences in their consequences. Prinzmetal et al. (2005) [171] demonstrated that voluntary and involuntary attention have different consequences for perception and performance, they serve different functions, and they affect different physiological processes. Based on behavioural, functional imaging, and electrophysiological evidence, they suggested that voluntary attention enhances the perceptual representation for objects and locations that are important for our current goals, whereas the function of involuntary attention is to select an object for response. The response may be an 'orienting response', such as a reflexive saccade, or the tendency to respond to an object in a particular location. Voluntary attention affects the allocation of perceptual processing resources so that more information accrues in attended locations [171].

Some of the above findings suggest that attentional capture by visual distractors can be supressed by the intentions and the strategies of a subject, the goal-directed attention in general. In the following section we will introduce and in section 3.6 we will further analyse what happens when attention is captured by auditory distractors. Are the latter easily supressed or is their 'distracting' power higher than that of visual distractors?

3.4 Attentional Selection between Spatially Defined Visual and Auditory Stimuli

Whereas at any given moment the human behavioural output is usually restricted to one single act or one single task, the senses simultaneously provide information about countless states and their current changes. Thus, the senses provide much more information than needed in order to appropriately perform an ongoing action or to accomplish a cur-

rent task. To deal with this problem, mechanisms have evolved that selectively facilitate the impact of only those stimuli on behaviour that are currently behaviourally relevant.

As we have mentioned before, attention has been compared to a spotlight that we selectively shine on things around us. It makes things stand out and be brought to our awareness. These things are then processed or interpreted by us. Attention can change swiftly, moving from one thing to another.

Research on the topic of selective attention has mainly concentrated on the spatial allocation of processing resources. One of the first key studies in the investigation of selective attention was that carried out by Colin Cherry (1953) [34] in his investigation of the cocktail party problem. Cherry devised a dichotic listening task to study what became known as *Focused Auditory Attention.* During his experiment participants would wear special headphones in which two different messages were presented, one to each ear. Their task was to concentrate on just one of the messages and to repeat everything they heard ("shadowing"). It was found that when they did this the listeners obtained very little information from the unattended ear, they often did not notice if the unattended message was spoken in a foreign language or even if the speech was reversed [262]. However, physical changes were noticed, for example if speech was replaced by a pure tone, or if the sex of the speaker changed, or if there was a change in the intensity (loudness) of the unattended message.

In the ensuing years, research by various investigators has uncovered some of the details of the processes involved in selectively attending to spatially defined stimuli. An important question for researchers in scene perception regards what determines where a person will attend in a scene. Visual psychology researchers, such as Yarbus (1967) [271], Yantis (1996) [267], and Itti and Koch (2000) [92], have demonstrated that the visual system is highly sensitive to features such as edges, abrupt changes in colour, and sudden movements.

In one of the most recent studies, Loschky and McConkie (2002) [117] found that there are certain pre-attentive processes that select a location for attention. These processes involve a basic visual analysis before object recognition, which brings to attention certain features of objects that make the target 'pop out', such as orientation, colour, spatial frequency, motion and size.

Below we will present the most influencing of these studies and how they dealt with important research questions regarding selective attention and attentional capture by irrelevant stimuli.

3.4.1 Attentional Capture by distractors

The ability to remain focussed on goal-relevant stimuli in the presence of potentially interfering distractors is crucial for any coherent cognitive function. However, people are often distracted by task-irrelevant stimuli, for example, objects that are novel, bright, colourful, and moving, and simply instructing people to ignore goal-irrelevant stimuli is not sufficient for preventing their processing. As Theeuwes et al. (1998) noted, "Our eyes do not always go where we want them to go" [222, page 379]. Daily life provides numerous examples: a fly hovering about might distract you while reading a book, an attractive billboard can distract a driver, and so forth.

In the laboratory, research that looked at the extent to which distractor processing can be prevented led to an enduring controversy. Mixed results as to whether focussing attention on task-relevant stimuli can exclude distractors from early perceptual processing (an 'early' selection effect) or can only prevent distractors from controlling behaviour and memory (a 'late' selection effect) has fuelled a long standing debate between early- and late-selection views of attention, see, for example, [47]. Recent evidence supports the notion that the effects of load on distractor processing depend crucially on the type of mental processing that is loaded.

Perceptual load Studies: behavioural experiments

Research on the role of perceptual load in selective attention was triggered by the hypothesis that perception has limited capacity (as in early-selection views) but processes all stimuli in an automatic mandatory fashion (as in late-selection views) until it runs out of capacity [106, 110]. This led to the predictions that high perceptual load that engages full capacity in relevant processing would leave no spare capacity for perception of task-irrelevant visual stimuli. In situations of low perceptual load, however, any capacity not taken up by task-relevant stimuli would be involuntarily allocated to the processing of task-irrelevant visual distractors. These predictions were tested in experiments that assessed the effects of varying perceptual load on distractor perception during task-relevant processing. These experiments found that increased perceptual load reduces, indeed typically eliminates, any visual distractor interference effects, in support of the perceptual load hypothesis [105, 106, 107].

Cognitive load Studies

Load on executive cognitive control functions, such as working memory, that renders them unavailable to actively maintain stimulus-processing priorities through-out task, seems to have the opposite effect to perceptual load: it increases interference by irrelevant low-priority distractors rather than decreases it. Behavioural studies demonstrate that high working-memory load can increase distractor response-competition effects on behaviour [109]. Attentional capture by a salient but task-irrelevant odd colour 'singleton' distractor, during shape-based search tasks, is also increased by high working-memory load [108].

The studies reviewed above illustrate the importance of considering the level and type of load involved in the task performed to determine interference by task-irrelevant distractors. Simply instructing people to focus attention on a certain task is not sufficient to prevent distractor interference. A high perceptual load that engages full attention in the task is also needed. In contrast with the effects of perceptual load, high cognitive-

control load increases distractor interference, suggesting that cognitive control is needed for actively maintaining the distinction between targets and distractors. These findings resolve in a way the long-standing early- versus late-selection debate and also clarify the role of cognitive control in visual selective attention: early selection critically depends on high perceptual load and cannot be achieved simply by exerting active cognitive control. Successful late selection, however, (namely, correct target responses despite perception of potent but irrelevant distractors, as in situations of low perceptual load) critically depends on active cognitive control functions being available for the selective attention task.

Crossmodal effects on perceptual load

It is important to determine whether attentional capacity is modality specific (such that, for example, visual load should have no effects on the perception of auditory distractors) or is shared between the modalities (such that load in one modality should determine distractor processing in another modality). With the prevailing emphasis in the last few decades on attention in vision, most load studies to date have been conducted in the visual modality, although a few studies have now examined crossmodal load effects.

One study replicated the within-modality visual load effects on distractor processing reported earlier by Lavie and Cox (1997) [107], but found that auditory presentation of the distractor letters resulted in greater distractor effects with high (versus low) load in the visual search task [219]. More specifically, using a response competition paradigm, Tellinghuisen and Nowak (2003) [219] investigated the ability to ignore target response compatible, target response incompatible, and neutral visual and auditory distractors presented during a visual search task. In three experiments, participants searched sets of one (easy search) or six (hard search) similar items. In Experiment 1, visual distractors influenced reaction time (RT) and accuracy only for easy searches, following the perceptual load model, according to which increased processing requirements prevent distractor processing. Surprisingly, auditory distractors yielded larger distracting effects for hard

searches than for easy searches. In Experiments 2 and 3, during hard searches, consistent RT benefits with response-compatible and RT costs with response-incompatible auditory distractors occurred only for hard searches. Tellinghuisen and Nowak suggested that auditory distractors are processed regardless of visual perceptual load but that the ability to inhibit cross-modal influence from auditory distractors is reduced under high visual load [219].

The subject of visual attentional capture by auditory distractors will be further analysed in section 3.6, where we will deal mainly with spatial attention and lay the foundations for the experiments regarding the spatial dimensions of visual attentional capture by auditory stimuli, presented in Chapter 6.

The effect of emotions on cognitive activity

A group of studies have investigated the role of emotions, and especially of arousal, in cognitive activity. Arousal typically refers to the degree of physiological activation or to the intensity of an emotional response [198]. Measures of arousal include adjectives that make reference to physiological states and intensity, e.g. vigor, activity, wakefulness. Discussions of the effects of emotional arousal on attention are assuming an increasingly prominent role of arousal in accounts of the experiential and behavioural consequences [7, 53, 60, 99, 103, 197, 238].

The original theoretical formulation was provided by Easterbrook (1959), who suggested that the effect of emotional arousal on attention is to narrow and focus the attentional field, by systematically reducing the range of cue utilisation. More specifically, it was proposed that responsiveness to peripheral or less relevant stimuli is diminished, while responsiveness to relevant or dominant cues is maintained, if not in fact augmented. This hypothesis essentially asserts that the effect of arousal involves capacity limitations in information processing and the activation of attentional control mechanisms [53].

Music has well-established psychological effects, including the induction and modification of cognitive states, moods and emotions. Many research studies have investigated the effect of music on arousal and mood and most of them favour the *Arousal-Mood hypothesis*, according to which listening to music affects arousal and mood, which then influence performance on various cognitive skills [67, 104, 123, 186, 198, 220, 264]. Physiological responses to music differ depending on the type of music heard: Exciting, fast tempo music has been found to increase arousal levels, see for example, [87]. In accord with that, Wakshlag, Reitz and Zillman (1982) [239] found that the presence of fast, appealing background music in educational television programs competes with program content for the viewer's attention and and results in reduced information acquisition.

The finding that music affects arousal and mood, which in turn imposes capacity limitations in the breadth of attention and information processing, inspired our first experiment (refer to Chapter 4), regarding the influence of music on the visual processing of the temporal dimensions of animated visual stimuli.

3.5 Theories of Selective Attention

Selective attention implies that we can consciously focus on one stimulus but usually at the expense of another. The *cocktail party phenomenon*, in which you are able to attend to one person at a party despite the high levels of background noise, is an example of selective attention. This section is concerned with some of the theories that have attempted to explain the selectivity of attention by using ideas from information processing theory. Of the most influential theories in the field, the majority fall into two broad categories: "bottleneck" theories and capacity model theories. It is worth noting that both bottleneck and capacity theories are based on the idea that humans have limited information processing capacity: i.e. we are never able to deal with of all the inputs that continuously flood into our processing systems from our senses and memory, and even if we were, we are limited

in the number of motor responses we can make. One can describe bottleneck theories as a strong version of this limited capacity idea, in that only one message at a time can enter consciousness, since at some point processing is reduced to a single channel. Capacity models, on the other hand, are a weaker version, in that information can be processed via many channels but that there is a fixed capacity limit to be distributed amongst the channels.

3.5.1 Early Theories of Attention: Bottleneck Theories

The clearest and the most influential bottleneck theories of selective attention are those of Broadbent (1958) [26], Treisman (1960,1964) [227, 228], and Deutsch and Deutsch (1963) [44].

The earliest bottleneck theory was developed by Broadbent (1958), in response to the problems encountered by air-traffic controllers in attending to more than one incoming message at a time. Broadbent's theory suggested that there is a narrow passageway in our working memory that limits the quantity of information to which we attend at any one time. Broadbent concluded that although a large amount of sensory information can be absorbed simultaneously, a selective filter (attention) reduces the input from one source while that from another source (information channel) is analysed by the brain [26]. Broadbent also argued that the selection of the channel to be attended is made on the basis of its physical properties only and not on the basis of the semantic content of the message.

Anne Treisman developed her *Attenuation Model* of a two-stage selection model of attention to account for experimental results that did not support Brodbent's model, for example [72], whereby it was claimed that the 'volume' of an unattended message was turned down but not off, such that an important word arriving in an unattended channel would nevertheless be perceived - having a lower activation threshold than other words.

Experimental support for attenuation theory was provided by Treisman (1960) in a dichotic listening study in which participants had to listen to a story in one ear and repeat it aloud, while ignoring the story in the other ear. At some point the stories being played to each ear would swap over and continue in the opposite ear. Treisman found that when this happened the participants would continue to shadow the story they started with and were often unaware that they had changed ears. In other words they were following the meaning of the story rather than just the physical characteristics of the incoming auditory information.

The *Pertinence model* was first proposed by Deutsch and Deutsch (1963) [44] and was later revised by Norman (1976) [153]. Deutsch and Deutsch's theory is based on the same empirical data as Treisman's theory, but is reinterpreted from a different perspective. These researchers argued that the two stage model in Triesman's attenuation theory is unnecessarily complex and in fact all information can be filtered semantically regardless of its physical characteristics. This semantic filtering determines if the information is pertinent or not and whether it will be attended to or ignored [44].

Later research showed that at least under conditions below information-overload, a capacity model of attention was probably more applicable than bottleneck theories.

3.5.2 Limited Capacity Theories

The definition of attention provided by William James implies that we can only attend to one thing at a time and Norman and Bobrow (1975) [154] have suggested the reason for this is that our attentional system has a limited capacity. So, instead of information being limited because it has to pass through filters, it is rather restricted because there is a limit to how much information we can attend to satisfactorily at any time. For example, having a conversation with a passenger while driving down a quiet street may not be a problem, but the same conversation in much more difficult driving conditions may not be possible.

There are two theories on how the limited amount of total available resources needed to complete tasks are allocated to attention: 1) Single-Resource Theory, and 2) Multiple-Resource Theory.

The *Single-Resource Theory* [99] argues that we have one single supply of undifferentiated resources available to all tasks and mental activities. Kahneman (1973) argued that many factors are implicated in the ability to divide attention between competing tasks, important among which being the familiarity or automaticity of the tasks being performed. A more automatic task- i.e., one which is more practiced- may take up very little of the available attentional resources, while more capacity can be allocated (to some maximum capacity) as the task difficulty and thus requirements in resources increase. If two tasks exceed capacity, there will be some performance decrement in one, or both [99].

Wickens (1984, 1992) provides a further alternative explanation for the findings of dual-task studies [256, 257]. In his *Multiple-resource Theory*, he argues that people possess multiple resources and proposes that there are three successive stages of processing: 1. Encoding. 2. Central processing. 3. Responding. Wickens' model makes two key assumptions: "There are several pools of resources. If two tasks make use of the different pools of resources, then people should be able to perform both tasks without disruption."

The Multiple-Resource Theory argues that tasks are processed based on multi-dimensional constraints. These constraints, involve the task's Codes (Spatial vs. Verbal), Modalities (Auditory vs. Visual), and Stages (Encoding, Central Processing, and Responding). As such, "...people have several different capacities with resource properties. Tasks will interfere more and difficulty-performance trade-off's will be more likely to occur, if more resources are shared" [257]. For example, two visually dominating tasks may compete for the same resources resulting in greater interference between the two tasks. But, if one task is visually dominating and one task is aurally dominating, they may not have to compete with each other, as they utilise separate resources.

There is reasonable evidence for the existence of multiple resources and many of the findings from dual-task studies can be accounted for by Wickens' model, see for example [3]. However, it cannot account for interference between an auditory and a visual task which, according to this model, require different pools of resources [22].

3.6 Crossmodal Attention

In everyday life, people commonly interact with a multisensory environment, in which attention becomes relevant for the selection and coordination of stimuli coming from different sensory modalities. Ryan (1940), in his survey of work concerning intersensory relationships, notes "It is a commonly observed fact that most objects of our everyday lives are perceived by means of two or more sensory modalities working in cooperation" [185].

Hence, of particular interest is the study of how selective attention gates the flow of information processing in the situations in which stimulation of different sensory systems is provided concurrently or in close temporal proximity. Although scientists have been aware for centuries that multiple senses act in concert in almost every aspect of behaviour and cognition, only recently has this difficult issue been investigated experimentally.

In the context of experimental literature, the concept of **crossmodal attention** is used to refer to capacities and effects involved in the process of coordinating - or 'matching' - the information picked up by multiple perceptual modalities [48]. For instance, during the multimodal perception of one unique speaker, the concept of 'crossmodal attention' is frequently used to refer to the capability to coordinate the information picked up by audition (speech perception) and vision (lip-reading).

It is becoming increasingly clear that multisensory processing is more of the rule than the exception [193] and that the different senses receive correlated input about the same external object or event, information which is often combined by our perceptual system to

yield a multimodally determined percept. In this regard, there is evidence of interference in processing stimuli coming simultaneously from different modalities, for example [54, 61, 134, 168].

Early work on how multisensory cues are combined favoured 'dominance' or 'capture' hypotheses and, in particular, visual dominance over discrepant auditory or haptic cues, for example, [37, 69]. An example of that is the **ventriloquism effect** [14, 16, 173], which creates the illusion during the synchronous presentation of auditory and visual events in somewhat separate locations, that the location of the sound is shifted in the direction of the visual stimulus.

However, other studies made clear that the 'visual dominance' view is over-simplistic and it overlooks many other cases of crossmodal interactions in which visual perception is systematically biased by auditory or tactile stimulation, see, for example, [61, 190, 250].

While many findings concerning threshold modulation in multimodal stimulation show contradictory results and are subject to methodological criticisms, see [252] for a discussion, crossmodal interactions occurring at higher levels of processing, i.e. levels requiring more attentional resources, seem to yield stronger effects, for example, the McGurk effect [134] where vision alters speech perception (for instance, the sound 'ba' is perceived as 'da' when viewed with the lip movements for 'ga', resulting in a percept which is different from both the visual and auditory stimuli). This observation led to the fundamental question of the importance of attentional mechanisms in audiovisual interactions. That is, whether attentional mechanisms need necessarily to be activated to some extent, to allow intersensory interactions to occur. Several psychophysical studies have reported crossmodal effects of attention, for example [201, 202, 203, 205].

Particularly, the possible role of audition in visual orienting, which has been little investigated until recent years, is now the point of convergence of many authors, for example [51, 164, 202, 203]. This interest is partly due to a growing body of neurophysio-

logical evidence for audiovisual interactions in structures known to be involved in overt orienting, such as the superior colliculus, for example [209]. Several investigators have argued that one of the primary functions of sound localisation is to direct the eyes toward auditorily specified events, for example [164]. This orienting reflex (*overt* visual orienting) involves the coordinated movement of the observer's eyes, head and body toward the sound source. It results in the foveation of the likely sound-emitting stimulus.

Beside the studies involving audiovisual interactions in the control of overt orienting, recent research is concerned with audiovisual interactions in *covert* visual orienting [202, 203]. Covert orienting occurs when attention is moved toward a stimulus without eye or head movements [167]. A distinction can be made between reflexive or *exogenous* covert orienting (induced by cues appearing directly at the target locus without predicting its position) and voluntary or *endogenous* covert orienting (induced by an informative cue, such as an arrow predicting the likely target side). Recently, Spence et al. (1997) [203] showed a one-way crossmodal dependence in exogenous covert orienting whereby audition influences vision, but not vice versa. They also showed strong audiovisual links in endogenous covert spatial attention.

3.6.1 Attending to one sensory modality versus another

The most basic crossmodal issue concerning attention is whether people can attend selectively to one modality at the expense of others, or whether the modalities are so independent that concentrating on one has no implications for the others. As long ago as 1893, Wundt [266], claimed that attention can speed up the processing of a relevant modality, while delaying irrelevant modalities; many subsequent authors have concurred. For instance, Massaro and Warner (1977) showed there is a small but significant limitation of attentional capacity during visual and auditory perception [127].

A longstanding issue is concerned with describing resources as a single, central, undiffer-

entiated pool [147] versus describing them as multiple pools allotted to specific domains such as sensory modalities [255]. Past studies of simultaneous attention to pairs of visual stimuli have used the "dual-task" paradigm to show that identification of the direction of a change in luminance is 'capacity-limited', while simple detection of these changes is governed by 'capacity-free' processes. On the basis of that finding, it has been suggested that the contrast between identification and detection reflects different processes in the sensory periphery.

Bonnel et al.'s study (1998) [21] questioned that assertion and investigated the contribution of central processing in resource limitation by applying the dual task to a situation in which one stimulus is auditory and one is visual. The results regarding the identification task demonstrated the tradeoff in performance, generally attributed to a limited capacity, but detection showed no loss compared with single-task controls. The data lend strong support to the idea of a single, higher order limitation on processing at a level that is central to either sensory modality alone. This adds still more credence to the idea of a central bottleneck on attention. Furthermore, the results imply that resources withdrawn from one sensory channel are fully applied to the other.

On these findings we based one of our key hypotheses that the redirection of part of the cognitive resources to the processing of the auditory stimuli embedded in animated computer graphics will reduce the total available resources for the processing of the visuals and therefore visual defects, both in the spatial and the temporal domains, might go unnoticed.

3.6.2 Crossmodal links in spatial attention

While there is a long history of studies investigating the non-spatial aspects of multimodal selection, there has been a recent growth of interest in studies examining the effects of crossmodal links in spatial attention. In particular, these have addressed the question of

what effect a shift of attention (either exogenous or endogenous) in one modality to a particular spatial location has on the spatial distribution of attention in other modalities.

Many studies document the existence of numerous crossmodal links in spatial attention [48, 55, 202, 203, 204, 207], especially links between vision and audition [243]. However, spatial locations are not the only features over which attention can operate. Robust evidence exists that, under certain circumstances, visual attention is deployed to objects rather than to spatial locations [188]. In addition, many researchers have proposed that the notion of "objecthood" could also apply to the auditory modality [24, 231]. For example, Turatto et al. (2005) [231] demonstrated that attentional crossmodal links between vision and audition can also be established at an object-based level, i.e. when auditory objects are provided during a visual task. Their results showed that two task-irrelevant auditory objects affected the distribution of visual attention, with faster discrimination when the cue and the target occurred in the same auditory object than when they occurred in two different auditory objects. They attributed it to a possible crossmodal perceptual grouping between the auditory object and the visual events occurring on the same side [231].

Studies distinguish between *overt* and *covert* shifts of attention. The former refer to receptor, such as eye, head or hand, movements toward regions of interest, as opposed to the latter which concern internal attentional changes without receptor shifts [167]. Our review concentrates on covert mechanisms. A further distinction is between endogenous or voluntary mechanisms of attention versus exogenous mechanisms, since several qualitative differences between these two mechanisms of attention have been observed in unimodal studies [167, 202, 203]. Therefore, these two forms of attention should be considered separately when assessing any crossmodal spatial links.

Crossmodal links in endogenous spatial attention

Spence and Driver (1996,1997) [202, 203] have developed the orthogonal-cuing paradigm as a method for measuring the spatial distribution of covert attention in vision, hearing and touch. Participants make speeded discriminations concerning the elevation, up versus down, for each of a series of targets, presented in a random order such that target modality and target side are unpredictable. Covert attention is directed toward one side, for instance by informing participants that targets are most likely there. Note that the lateral direction of attention is orthogonal to the required up/down decision, and thus should not induce any response bias. Elevation judgements are typically faster and/or more accurate on the expected side, even though no receptor shifts are allowed, and even when target modality is uncertain. This suggests that stimulus localisation improves within endogenously attended regions.

Importantly, when a target is strongly expected on a particular side in just one modality, such as audition, up/down discriminations in other modalities, for example vision, also improve on that side, suggesting a tendency for common shifts in endogenous spatial attention across the modalities [202]. However, the spatial effect for secondary modalities is smaller than in the primary modality for which the spatial expectancy applies. Finally, some of their experimental results suggest that information from different sensory modalities may be integrated preattentively, to produce the multimodal internal spatial representations in which attention can be directed. Such preattentive crossmodal integration can, in some cases, produce helpful illusions that increase the efficiency of selective attention in complex scenes.

The above results suggest substantial crossmodal constraints on the allocation of endogenous attention.

Crossmodal links in exogenous spatial attention

Extensive crossmodal links have now also been reported for shifts of exogenous spatial attention, triggered by task-irrelevant but salient peripheral stimuli, rather than by spatial expectancies. When testing this issue, it is important that the triggering stimulus (or cue) does not predict the likely location of subsequent targets. Recent work shows that an entirely non-predictive cue in one modality can attract exogenous covert attention to its location in other modalities [203, 206], improving performance in speeded up/down discriminations for targets on the side of the cue. Although crossmodal links in exogenous spatial attention have primarily been studied using speeded responding paradigms, several recent psychophysical studies have also shown that crossmodal shifts of exogenous spatial attention lead both to a 'speed-up' in the relative time of arrival of cued stimuli and also to the increased perceptual saliency of cued stimuli, for example [206].

The studies of crossmodal attention discussed above invariably concerned situations in which the receptor systems for the various modalities were aligned in one particular default posture. The participant's head and eyes were fixed straight ahead, with each hand resting on a table in its usual hemispace, i.e. left hand on the left, right hand on the right. However, in daily life we can adopt many different postures, and the important point is that these spatially realign the receptors for the different modalities. A recent study of normal performance shows that links in exogenous attention between vision and audition are similarly maintained spatially across different postures, such as when the eye deviates to one side in the head [49]. Thus, an abrupt auditory event at a particular location does not attract visual attention merely by activating a fixed retinal location. Instead, the sound activates a representation of whichever retinal region currently corresponds to the external location of the sound. In turn, this depends on how the eye is currently deviated in the head, as signalled proprioceptively [49].

Taken together, the results from all the studies presented above converge on the conclusion

that there are extensive crossmodal links in attention, both exogenous and endogenous, between audition and vision. There is no single answer to the general question of how attention is coordinated across modalities. Instead, the exact nature of the crossmodal links seems to depend on the particular type of attention involved, e.g. covert versus overt or endogenous versus exogenous, and on the particular modalities concerned. Nevertheless, some generalisations can be made. Shifts of covert attention in one modality tend to be accompanied by corresponding shifts in other modalities, with just a few exceptions. The correspondence between the modalities in the direction of attention with respect to external space is largely maintained across changes in posture, even when these realign receptors for the different senses. Some crossmodal integration can apparently arise before attentional selection is completed, and this contributes to the construction of the representational space(s) in which attention is directed. Also, given that crossmodal links have been demonstrated for both exogenous and endogenous attention when studied in isolation, an important question for future research will be to examine how these two forms of spatial attention interact to control multisensory selection in more realistic settings.

3.7 Spatial Attention and Eye Movements

Shifting visual attention to an object can be performed with eye movements towards this object, in which case attention is foveally centred by means of a fixation or smooth pursuit eye motion (overt orienting). Nevertheless, for tasks that require only limited acuity, attention can be shifted without eye movements (covert orienting). In fact, even at present it is unclear whether the attentional effects reported in many of the studies on attention reflect the consequences of covert orienting, overt orienting, or some unknown combination of the two [208].

Extensive research into the role of eye movements in human visual perception has been carried out for many years. Groner and Groner (1989) [74] reviewed recent findings

(both experimental and theoretical) from neurophysiological and psychological research on attention and eye movements and how these are related to cognitive theorising.

Although the mechanisms of visual search have been researched for many years, accurate measurements of eye movements have only become possible in the last few decades. Recent advancement in eye tracking hardware design has allowed the implementation of visual search experiments in both laboratory settings and public displays.

In this section, we describe the most common eye movement and some of the major models for the eye movement during scene perception. We are particularly interested in how attentional shifts are related to eye movements, since for the 'spatial' aspect of our work we hypothesised that sound spatially captures visual attention and results in the fovcation of the perceptual origin of the sound. To test our hypothesis we conducted a suitable eye-tracking experiment, described in section 6.3.

3.7.1 Types of Eye Movements

Human eyes do not have a uniform visual response, in fact, the best visual acuity is only within a visual angle of one to two degrees. This is called foveal vision, and for areas that we do not direct our eyes towards when observing a scene, we have to rely on a cruder representation of the objects offered by non-foveal vision, of which the visual acuity drops off dramatically from the centre of focus. The intrinsic dynamics of eye movement are complex, but the following movements are common:

- **Saccadic:** This is the rapid voluntary eye movement from one point to another used to reposition the fovea to a new point in the visual scene.
- **Miniature:** These are a group of small involuntary eye movements, of less than 1 degree visual angle, that cause the eye to have a dither effect. They include drift,

micro-saccades and eye-tremor.

- **Pursuit:** It is a smooth involuntary eye movement that acts to keep a moving object foveated.

- **Compensatory:** These movements are similar to the pursuit movement but they maintain a fixation while the head is moving.

- **Vergence:** This is how the eyes converge to a single point, depending on whether the viewer is focussing on a near or far object.

- **Optokinetic:** This is an involuntary saw-tooth movement that the eye performs when observing repeated moving patterns.

A *Saccade* has a fast acceleration at approximately 40,000 degrees/$second^2$ and a peak velocity of 600 degrees/second. As saccadic eye movements are so fast, visual input is disrupted for a brief period of time while a saccade takes place, primarily due to the very fast motion of images across the retina. This occurrence is called saccadic suppression. Saccades are observed in visual searches in the range of 1-40 degrees visual angle and they range in duration from about 10ms to 100ms [50].

Miniature eye movements impose a lower limit on the accuracy of eye tracking techniques. An accuracy of more than 0.5 is generally thought not to be required. The other types of eye movement are employed when following moving objects and also for the viewer to be able to maintain reference points during head motions.

When we try to understand a scene, we do not scan it randomly. Instead, we fixate our eyes on particular areas and move between them. During visual search, a saccade moves the gaze of the eye to look at the current area of interest. This area of interest normally needs to be dwelled on for longer than 100 ms so that the brain can register what is in that area. When this happens, the point is called a fixation. Fixations stabilise the retina over a motionless object of interest. 90% of viewing time is employed in fixations. Saccades and

fixations are considered to be the determinants of visual attention. Specifically, saccades tend to be evidence of the desire to voluntarily change the focus of attention, and fixations of the desire to retain gaze onto an object of interest. It has now been widely accepted that fixations are governed by intention, and therefore areas containing more information are preferred. In practice, however, it is also common that some areas are completely ignored.

Visual acuity and colour sensitivity are highest at the fixation point and fall off rapidly and continuously with visual eccentricity. Cognitive processing and memory encoding are also most complete for the region of the scene at fixation and drop off monotonically from that point [148, 80]. Because human vision is structured with a high resolution central region and a lower resolution visual surround, human scene perception is active and dynamic: Three to four times each second the viewer selects, via a saccadic eye movement, a specific region of the scene to receive priority for perceptual and cognitive analysis (see [79] for a review).

Kahneman [99] provided a different classification of eye movements depending on the situation when they occur:

- **Spontaneous looking** occurs when the subject observes any scene without any specific task in mind.
- **Task-relevant looking** is guided by a particular task during observation.
- **Orientation of thought looking** arises as a consequence of the observer not paying attention to the actual visual field under consideration but rather his/her attention is allocated to an inner thought.

3.7.2 Eye Movement Control

During visual perception and recognition, human eyes do not scan a scene in a raster-like fashion, but they rather 'jump' and successively fixate at the most informative parts of the

image under the control of visual attention [27, 156, 271], see illustrations in Figure 3.2.

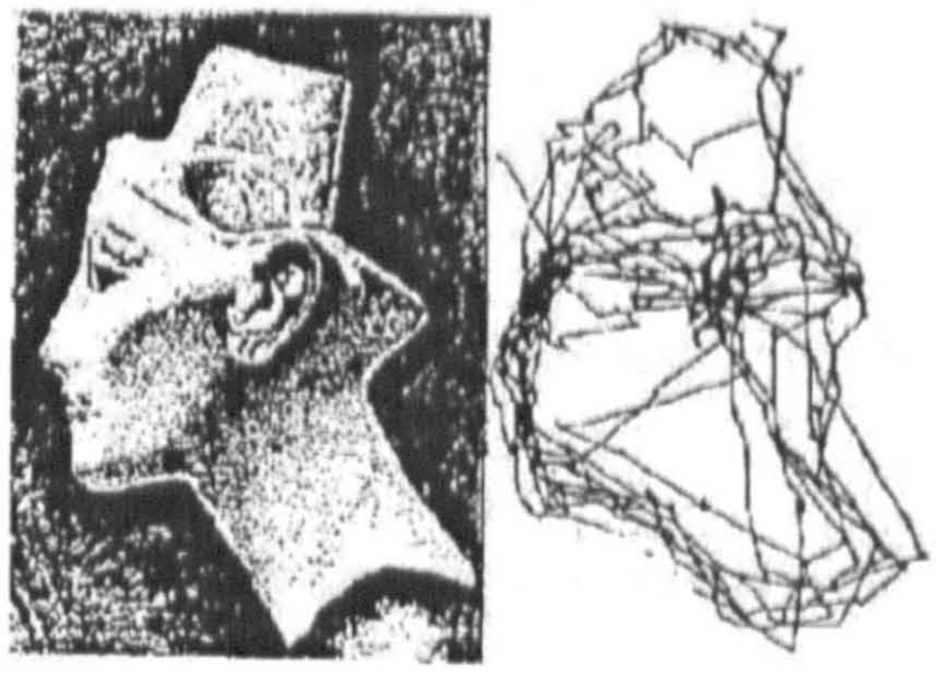


Figure 3.2: During visual perception and recognition, human eyes move and successively fixate at the most informative parts of the image (from [271]).

Yarbus (1967) [271] found that suggestions or hints can change the manner with which an observer 'scans' a scene and therefore can modify the corresponding scan path. In his famous experiment in the 1960s he recorded the eye movements of a subject while he was observing a painting by Repin named "An Unexpected Visitor" (Figure 3.3). During the experiment, the pattern of eye movements changed substantially as the viewer was given a number of prompts (Figure 3.4). Such prompts turned out to guide the way the subject observed the painting.

Figure 3.3: "An Unexpected Visitor" by Repin.

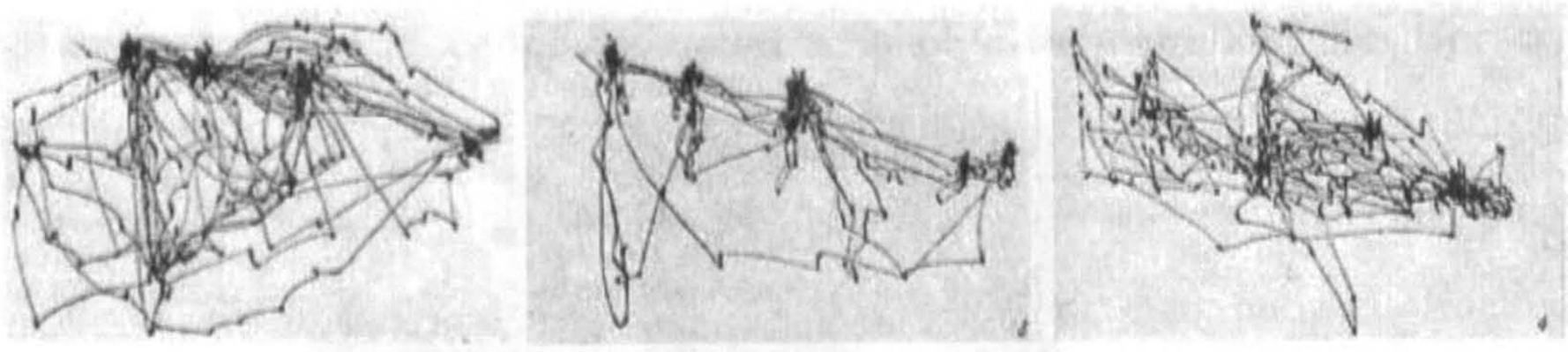


Figure 3.4: Recordings of saccadic eye movements scanning "The unexpected Visitor" (from [271]).

Figure 3.4 depicts three eye movement maps. Although they belong to the same individual looking at the same image, the patterns of movement are different, because the viewer was given different instructions in each case. First, the viewer was told to look at the image, but was given no specific instructions about what to look for (left map, Figure 3.4). Second, he was instructed to identify the ages of the people in the picture (centre map, Figure 3.4). Third, he was asked to determine what the people in the picture were doing before the visitor arrived so unexpectedly (right map, Figure 3.4). Each instruction required a different viewing strategy, which resulted in a different pattern of eye movement.

Extensive experiments have been conducted by Noton and Stark (1971) [156] based on the scan path theory. Their approach was based on tracking the eyes of a subject when observing an image for the first time. Their studies of eye movement scan paths over images showed that different subjects fixate similar regions of interest, although there are often variations in the temporal order in which fixation points are viewed by different observers.

It is important to note that prior knowledge about the scene being observed largely affects the way that the image is scanned. Noton and Stark [156] performed another test to determine how the eyes moved when the subjects were presented with an image that they had seen before, by comparing the individual scan paths of human eye movements in two phases: during image memorising, and during the subsequent recognition of the same image. They found these scan paths to be topologically similar and suggested that each object is memorised and stored in memory as an alternating sequence of object features

and eye movements required to reach the next feature, see Figure 3.5.

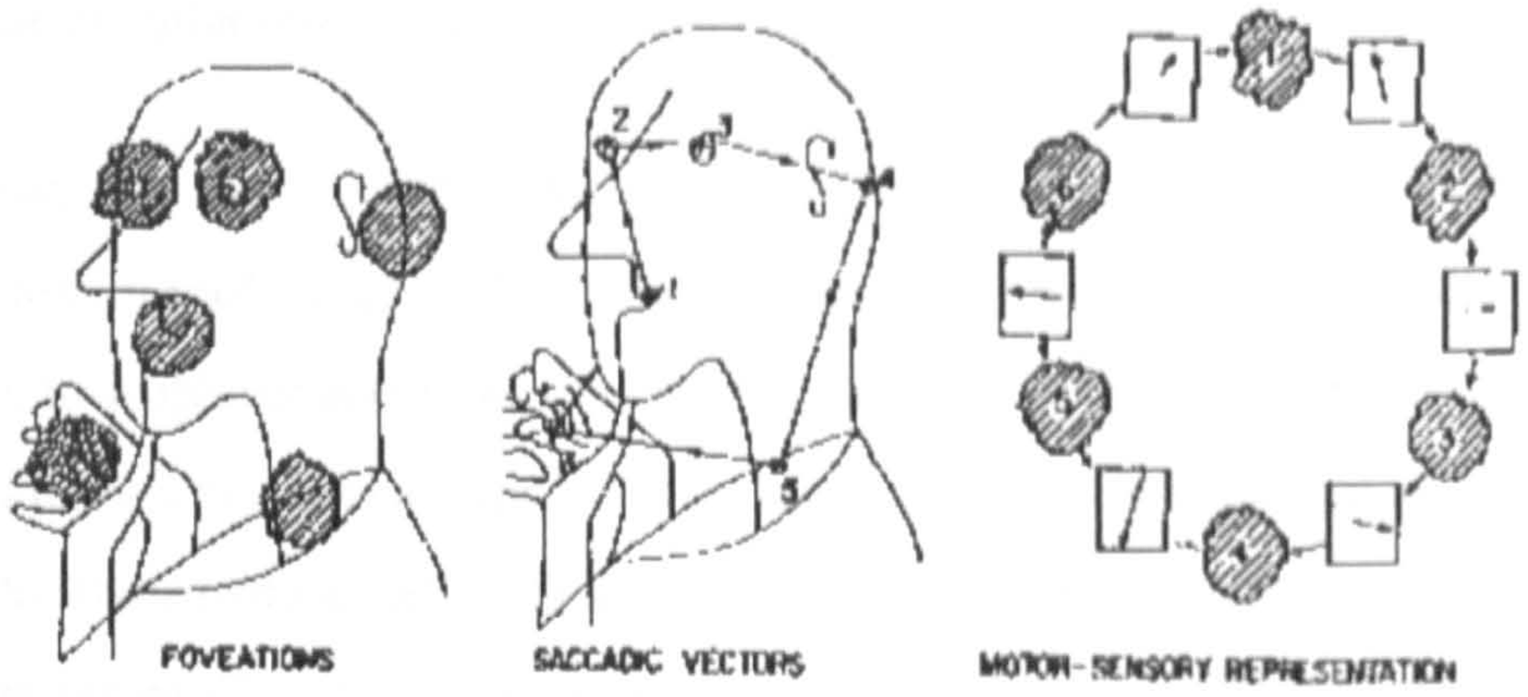


Figure 3.5: According to Noton and Stark (1971) [156], each object is memorised and stored in memory as an alternating sequence of object features and eye movements required to reach the next feature.

Findlay et al. (1992) [62] presented results from an experiment in which subjects' eye movements were recorded while they carried out two visual tasks with similar material. One task was chosen to require close visual scrutiny; the second was less visually demanding. The oculomotor behaviour in the two tasks differed in three ways. (1) When scrutinising, there was a reduction in the area of visual space over which stimulation influences saccadic eye movements. (2) When moving their eyes to targets requiring scrutiny, subjects were more likely to make a corrective saccade. (3) The duration of fixations on targets requiring scrutiny was increased.

3.7.3 How Shifts of Attention are related to Eye Movements

The relationship between selective spatial attention and eye movements has been the subject of speculation over the years. Even though there is a large body of evidence which suggests that the appearance of a new object may capture attention, it was, until recently at least, largely unknown whether such an event also triggers a subsequent eye movement.

Recent theories of visual attention, such as the oculomotor readiness theory of Klein

(1980) [101] and the sequential attention theory of Henderson (1992) [78], propose a link between shifts in spatial attention and the generation of saccadic eye movements.

On the contrary, other studies have demonstrated that it is possible to dissociate the line of attention from the gaze direction (see [58, 96, 177]). Remington (1980) [177], for instance, performed four threshold detection experiments to address issues concerning the relationship between shifts of spatial attention and saccadic eye movements. The results from these experiments support the contention that the mechanisms that shift attention are separate from those that control saccadic eye movements. Relevant events in the visual field periphery, however, triggered both a saccade and attention shift. The attentional response to such events did not appear to be under subjects' control.

We will next present research findings related to the link between attentional shifts and eye movements. In our discussion we will look separately into voluntary and involuntary attention.

Voluntary Attention and Eye Movements. Deubel and Schneider (1996) [43] and Hoffman and Subramaniam [82] have demonstrated that attention precedes voluntary eye movements to the target of the saccade. The eye typically will land at the position at which attention is directed [43]. More specifically, Hoffman and Subramaniam (1995) investigated the relationship between saccadic eye movements and visual spatial attention in two experiments [82]. In the first experiment, subjects were required to make a saccade to a specified location while also detecting a visual target presented just prior to the eye movement. Detection accuracy was highest when the location of the target coincided with the location of the saccade, suggesting that subjects use spatial attention in the programming and/or execution of saccadic eye movements. In the second experiment, subjects were explicitly directed to attend to a particular location and to make a saccade to the same location or to a different one. Superior target detection occurred at the saccade location regardless of attention instructions. This finding suggests that subjects cannot move their eyes to one location and attend to a different one. They concluded

that visuospatial attention is an important mechanism in generating voluntary saccadic eye movements [82].

Involuntary Attention and Eye Movements. More recently, Theeuwes et al. [223] examined whether whether attention precedes involuntary eye movements and demonstrated that the sudden appearance of a new object can lead to involuntary eye movements to be made to that new object rather than to the intended target. They reported that a goal-directed eye movement toward a uniquely coloured object is disrupted by the appearance of a new but task-irrelevant object, unless subjects have a sufficient amount of time to focus their attention on the location of the target prior to the appearance of the new object.

In many instances, the eyes started moving toward the new object before gaze started to shift to the target (colour singleton). The eyes often landed for a very short period of time (25-150 ms) near the new object. Depending on which eye movement program is ready first, the eyes will start moving in the direction of the onset or in the direction of the colour singleton.

The results suggest parallel programming of two saccades: one voluntary, goal-directed eye movement toward the target and one stimulus-driven eye movement reflexively elicited by the appearance of the new object. These results also indicate that in everyday life, particular events or objects may catch a person's eyes even when they run counter to the person's intentions. When a new object appears in the scene, it can interrupt ongoing goal-directed eye movement behaviour and elicit an eye movement to its location. Such a mechanism is ecologically beneficial because new objects are potentially important to the organism.

Information integration across saccadic eye movements. The visual world contains more information than can be perceived in a single glance. Consequently, one's perceptual representation of the environment is built up via the integration of information across saccadic eye movements. Irwin (1991) [88] investigated the properties of transsaccadic

integration in six experiments. Subjects viewed a random dot pattern in one fixation, then judged whether a second dot pattern viewed in a subsequent fixation was identical to or different from the first. Inter-pattern interval, pattern complexity, and pattern displacement were varied in order to determined the duration, capacity, and representational format of transsaccadic memory. The experimental results indicated that transsaccadic memory is an undetailed, limited-capacity, long-lasting memory that is not strictly tied to absolute spatial position. In all these respects it is similar to, and perhaps identical with, visual short-term memory.

This is very important to our work, since we can expect viewers not to be able to perceive variations in the rendering quality of an animated scene during an involuntary saccade toward a sound-emitting 'distractor' object in the scene. Such an object, according to the findings regarding the attentional capture by auditory distractors presented above, will catch a viewer's eyes despite his intentions or task-relevant strategies. Therefore, we could selectively render the scene with the highest quality around the sound 'source' and at a reduced quality everywhere else, when the sound is heard. The undetailed transsaccadic memory combined with inattentional blindness may prohibit the observer from noticing this variation in the rendering quality.

3.7.4 Tracking of Eye Movements

Eye tracking is becoming increasingly important in understanding visual attention and cognitive processing of visual information in humans. The objective of eye tracking systems is to determine when and where fixations occur. The time order of the list of fixations represents the actual visual search that takes place.

As described by Andrew Duchowski (2000) [50] there are two main types of eye movement monitoring techniques used most commonly in eye movement experiments. These are methods that measure the position of the eye relative to the head, and those that mea-

sure the orientation of the eye in space. There are various methods of tracking the eye position relative to the head. Assuming the head position is known, then, this can be used to follow the gaze of a subject. The methods that measure the orientation of the eye in space are more popular for identifying elements in a visual scene [50].

Eye tracking results are represented as lists of fixation data. This data contains, for example, the fixation location, its duration and the start and end times of each fixation. The results are analysed to calculate the areas which received most attention, i.e. the areas which exhibit high fixation density.

In the last few years, eye tracking techniques are attracting increasing interest from the computer graphics community due to the rich sets of visual stimuli generated, ranging from 2D imagery to 3D immersive virtual scenarios. Visual attention and perception can influence the way scenes are constructed. One of the possibilities opened to the community is the use of eye tracking for gaze-contingent applications for which the resolution and quality of an image or animation is adjusted according to the viewer's saccades and fixations. Another line of research not much studied yet is the analysis of the perception of synthesised images and animations with the objectives of optimising their perceived 'realism' and improving the efficiency of graphics algorithms [159].

Eye tracking techniques are vital for the design and development of the future virtual environments. It is foreseeable that current research in binocular eye tracking systems for determining a viewer's gaze in 3D space will become an integral part of future immersive media systems. Both research and development are underway for providing gaze-enabled interactive environments, as there is still a range of system level problems that need to be addressed. Most existing eye tracking systems still require a cumbersome calibration process, and issues related to accuracy, reliability, and user friendliness can potentially become problems when eye tracking is used for general mass-market applications.

3.8 Human Perception and Computer Graphics

As we discussed above, with the help of pre-attentive processes, we decide what to pay attention to and what to filter out and ignore. Attention filters and feeds information about the world around us into our minds. This filtering implies that objects or features that are outside the scope of attention are not noticed by the viewer.

In recent years, computer graphics techniques have incorporated perceptually-based rendering methods, which take advantage of shortcomings of the human visual system (HVS) and characteristics of human perception and attention, in general, in order to render more realistic images or gain time. Since it is the human observer who eventually judges the fidelity of the rendered images, the aim is to minimise the perceivable differences between the generated images and their real world counterparts. Thus, visual perception issues should be considered at various stages of computation, rendering and displaying [33]. This research direction has attracted much attention from the computer graphics community [73], whose research has been motivated by the progress in psychophysics. Perceptually adaptive graphics involve the investigation of the above issues and will be the subject of this section, as our research aims at bringing new insights into the perception of graphics.

This section is devoted to interdisciplinary research in the fields of computer graphics and perception, related to this thesis. More specifically, we discuss here relevant principles of human visual perception which can be exploited in order to develop perceptually-based metrics for the assessment of the quality of rendered scenes and algorithms which make the rendering procedure of realistic computer graphics more efficient.

In this section, principles of human visual perception that are employed in this thesis will be reviewed. Also, research that utilises principles of human visual perception towards e and image quality metrics will be mentioned. It is useful to demonstrate that such

knowledge is invaluable for the progress of the computer graphics field.

3.8.1 Rendering Quality/Fidelity Perception

An important question when aiming at rendering realistic images and animations is whether we are representing reality in a faithful or at least plausible manner and we are not just producing visually pleasing images. For real-time rendering, on the other hand, it is crucial to make the right speed-accuracy trade-offs in order to minimise the perceptibility of any resulting visual defects. Researchers are particularly interested in the types of visual artifacts that are most noticeable. They also focus on discovering new ways to 'fake' reality and use them in a methodical way to adapt graphics to the perception of the viewer.

Perceptually-based Image Quality Metrics

It is now possible to simulate the distribution of light energy in a scene with high accuracy, however physical accuracy in rendering does not ensure that the displayed images will look 'real'. Even if we assume that we simulate light distribution correctly (to within a given tolerance), there are problems with the way human observers perceive the resulting images. One of the reasons for this is the fact that current display devices can deliver only a limited range of intensities and most renderers cannot compensate for these limitations [230].

Therefore, validated visual models which predict image fidelity from a perceptual point of view would be more appropriate. Such models would enable researchers and developers work towards greater efficiency and speed in rendering, since they would know that the resulting images would look 'real' to the human eye [135]. Current global illumination algorithms, which deliver more physically-correct lighting effects, usually rely on energy-based error metrics, that do not necessarily correspond to the noticeable improvements of the image quality [112], in contrast to perceptually-based error metrics. Rushmeier et al.

(1995) explored a number of such perceptually based metrics, using ideas from the image compression literature, and concluded that perceptual metrics may be used to numerically compare rendered and captured images in a manner that approximately corresponds to human contrast perception [183].

One approach is to predict the visual impact the errors may have on the perceived fidelity of the rendered images. Another approach is to develop a perceptual metric which operates directly on the rendered images. This second approach takes a captured image of the real scene in question and a rendered image of the same scene and uses numerical techniques to determine the perceptual differences between the two, for example [183]. An example of an advanced image fidelity metric that operates directly on rendered images and incorporates a complex HVS model is the Visible Differences Predictor (VDP) [38], which 'fakes' many characteristics of human perception. The VDP takes as input a pair of images and generates as output a map of probability values that characterise if these differences could be perceived by a human observer.

Myszkowski (1998) [141] completed a comprehensive validation and calibration of the Visible Differences Predictor (VDP) response, using psychophysical experiments. More specifically, two experiments were conducted to determine the correspondence between VDP predictions and viewers' subjective reports of visible difference between images. Their results showed a good correspondence between human observations and VDP predictions [141].

In a more recent approach, McNamara et al. (2000) [136] introduced a method for measuring the perceptual equivalence between a real scene and a computer simulation of the same scene. The developed model was based on the outcome of psychophysical experiments which compared human judgements of lightness when viewing a real scene, Figure 3.6 (right), a photograph of the real scene and nine different computer graphics simulations, e.g., Figure 3.6 (left), including a poorly meshed radiosity and a raytraced image. They showed that certain rendering solutions, as the tone-mapped one, were of

the same perceptual quality as a photograph of the real scene [136].

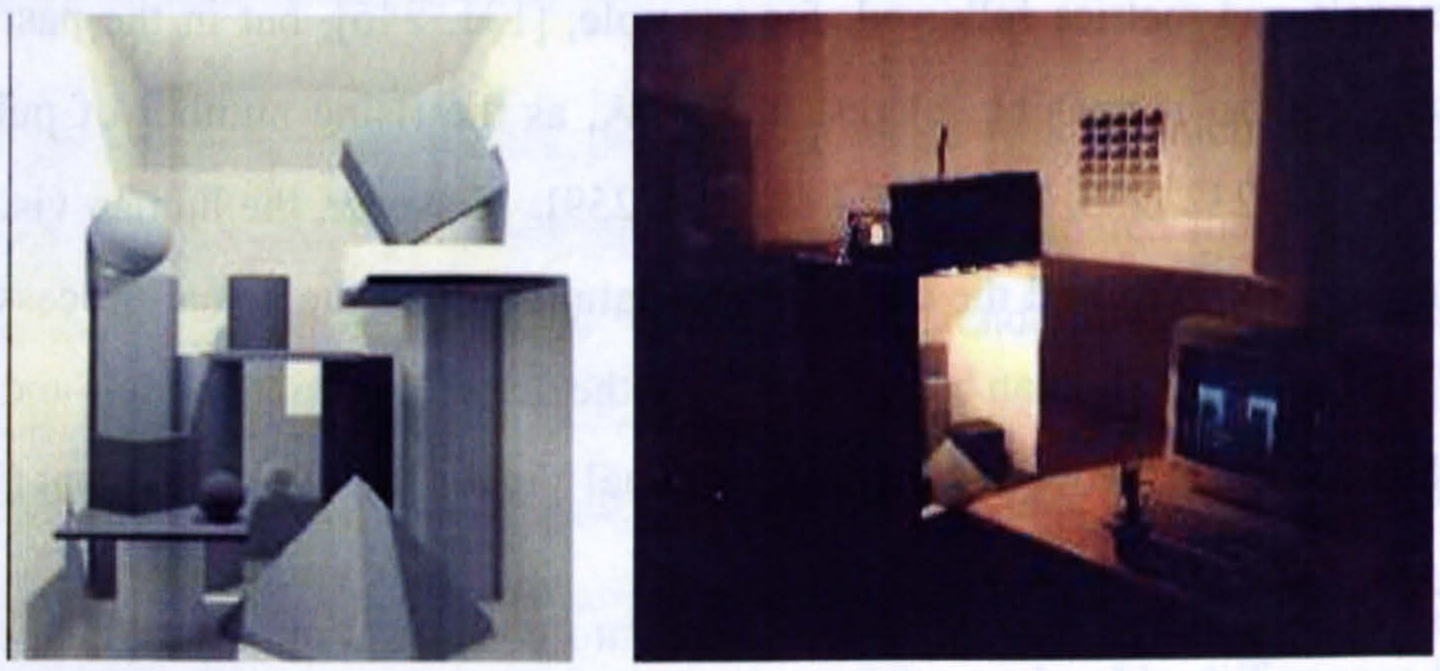


Figure 3.6: Comparing real and synthetic scenes using human judgements of lightness perception. Image courtesy of [136].

Visual Quality assessment for animations and videos

Because of compression, digital video systems exhibit artifacts such as, blockiness, blurriness, colour bleeding and motion compensation mismatches [5, 274]. The amount and visibility of these artifacts strongly depend on the image content.

Researchers initially used simple error measures, such as *Root Mean Squared Error* (RMSE), which operate solely on a pixel-by-pixel basis and do not take into account the image content and viewing conditions. Many experiments have confirmed their poor results compared to the perceived quality, for example [163].

These problems call for methods of more accurate video quality assessment. The ideal objective quality assessment system should rate video impairments like a human being. Considering the variety of compression algorithms available and the rapid change of technology in this field, a quality metric that is independent of the particular algorithm is preferable. Metrics based on models of the human visual system are one way to achieve this technology independence, because they are the most general and potentially the most accurate ones. Lukas and Budrikis (1982) [121] were the first to propose a comprehen-

sive metric based on a spatiotemporal model of the human visual system. A few other models and metrics followed, for example, [131, 246], but in the past few years there has been an increasing interest in PVQA, as the rising number of publications shows [85, 111, 232, 233, 218, 247, 249, 254, 259]. However, the human visual system is extremely complex and the existing uncertainties about the actual processing of visual information in the human brain complicate the design of vision models and explain many of the differences between existing Perceptual Video Quality Assessment (PVQA) systems.

A reliable PVQA system, should take into consideration what "quality" means to the viewer. A viewer's enjoyment when watching a video depends on many factors. One of the most important is, of course, its content and material. Provided the content itself is at least 'watchable', video and sound quality play a key role. Research has shown that video quality also depends on display size, resolution, viewing distance, brightness, contrast, sharpness, colour rightness, naturalness and other factors, see for instance [2]. In addition, it has been demonstrated that there is a difference between fidelity, the accurate reproduction on the display, and perceived quality. For instance, subjects prefer slightly more colourful images despite realising that they look somewhat unnatural [39, 274]. The accompanying sound has also been shown to influence perceived video quality: subjective quality ratings are generally higher when the test scenes are accompanied by good quality sound, which apparently lowers the viewer's ability to detect video impairments [180].

Visual Quality assessment for Interactive Computer Graphics Scenes (VR environments, games, simulations)

The mapping from a real scene to the computer graphics environment is mediated by environmental or visual fidelity. The term "environmental fidelity" refers to the degree to which visual features in the virtual environment conform to visual features in the real scene. Increases in fidelity, though, could prove to be computationally demanding and could impair the responsiveness of an interactive system. Greater efficiencies can be re-

alised by taking advantage of the limitations of the human visual system and not rendering scene features that will be imperceptible.

The perceptually-based approaches, presented in section 3.8.1, which use computational models of visual thresholds to efficiently produce approximated images that are indistinguishable from the highest quality "gold standard" renderings, are promising, but two factors limit their usefulness for interactive rendering. First, the computation of the metrics is itself a computationally intensive process that can take seconds or minutes, which make their use prohibitive for interactive rendering due to the time constraints. Second, these metrics are based on threshold measures of the noticeable differences between a rendered image and a "gold standard" image. In an interactive rendering scenario, time and resources are typically so limited that the differences will be well above threshold. Here, the appropriate question is not "how can I create an image that is visually indistinguishable from the gold standard", but "how can I make an image of the highest possible quality given my constraints" [124].

Padmos and Milders (1992) [161] presented a long list of quality criteria for simulator images. This list includes criteria based on: Visually Perceiving the Environment, Physical Image Properties, Image Capacity, Appearance of Surfaces, Visibility and Light Effects, and other miscellaneous features. The target simulator for these quality criteria is that of the vehicle simulator, but the same criteria would apply equally well to virtually any type of simulator image.

Horvitz and Lengyel (1997) [86] reviewed findings on visual search, attentional focus and their implications and introduced decision-theoretic models for the control of rendering approximations and their expected perceptual cost, when the available computational resources are varying or limited. To compute the expected cost, these models take into account the perceptual cost of degradations and a probability distribution over the attentional focus allocation of viewers.

More recent methods that order possible rendering operations to achieve high quality within system constraints offer a promising solution, for instance, [41, 52].

3.8.2 Crossmodal Interactions on the Perception of Quality

When one talks about using both audio and visual displays for some kind of simulation, game, VR, etc., some people will say that the use of high quality sound positively influences their perception of the visual images.

For example, Brenda Laurel states that "...in the game business we discovered that really high-quality audio will actually make people tell you that the games have better pictures, but really good pictures will not make audio sound better; in fact, they make audio sound worse" [224]. The reason is probably because simulations, games, VEs, etc., all started out as having only visuals and sound was added later. The addition of sounds, then, adds to the overall perception of the experience and as a result, the visuals appear better. The reverse is rarely reported, i.e. that the use of high-quality visual images positively influences the perception of auditory displays, probably because we are used to games and simulations which are primarily based on the visual displays.

In the following section we report experimental findings in which auditory displays influenced the quality perception of visual displays or vice versa. These findings are directly related to our research, as we are investigating the way sound alters the perception of the visuals (in both the spatial and the temporal domain) in a graphics environment.

Experimental Findings Assessing Complimentary Audio and Visual Quality

Audio and video quality are usually assessed as separate entities, despite the fact that it has been demonstrated that the quality of one medium can have an impact on the user's perceived quality of the other [84, 181, 244].

Neuman (1990,1991) [149, 150] conducted an experiment to measure the effect of variations in audio quality on the visual perception of High-Definition Television (HDTV). According to the experimental design, the quality of the auditory stimuli was manipulated, while keeping the quality of the visual stimuli constant. The auditory conditions were the following: low fidelity (very low-quality speaker system) vs. high fidelity (very high-quality speaker system); monaural vs. stereo sound. Three types of television programs were employed: sports, situation comedy, and action-adventure. Subjects were presented a short video clip along with one of the auditory conditions and they were then asked to rate a) their liking, b) their level of interest, c) their psychological involvement in the program, d) picture quality, and e) audio quality. The results indicated that subjects "...had a difficult time distinguishing mono from stereo and even low-fidelity from high-fidelity sound. ...[and] video with better quality and stereo sound were consistently rated as more likeable, interesting, and involving" [150]. Perhaps the most interesting finding- and certainly the most relevant to our research- was that a few subjects perceived an increase in visual quality when a video clip was coupled with better audio even though the visual quality remained constant throughout the experiment. This finding, however, was not statistically significant. Moreover, it did not apply to all three presented types of television programs, but only to one video clip type.

Hollier and Voelcker (1997) [83, 84] conducted two experiments investigating the influence of video quality on audio perception. Subjects watched video clips accompanied by audio (speech) commentaries. Results showed that an audio segment would receive different quality ratings depending on the quality of the corresponding video clip. More specifically, their results indicated that 1) when no video was present, the perceived audio quality was always worse than if video was present, and 2) although only small differences were noted, a decrease in video quality corresponded to a decrease in perceived audio quality. Also, the audio quality was found to have an influence on the perceived quality of the video. Hollier and Voelcker noted that "for a majority of applications both in the communications and entertainment industry separate evaluation of audio or video

quality is likely to become of limited value" [83].

Two companion papers by Woszczyk et al. (1995) [265] and Bech et al. (1995) [10] discuss the design and results of an experimental procedure examining the interaction between the auditory and visual modalities in the context of a home theatre system. Their approach acknowledges that "...experiments involving both modalities require a novel approach that recognises domains of cooperative interaction between the senses" [265]. With the growing interest and development of virtual reality systems, Woszczyk identifies the need for testing the interaction of audio and visual displays in order to bring about "substantial improvements in the integration of various audio and video parts of these [virtual reality] systems, and thereby provide important perceptual benefits that enhance [the] audio-visual experience of the viewers" [265]. The investigation of auditory-visual interaction is critical because "auditory and visual channels work both independently and in mutual cooperation on both cognitive and sensory levels of perception" and also "the matching of auditory and visual data triggers perceptual synergy between modalities and promotes intermodal fusion" [265].

Bech et al. agree that in order to study the interaction between the audio and visual sensory modalities "it is necessary to focus on the total experience and not on the two modalities individually" [10]. In their experiments, subjects assessed audio-visual reproductions using the subjective dimensions of action, space, mood, and motion while asking specific questions focussing on quality, magnitude, degree of involvement, and audio-visual balance. Quality was defined as: distinctness, clarity, and detail of the impression. One of their findings of particular interest is that both visual and audio perceived quality increased with increasing screen size. To further explore auditory-visual interaction, Bech conducted another two experiments to investigate the influence of stereophonic audio width on the perceived quality of an audiovisual presentation, using multi-channel surround sound systems. During the experiments, the subjects were asked to evaluate the quality (fidelity) of the spatial information contained in audio-visual reproductions. The

results indicate that "the quality of [perceived] spatial reproduction increases linearly with an increase in the stereophonic [audio] width" [9].

In the most comprehensive study, so far, on the effect of audio on the impression of visual images and vice versa, Storms [210] measured the impact of auditory-visual crossmodal perception phenomena by varying the quality (fidelity) of both auditory and visual displays. His overall findings strongly suggest the following:

1) When attending only to the visual modality, a high-quality visual display coupled with either a medium- or high-quality auditory display causes an increase in the perception of visual quality compared to visual-only quality perception evaluations.

2) When attending only to the auditory modality, a low-quality auditory display coupled with either a medium- or high-quality visual display causes a decrease in the perception of auditory quality compared to auditory-only quality perception evaluations.

3) When attending to both auditory and visual modalities, a high-quality visual display coupled with a low-, medium-, or high-quality auditory display causes an increase in the perception of visual quality compared to visual-only quality perception evaluations.

Winkler and Faller (2005), as well, investigated the factors which affect the evaluation of audiovisual quality and found that both audio and video quality contribute significantly to the perceived audiovisual quality [260, 261].

Users' perception of quality is also likely to vary with the task [84]. For instance, video quality may be more important in an intense interview situation than it might be in other relaxed scenarios. The perception of audio and video quality may also be directly linked to the level of quality a user assumes is necessary for the situation.

All the experimental results presented above provide the empirical evidence to support what most people in the gaming business, multimedia industry, entertainment industry,

and VR community empirically know: that audio can influence the quality perception of video and vice versa. These results also indicate that although we can divide our attention between audition and vision, we are not consciously aware of potentially significant intersensory effects. Nevertheless, the results from the field of crossmodal perception and intersensory interactions are still too few and computer graphics would benefit a lot from further study in this topic.

3.8.3 Perceptually-Aware Rendering Techniques

In most of today's systems the amount of computation done to produce a final image, is much higher than the amount of information the human visual system can process. Perceptually-based rendering uses all known information about what the eye actually 'sees', in order to reduce excess computation without compromising the resulting image quality as perceived by a human observer [275]. *Perceptually-adaptive* graphics improve image rendering efficiency by only computing information that is actually noticeable by the human observer.

Much of the work on perceptually-based rendering has focussed on two goals:

1) developing perceptual metrics that can be used to establish stopping criteria for high quality rendering systems, for example, [70, 138, 170, 175]

2) using perceptual metrics to optimally manage resource allocation for efficient rendering, for example, [19, 20, 52, 71, 75, 144, 145, 226, 273]

In the following sections we review some of the most recent perceptually-driven rendering approaches, which focus on the features that are more readily perceivable, under given viewing conditions, in order to accelerate interactive rendering. We first describe solutions for global illumination computations in dynamic environments and then examine perception-aware algorithms which take the focus of attention and eye movements

into account.

Perceptually-Guided Rendering of Global Illumination Solutions for Dynamic Environments

Global illumination effects, while essential for improved realism, are often omitted because of their high cost and low quality interactive alternatives are employed. In that case, while the quality of the resulting images can be satisfactory for less demanding applications, such as computer games, the visual fidelity of lighting effects, compared to their real world counterpart, is poor. On the other hand, adding more physically-correct lighting to a scene substantially increases the computational cost. Due to this cost, in dynamically changing environments only a limited number of pixels can be updated for every frame without impairing the sense of interactivity for the user. Thus, the question is how to efficiently allot the system resources in order to minimise the perceptual distance between the images obtained during an interactive session and the corresponding global illumination renderings [75]. The goal of the latest research in the field is to enable the production of high quality global illumination renderings at interactive rates by developing rendering algorithms that are perception-aware. This research approach makes the compilation of the rendering solution more efficient, by focussing on those scene features that are readily perceivable under given viewing conditions. The features that are below perceptual visibility thresholds can be simply omitted from the computation without causing any perceivable difference in the final image appearance [33].

Myszkowski et al. (1999, 2001, 2002) [142, 143, 145] described a perceptual Animation Quality Metric (AQM), which is based on the widely used original VDP by Daly [38]. To introduce the time dimension in the metric, they replaced the purely spatial "Contrast Sensitivity Function" (CSF) with a spatiovelocity CSF, which expresses the sensitivity of the eye to contrast as a function of the spatial frequency of the stimulus and its speed over the retina. A spatiovelocity CSF, rather than a spatiotemporal, is chosen because the AQM

was principally designed for the rendering of animation sequences. This means that, since camera pose, range data and frame rate are known, the pixel flow- and therefore velocities of each pixel- can be easily computed through image warping.

Myszkowski's first application of the AQM was to speed up the rendering of walkthroughs of static environments [143]. The main idea is to use computationally cheap *Image Based Rendering* (IBR) techniques to compute as many pixels of the in-between frames as acceptable, by using keyframes as reference views. Keyframes have to be chosen intelligently so as to minimise the number of pixels that have to be rendered. Warping gaps and areas where the pixel flow is slow are accurately rendered. The second application, first described in [145], aims at keeping noise below noticeable thresholds in stochastic global illumination rendering of dynamic scenes, by taking advantage of the temporal coherence in lighting distribution. The method is embedded in the framework of stochastic photon tracing and density estimation techniques [145]. An energy-based error metric, which operates locally, is used to prevent photon processing in the temporal domain for the scene regions in which lighting distribution changes rapidly. As a result, a perceptually-consistent quality across all animation frames is obtained, both in the spatial and temporal dimensions. Furthermore, the computation cost is reduced compared to the traditional approaches which operate solely in the spatial domain.

Dumont et al. (2003) [52] presented a general framework, based on a decision theory approach, which uses perceptual criteria to handle resource constraints in interactive rendering of pre-computed global illumination solutions. The rendering of a frame is seen as the result of a set of rendering actions, each with an associated cost and utility. The cost represents the amount of resources needed to take the action. The utility measures the contribution of the action to the result. It is defined as a measure of fidelity to the 'gold standard' solution, provided by a Visual Difference Predictor (VDP). Resource constraints are met by running a resource allocation algorithm that will maximise utility [52]. A key point is that, at equal cost, an ordering of the utilities of rendering actions is sufficient,

as opposed to an absolute estimation. An important issue is the choice of image regions for the global illumination computation, according to their potential ability to attract visual attention. It turns out that regions that are strong attractors of visual attention are highly correlated between subjects and usually affect a limited screen area [157, 271]. Thus, ordering the computations according to the predicted saliency of objects could be a promising strategy, which should reduce the noticeable image artifacts.

Yee (2000) [272] was the first to use a visual attention model to improve the efficiency of indirect lighting computations in the *Radiance* system [242] for dynamic environments. *Radiance* uses an ambient accuracy parameter as an error tolerance threshold when interpolating values from its irradiance cache, as opposed to computing them accurately. Yee et al. [273] modified this parameter on a per pixel basis, by using a spatiotemporal error tolerance map computed on estimate renderings of each frame. For less salient image regions, greater errors are tolerated and the indirect lighting is interpolated across a larger neighbourhood. This makes caching more efficient at the expense of blurring details in the reconstructed illumination solution. However, variability in the selection of the regions of interest (ROI) for different observers, or even for the same observer from session to session, can lead to some animation quality reduction in regions that were not considered as important attractors of the visual attention [273]. The tolerance map is computed in a similar manner to Myszkowski's AQM, the major difference being that Yee et al. incorporate a saliency map with the topographic locations of attention 'attractors'.

The saliency map is based on Itti et al.'s visual attention model [89, 91, 92, 93], which was originally designed for static images, but was later extensively validated in many applications, including both natural [89] and rendered scenes [273]. Based on a low-level architecture initially proposed by Koch and Ullman [102] and by Nieber and Koch [151], this visual attention model attempts to account for the automatic (exogenous) mechanisms responsible for attracting our attention to the salient locations in our environment.

The model, see Figure 3.7, processes an input image, calculating local contrast for in-

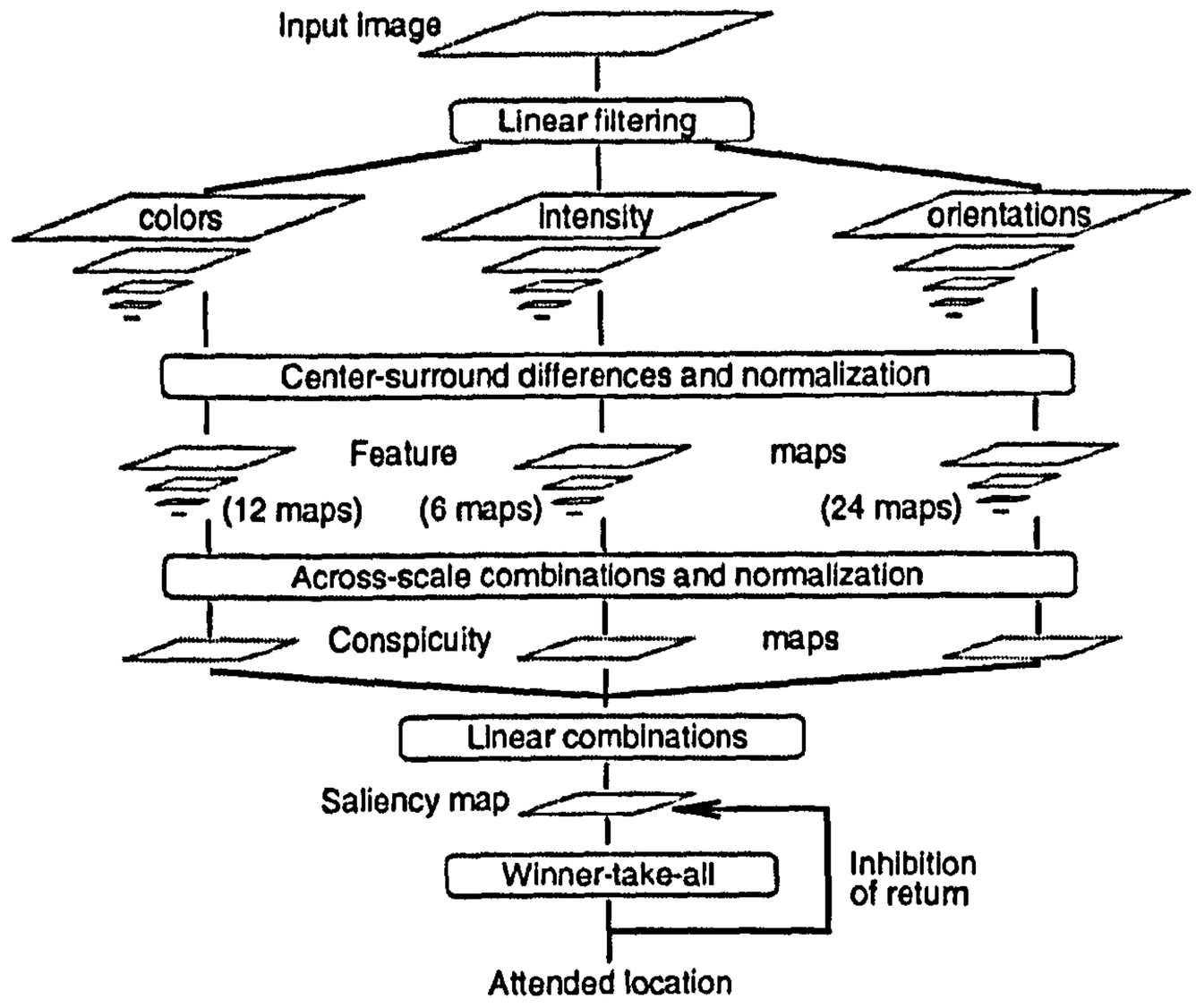


Figure 3.7: The initial bottom-up model of attention introduced by Itti, Koch and Niebur [91, 92, 93]. The input image is separated into three parallel feature channels (color, intensity, and orientation) and sampled at a series of spatial scales. Feature activity is propagated to the next level and reorganised into a center-surround fashion. Activity is normalised within each feature channel and linearly summed to form the salience map. Attentional focus is determined through a winner-take-all network. Once attended to, the current location is transiently inhibited in the saliency map by an inhibition-of-return (IOR) mechanism. Image courtesy of [93].

tensity, orientation and colour features, respectively. An input image is decomposed into four constituent channels, one for intensity, one for orientation and two for colour. Each channel is used as the first level in constructing a dyadic image pyramid, which is a set of images where each successive image is a filtered and decimated version of its predecessor. For the intensity and orientation channels, a Gaussian filter is applied. The orientation channel is filtered with Gabor filters of angles 0^o, 45^o, 90^o and 135^o [92]. Feature maps representing "centre-surround" differences are obtained from the filtered images. This makes the system sensitive to local feature contrast rather than feature amplitude. The feature maps for each feature are then combined respectively into three "conspicuity"

maps. Each conspicuity map provides a measurement of scene areas that 'pop out' for that feature type. Combining the conspicuity maps results in a saliency map. The final stage in the visual attention model is a winner-take-all (WTA) network. The WTA network finds the maximum of the saliency map at any time, which corresponds to the current most salient location in the scene.

In the absence of any further control mechanism, this system would direct attention, in the case of a static scene, constantly to one location, since the same winner would always be selected. To avoid this undesirable behaviour, Itti et al. [92] established an inhibitory feedback from the winner-take-all (WTA) array to the saliency map, in order to reduce the chances of 'success' of previous winners. Once attended to, the current location is transiently inhibited in the saliency map by this inhibition-of-return (IOR) mechanism. Thus, the WTA network naturally converges to the next most salient location, see [89] for details.

The success of a visual attention model which predicts fixations depends strongly on the similarity between the predicted and actual regions foveated by the observer. Marmitt and Duchowski (2002) [125] developed and evaluated a new method for the comparison of human and artificial scan paths recorded in virtual reality. Their method compares the sequence of regions of interest identified using Itti et al.'s attentional model [92] with those recorded from a human observer. They experimented with three different scenarios; a simple cube, a panorama, and a more complex graphical environment, which participants were allowed to free-view. They showed that, for all three situations, the similarities between the human and the artificial scan paths predicted by Itti et al.'s model are less than expected. They also found that this model assigns attention to a wider area of the image, whereas observers pay more attention to the central region of the display [125].

Itti et al. (2003) [90] extended their initial attention model to account for dynamic video clips. The extended model predicts the spatiotemporal employment of gaze onto any dynamic visual scene. Their extensions to the initial model include mainly a *flicker feature*

that detects temporal change, such as onset and offset of lights, and a *motion feature* which detects moving objects. Flicker is computed from the absolute difference between the luminance of the current frame and that of the previous frame, yielding a flicker pyramid. Motion is computed from spatially-shifted differences between Gabor pyramids from the current and previous frames. The same four Gabor orientations as in the orientation channel are used. Six feature maps are computed for each type of feature and therefore, 72 feature maps are computed in total: Six for intensity, 12 for colour, 24 for orientation, 6 for flicker and 24 for motion. The feature maps for each feature channel are normalised and summed into five separate "conspicuity" maps. For each spatial location, a single scalar measure of salience is calculated from the activity in the five "conspicuity" maps. The interested reader is referred to [90] for more details.

Haber et al. (2001) [75] also applied an advanced visual attention model to identify and order the image regions that are likely to be attended by the user. Their system uses that order to generate ray traced point samples that are "splatted" into the image plane using a stencil buffer test. Their approach (corrective splatting) is used for high quality rendering during interactive walkthroughs in environments containing objects with arbitrary light scattering characteristics. Due to the restricted computation time available during interactive rendering, their approach incorporates several aspects to obtain the best image quality as perceived by a human observer [75]. As a result of their processing, potential shading artifacts during camera motion are more likely to appear in less salient regions and will therefore be considered less annoying by the user. Their implementation delivers good results when interactively navigating through scenes of medium complexity at about 10 fps. In contrast to other approaches that require massive parallel computations, such as [241], their method performs very well on operating platforms with a limited number (4-8) of processors. Moreover, they can display high quality soft shadows and indirect illumination and they do not introduce any kind of rendering artifacts for purely diffuse objects [75].

Recently, Longhurst (2005) presented a new method for accelerating the production of synthetic images, which he termed "SnapShot" [113]. The basis of this approach is the modulation of sampling in a distributed ray tracing system according to a saliency map, generated in real-time using graphics hardware support. He introduced a new model for the prediction of saliency, which considers the saliency of objects due to both their distance from the observer and a measure of the viewer's "familiarisation" (habituation) with them, based on a prior history of visibility. The main reported advantages of the model are: a) The saliency estimate of a scene is generated without the need for an image to be produced using computationally expensive global illumination calculations, b) It allows for easy identification/ segregation of individual objects, c) The model predicts areas which are likely to be salient due to aliasing artifacts, d) Motion saliency is calculated from the 3D scene description rather than an image-based estimate of motion, e) Depth saliency is also calculated from the 3D scene description, f) Saliency due to habituation is calculated per pixel on a per object basis and g) This is all performed at real time rates due to efficient use of the GPU. He validated his model with a sample application which directed the sampling in a distributed ray tracing algorithm and showed a marked reduction in rendering time whilst maintaining the high perceptual level of quality.

Work by Cater et al. (2003) [31] supports the suggestion that visual attention is largely controlled by the task. They argue that task semantics can be employed in order to selectively render in high quality only the details of the scene that are being attended to. They introduce the concept of a task map, which is a two-dimensional map highlighting a given task. They show experimentally that it is possible to render scene objects not related to the task at lower resolution without the viewer noticing any reduction in quality [31].

Sundstedt et al. [214, 215, 216, 217] extended Cater et al.'s work by introducing the idea of an importance map, which is a combination of a task map and a saliency map, for the selective rendering of frames. The task map is calculated based on knowledge of what visual task the user is performing in the scene and the saliency map is generated

using a bottom-up visual attention model similar to [93]. The user can specify weighting parameters in the importance map for the relative importance of the task and the saliency maps. Sundstedt et al. [214, 217] showed experimentally that viewers performing a visual task while watching animations, consistently fail to notice the difference between high quality animations and animations selectively rendered based on their importance map. Using this technique they were able to accomplish considerable computational savings whilst maintaining a high perceptual visual fidelity.

Gaze-contingent approaches

In the past few years there has been a notable rise in the need for computer displays that have fast frame rates, high resolution and a large field of view in order to augment the viewer's perceptual experience. For example, aircraft simulators often require high enough resolution to identify aircraft from several miles away, and large enough displays to provide over 20 degrees of visual angle, which is necessary for certain aerial maneuvers. Remote piloting applications also, require fast update rates, high-resolution to identify terrain and large displays to provide accurate motion perception. However, these needs in computer displays are often constrained by the available transmission bandwidth or processing limitations, which lead to an extensive amount of time taken to render an image.

Instead of opting for more hardware, the user's attentional focus can be taken into account and display and computation resources can be instead directed to where they are needed most (Attention Driven Rendering). A single user can only focus on a small portion of the display at a time. Also, visual resolution decreases with an increase in retinal eccentricity, that is, the visual system perceives lower resolution in regions that are further away from the centre of vision [119]. Therefore, if a display system were to assign high resolution to the centre of gaze and reduce resolution elsewhere, it may not be noticed by the viewer.

Baudisch et al. (2003) [8] present several different "attentive display" approaches, such as the *gaze-contingent multi-resolution displays* (GCMRDs) which benefit from the variable resolution of the human visual system by varying display resolution according to the viewer's visual focus. More specifically, they present high-resolution imagery only at the location where the gaze is directed, with lower resolution elsewhere, Figure 3.8. An imperceptible degradation is difficult to achieve but often, in visual search tasks, the reduction in quality has no effect on performance even if it is noticeable. Furthermore, one should be able to optimise such a system, so that bandwidth savings are maximised while perceptual difficulties are minimised, if the GCMRD is closely tuned to the relevant parameters of the human visual system [152].

Figure 3.8: An example of multi-resolutional images used in GCMRDs, where high resolution information is put only where the user is looking at each moment, and lower resolution everywhere. Image courtesy of [119].

Loschky et al. (2001) [118] have argued that gaze-contingent multi-resolution displays can result in considerable computation and bit rate savings in single-user applications [118]. Given a system with sufficiently precise tracking of the centre of visual attention, a fast image update rate and a suitable image filtering method [118], one can produce a GCMRD without any negative impact or even sense of reduced image quality on the part of the human observer. Their study measured viewers' image quality judgements and their eye movement parameters, and found that photographic images filtered- as a function of contrast, spatial frequency, and retinal eccentricity- at a level predicted to be at or below

perceptual threshold produced results statistically indistinguishable from that of a full high-resolution display.

There are several techniques for the rendering of peripheral image [116]. The simplest is to display all peripheral information with the same resolution. Another technique is to reduce the resolution at predetermined distances from the fixation point. This technique can be extended to smoothly degrade the image. This smooth degradation attempts to match the spatial resolution function of the human eye.

3.9 Summary

To recapitulate the substance of this chapter, in the real world where stimuli flood our senses, competition between candidate objects often results in processing failures or confusion. Attention, in a very general sense, can be seen as a set of mechanisms and strategies designed to overcome these limitations and therefore it is closely linked to what we actually perceive. While the retina potentially embraces the entire scene, attention can only focus on one or a few elements at a time, and thus facilitate their perception, their recognition, or their memorisation for later recall. This results in 'peculiar' phenomena, such as inattentional blindness, which reflect, in fact, the limited capacity of attention.

When competition among inputs prevents the correct functioning of the visual system, attention can act by selectively biasing the competition towards one of the competing representations, i.e. one object, its location or another target feature. Studies on attentional selection in vision and relevant theories were presented.

We subsequently turned our focus to the evidence concerning how stimulus events may capture attention. At issue was the extent to which stimulus events can control the distribution of attention, independently of the goals and intentions of the observer. Attention can be directed to locations in space by a conscious and voluntary effort. It can also

be captured by abrupt onsets and other stimulus events. Stimulus-driven (bottom-up) attentional control is automatic, faster and more potent than goal-driven (top-down) attentional control. Several studies showed that attentional control results from an interaction between the observer's intentions and stimuli properties.

The role of emotions, and especially of that arousal, in cognitive activity was also discussed. Perception findings suggest that music induces and changes the levels of arousal, which then narrows and focusses the attentional field, by systematically reducing the range of cue utilisation. On these findings and findings regarding demonstrated cross-modal auditory-visual interactions, discussed briefly here and in depth in section 2.3.2, we based the 'temporal' path of our experimentation, presented in Chapter 4 and the first part of Chapter 5, where we investigated the effect of music on the perception of computer-generated visual stimuli in the temporal domain (i.e. perception of duration, motion velocity and delivery rate of the animated visuals).

We then went on to examine the most basic crossmodal issue concerning attention: whether people can attend selectively to one modality at the expense of others, as if there is a single higher order limitation on processing resources, or whether the modalities are so independent that concentrating on one has no implications for the others. A comprehensive report of findings, which indicate that there are limitations in attentional capacity during visual and auditory perception and processing resources applied to one sensory channel reduce the available resources for the other sensory channels, for example [127, 266], was given. One finding which has been very important to our research, was that of Tellinghuisen and Nowak (2003) [219], who investigated the ability to ignore visual and auditory distractors presented during a visual search task and concluded that auditory distractors are processed regardless of visual perceptual load and also that the higher this load is, the more 'distractive' the auditory stimuli can be [219].

The role of audition in visual orienting was highlighted next. Experimental results were presented, according to which there are extensive crossmodal links in spatial attentional

focus between audition and vision. Very important to our work are the findings which show that not only audition influences reflexive spatial attention orienting, but there are also strong audiovisual interactions in endogenous spatial attention, for example [203, 243], which lead to the function of sound to direct the eyes toward a likely sound-emitting source in a scene. These interactions have been attributed to a possible crossmodal perceptual grouping between the auditory event and the visual objects occurring at close proximity, for example [231]. The role of audio in visual attention orienting, coupled with the conclusions of Tellinghuisen and Nowak (2003) [219], formed the basis for our selective rendering approach, presented in Chapter 6, which exploits the 'distractive' power of audio in order to selectively render a scene and significantly speedup rendering, without any noticeable difference to the viewer, even when the viewer is engaged in a demanding visual task.

Section 3.7 presented the types of eye movements and also discussed how shifts of attention are related to eye movements and the role of the latter on scene perception. It also briefly overviewed eye tracking, a major tool for the investigation of the connection between eye movements, attention and visual perception.

We subsequently discussed relevant interdisciplinary research in computer graphics and human perception, with a focus on the application of attentional and perceptual models within the context of rendering. Taking into account that it is the human observer who finally judges the fidelity of the generated graphics, perceptually-based methods for the visual quality assessment of static images, animations/videos and interactive scenes, such as VR worlds and games, were subsequently presented.

Crossmodal auditory-visual interactions on the perception of quality were also overviewed. Although audio and visuals are usually assessed as separate entities, there is substantial evidence from experiments assessing complimentary audio and visual quality that the quality of the one medium can have an impact on the user's perceived quality of the other [150, 210] and that both audio and video quality contribute significantly to the

perceived audiovisual quality [260, 261]. Therefore, it is important to take crossmodal relationships into account when developing systems for multisensory display, in order to enhance the overall experience of the viewers. What is of particular interest to our research, is the fact that auditory stimuli have a strong impact on the perception of the visuals, both in the spatial and the temporal domain, see for example [210]. For instance, it has been shown that medium or high quality auditory displays coupled with visual displays increase the perceived quality of the latter [210].

We next reviewed current perceptually-driven techniques which enable the interactive, or nearly interactive, rendering of physically-based lighting effects. These methods employ knowledge about the characteristics and shortcomings of the human visual system, in order to accelerate the computation of the rendering solution. We concluded with perception-aware algorithms which take attention and eye movements into account in order to apply varying display resolution to different parts of the image, as a function of where the viewer's gaze is directed.

Despite the recent growth in the development of perceptually adaptive graphics techniques and attention-driven methods in the computer graphics field, researchers have restricted their focus on visual perception and the characteristics of the human visual system (HVS) only and have excluded multimodal criteria from their research. The main reason for that is the fact that too few results are currently available in the interdisciplinary field of computer graphics and crossmodal perception and therefore this topic needs considerable further study. The above findings and theories, coupled with relevant crossmodal research findings, inspired the 'spatial' direction of our research; taking into account the limits of processing resources, phenomena like inattentional blindness and the demonstrated visual attentional capture by auditory distractors, we experimented with selective rendering of 3D scenes which contain sound emitting objects that may attract visual attention.

In the following chapters we will show how graphics developers could take advantage of auditory stimuli, in order to reduce the computational load for the rendering and the

delivery of 3D graphics. These chapters will introduce our rationale and will present in detail the experimental work of this thesis.

Chapter 4

Preliminary Investigation of Temporal Perception

4.1 Introduction

The general conclusion from the research in auditory-visual perception, presented in sections 2.2.3, 2.3, 3.6 , is that sound influences visual perception and vice-versa (e.g. [9, 210, 245, 261]). According to some of these studies (see section 3.6.1), audio stimuli can potentially attract a part of the user's attention away from the visual stimuli, resulting in the reduced cognitive processing of the latter. A common example of this is turning down the radio in a car while looking for a particular street sign. Furthermore, Welch et al. (1986) [250] and Recanzone (2003) [176] found that the perceived rate of an audiovisual stimulus is determined primarily by audition.

Based on the perception findings regarding the 'auditory driving' phenomenon, presented in section 2.3.2, and the well-established effects of music on arousal and mood, discussed in section 3.4.1, which then influence performance on various cognitive skills, we decided

to further investigate the influence of musical stimuli on the human perception of temporal rate and duration, during the rendering and display of CG animated scenarios. As a first step towards this direction we examined whether music, and more specifically musical tempo and emotional suggestiveness of music, can serve as a 'driving' distractor of the presentation rate of the visual information.

This chapter outlines the experimental methodology employed and the relevant results for the first, informally designed, preliminary study which examined the influence of musical tempo and emotional suggestiveness of music on the perception of motion and time duration in a computer graphics environment [130]. The purpose of this study was to investigate whether music would be a significant distractor, allowing us to display frames at a slower rate without any perceivable difference to the user (Figure 4.1). Work in this chapter has been published in [130].

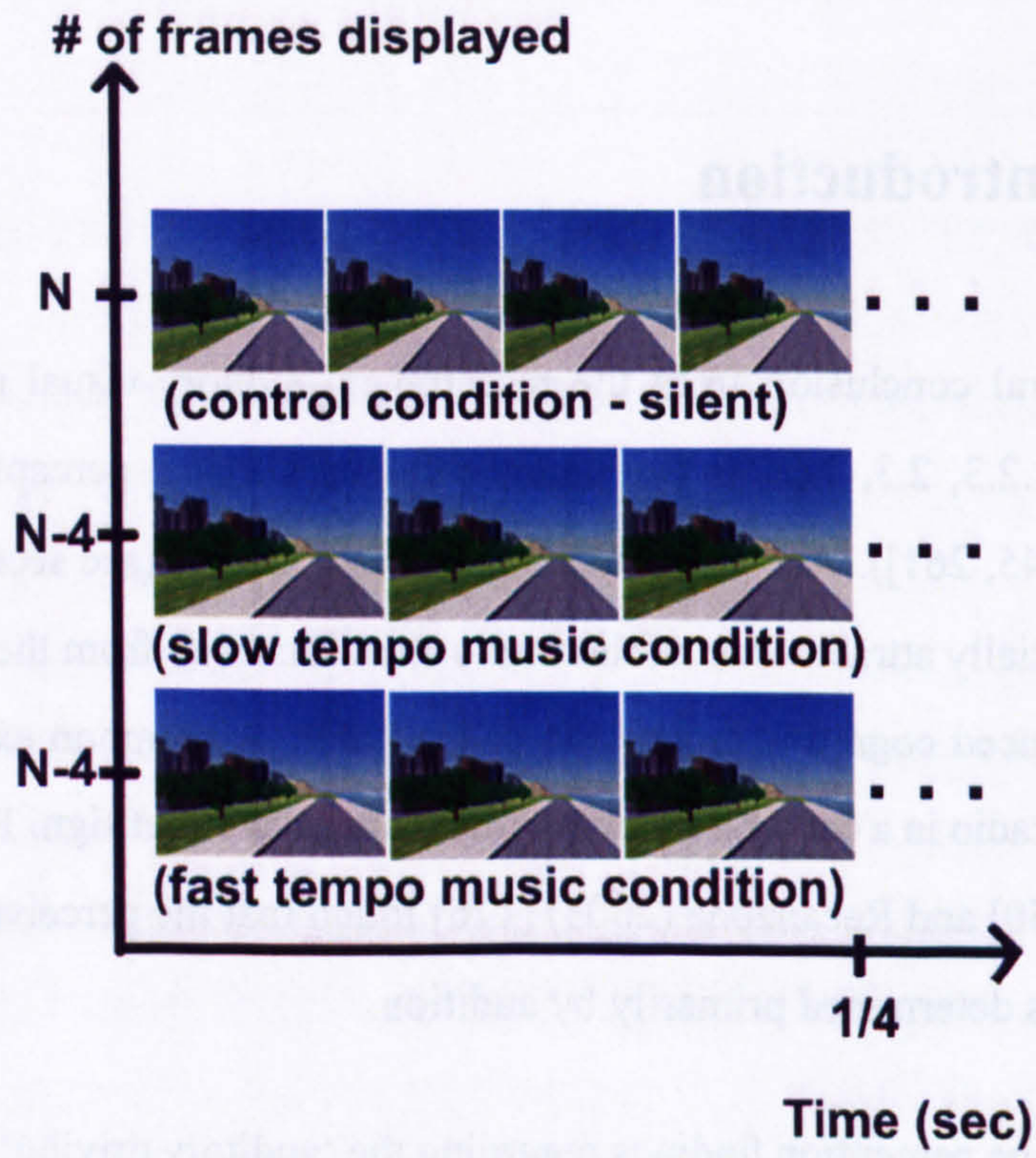


Figure 4.1: Less number of frames per second may be displayed when musical 'distractors' applied, without any noticeable difference to the observer.

Since this was an initial exploratory study which aimed to investigate the feasibility of our methodology, the experimental design was not very strict. A set of issues arising from this study will be discussed at the end of the Chapter. A formal experimental design, which took these issues into consideration, was employed for the two major studies presented in the following Chapters.

4.2 Experimental Methodology

In this section, the experimental methodology is going to be described in detail. Our hypothesis and the design of the study will be presented, in addition to three small pilot studies which were conducted before the main study, in order to test experimental materials and procedure and spot potential flaws. Participants, conditions tested, apparatus and materials of the study are also discussed.

4.2.1 Hypothesis

We hypothesised that the addition of music to an animated sequence of images would create to the subjects the impression that the scene temporal rate is higher (because of the 'auditory driving' phenomenon) and therefore the duration of the audiovisual clip would seem shorter. We based our hypothesis on the psychological findings discussed: the existence of the *auditory driving* phenomenon and the fact that audition plays a more significant role in temporal rate perception than vision.

We expected that the influence of music will be more profound in the case of fast tempo music, because of all the qualities of music it is tempo that mostly contributes to the

impression of movement [64]. Nevertheless, we decided to investigate the effect of both slow and fast tempo music.

4.2.2 Participants

48 participants (divided into 6 groups) from the University of Bristol, undergraduate and postgraduate student population (their ages ranging from 18 to 35) volunteered to participate in this study. All use computers a great deal in their daily activities. They were randomly assigned to each group. Participants were informed that they could withdraw at any time during the experiments and they were naive as to the purpose of the experiment. They had either normal or corrected-to-normal vision.

4.2.3 Design

An independent samples design was utilised, in which statistically independent samples are drawn from each population and comparative information about the different populations is derived from the analysis of the samples. The samples are considered independent because each participant is tested separately and contributes data to only one of the conditions [23].

The dependent variables were the perceived duration and the scene velocity of each animated sequence (i.e. temporal rate perception), measured retrospectively. The independent variable was the type of music.

During each trial the subjects watched two animations of similar content (navigations in

'virtual' outdoor scenes), one silent and the other with a musical background, and they had to judge which was longer and in which of the two the scene velocity was higher. The subjects were divided into groups so that for some of them the audiovisual animation had an exciting musical background of fast tempo and for the others the music clip used was relaxing and of slow tempo.

There was also another level of subjects' grouping, into those who watched an animation displayed at 16 fps and a second at 20 fps and those who watched an animation displayed at 12 fps and a second at 16 fps, investigating whether the display frame rate would affect the degree of distractive influence of the music. The two animations in every pair were constructed with the same number of frames but were displayed at different rates, without frame dropping, thus resulting in different durations.

There were also two control groups (one for each pair of frame rates), for which both the animations were silent. According to the group to which they were assigned, participants watched the two animations one-after the other in random order, in one of the conditions shown in (Figure 4.2).

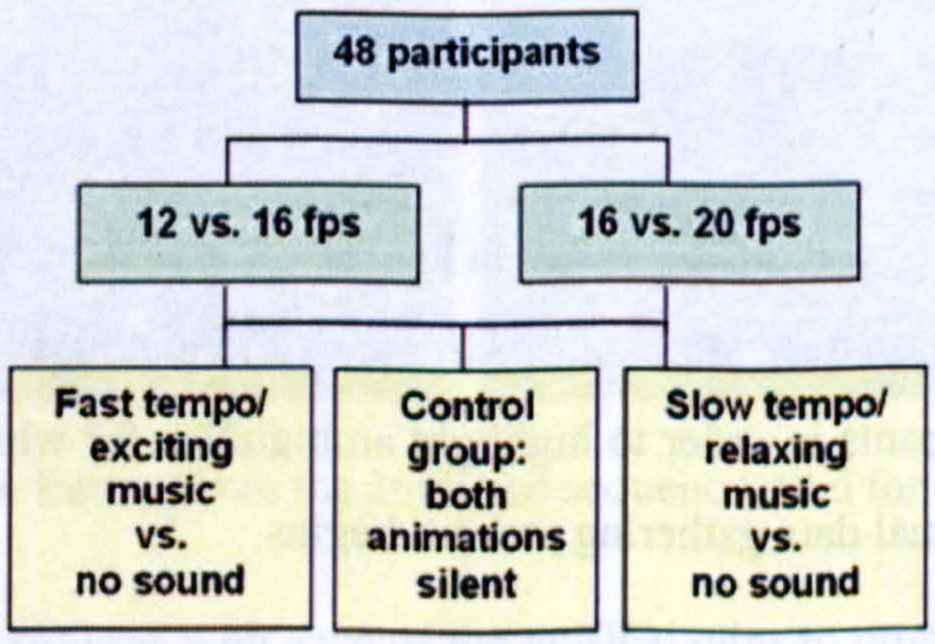


Figure 4.2: Experiment on Temporal Perception - The conditions tested.

Before doing the experiments we had to make an important decision about the experimental procedure to be followed, choosing between the retrospective and prospective

any flaws that could potential bias our results in the main study. Twelve subjects participated, 50% of whom were male. Their ages ranged from 25 to 60. All subjects had normal or corrected-to-normal vision. Subjects came from different professional backgrounds and had little to moderate experience in computers. They were not aware of the purpose of the pilot study. The materials used for the study included one computer-generated animated sequence of images (of duration 1 min 14 sec) and 2 musical clips one of slow tempo and the other of fast tempo (none of them had lyrics). The animated sequence depicted an outdoor 3D scene, which was modelled and rendered with Maya Alias Wavefront Software at 640×480 resolution. The fast tempo clip was an excerpt from "Pont des Arts" (Genre: Electronic-Dance Artist: St. Germain, Album: Tourist (2000)), while the slow tempo clip was an excerpt from "Alfonsina Y El Mar" (Genre: Instrumental Pop/Ethnic Artist: Ocarina, Album: El mejor disco de relajacion (1997)). In Figure 4.3 you can see example frames from the animated sequence we used for the first pilot study.

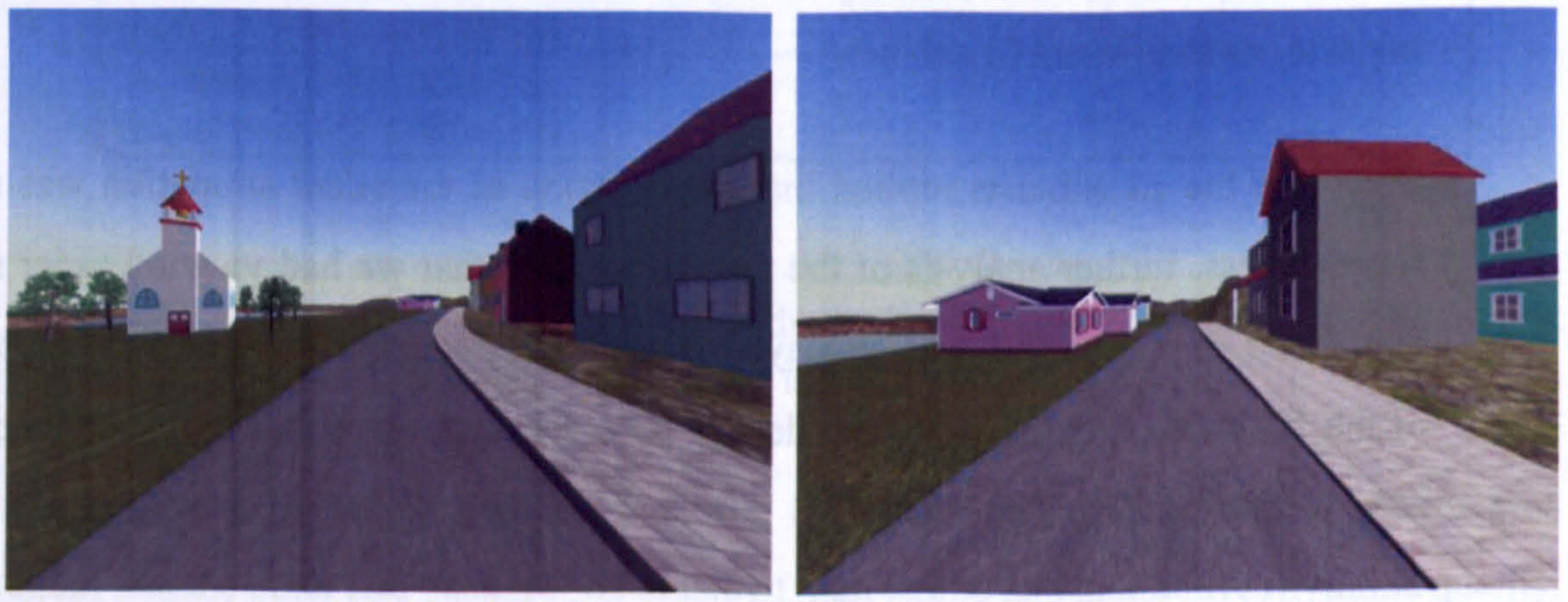


Figure 4.3: Example frames from the animated sequence used for the first pilot study.

The animation was combined with each of the musical clips to produce the three animations used in the experiment: a silent, another with a relaxing musical background (the slow tempo clip) and the third with an exciting musical background (the fast tempo clip). Each subject watched the silent animation and one of the other two in random order and was given, upon completion, a questionnaire. For the display of the animation pairs, Win-

paradigms. In the former, participants do not know in advance about the purpose of the experiments. The disadvantage of this paradigm is that the subjects can be used only once, because after the first retrospective evaluation they are sensitised to duration and temporal rate evaluation and therefore succeeding evaluations can not be considered retrospective. In the prospective paradigm subjects are informed in advance that they will have to judge the durations and rates of animated sequences. Sometimes they are also trained with animations of different durations and/or scene velocities. A disadvantage of this paradigm is that subjects may look for visual cues in the animations to help them decide about their relative durations and scene velocities. In an applied setting of varied display rate, the observers would be preferrably naive about the variations in the temporal rate and therefore the retrospective paradigm was found to be more appropriate for our study.

We conducted three pilot studies to test and refine our experimental design. The outcome of these studies led to us choosing the 4fps difference between the two animations and the actual frame rate pairs (12 & 16 fps and 16 & 20 fps).

Pilot Studies

Pilot trials are common and useful in human-centred experimentation. Experimental procedures such as a questionnaires, instructions and methods may be tested on a small sample of participants in order to highlight ambiguities for which adjustments can be made before the actual data gathering process begins.

Three small pilot studies were designed for the main experimental study described in this chapter. The participants of the pilot studies were different to the population of the main experiment.

The first pilot study was conducted in order to test our experimental design and discover

dows Media Player software was used. All animations were displayed in the centre of the screen with a black background. The subjects were asked to decide which of the two versions (the silent or the audiovisual one) of the animation lasted longer and in which the velocity of the motion was higher. They also had to answer whether they found the music of the audiovisual animation "Pleasant", "Indifferent" or "Unpleasant".

The results of the 1rst pilot study, summarised in Figure 4.4, reveal the following:

Only 2 of the 12 participants realised that the visual part of both the animations they had watched was the same (in terms of duration and motion speed) and only another 2 of the subjects replied that they found the motion faster in the silent animation. The questionnaires revealed that, as expected, the musical clip of high tempo was the one that mostly contributed to the participants' perception that motion was faster in the audiovisual animation. We should note here that the motion speed in the audiovisual animation was perceived as higher even in the cases where the subjects found the music unpleasant (3 subjects) or indifferent (2 subjects).

Only 4 of the 12 subjects replied that the duration of the silent animation was shorter. Nevertheless, further analysis of the results revealed that we had viewing order side effects: The animation that was displayed second was generally perceived as shorter (in 8 of the 12 cases). Psychological models of time perception account for this phenomenon (see, for example, [158]).

After the promising results of the first pilot study, a second study with 26 participants (sharing the same population characteristics with the previous twelve subjects) was conducted to establish the fps difference between the two animations presented to each participant.

In this study we focussed on the subjects' perception of the scene velocity of the animated sequences. To avoid the viewing order effects, different animations of scenes with similar

Results of the 1rst Pilot Study			
	Perceived Duration		
Condition	Silent Animation longer	Audiovisual Animation longer	No difference
Animation with exciting music + silent animation	4	1	1
Animation with relaxing music + silent animation	1	3	2
	Perceived Motion Velocity		
Condition	Silent Animation faster	Audiovisual Animation faster	No difference
Animation with exciting music + silent animation	0	6	0
Animation with relaxing music + silent animation	2	2	2
	Music		
Condition	Pleasant	Indifferent	Unpleasant
Animation with exciting music + silent animation	3	0	3
Animation with relaxing music + silent animation	4	2	0

Figure 4.4: The results of the first pilot study on Temporal Rate Perception. The figures represent numbers of participants who gave the corresponding answers.

contents were displayed in pairs, one after the other. In each pair, one animation was silent and the other had background music of high tempo and distinctive rhythmic quality. We used the fast tempo musical clip that was employed in the previous pilot study. The frame rates of animation pairs under trial were 16 and 12, 12 and 8 fps respectively. The silent animation was in all cases the one of the higher motion speed (although the subjects were ignorant about that). Again, Windows Media Player was used for the display of the two animations in each test pair, in the centre of the screen with a black background.

All the test animations used were based on two sequences of images, originally rendered

at 24 fps (we will call them *Base1* and *Base2*), which were slowed down respectively to get the various versions of the slower display rates (i.e. slower camera motion in the scene). The two corresponding 3D scenes were modelled using Maya Alias Wavefront and were both rendered flat-shaded, at 640 × 480 resolution. This preliminary study was designed to acquire a basic set of data for the simplest rendering. This set of elements will be built up in the main studies of the following chapters to include more photorealistic rendering. Refer to Figure 4.5 for example frames from the *Base1* and *Base2* sequences of images.

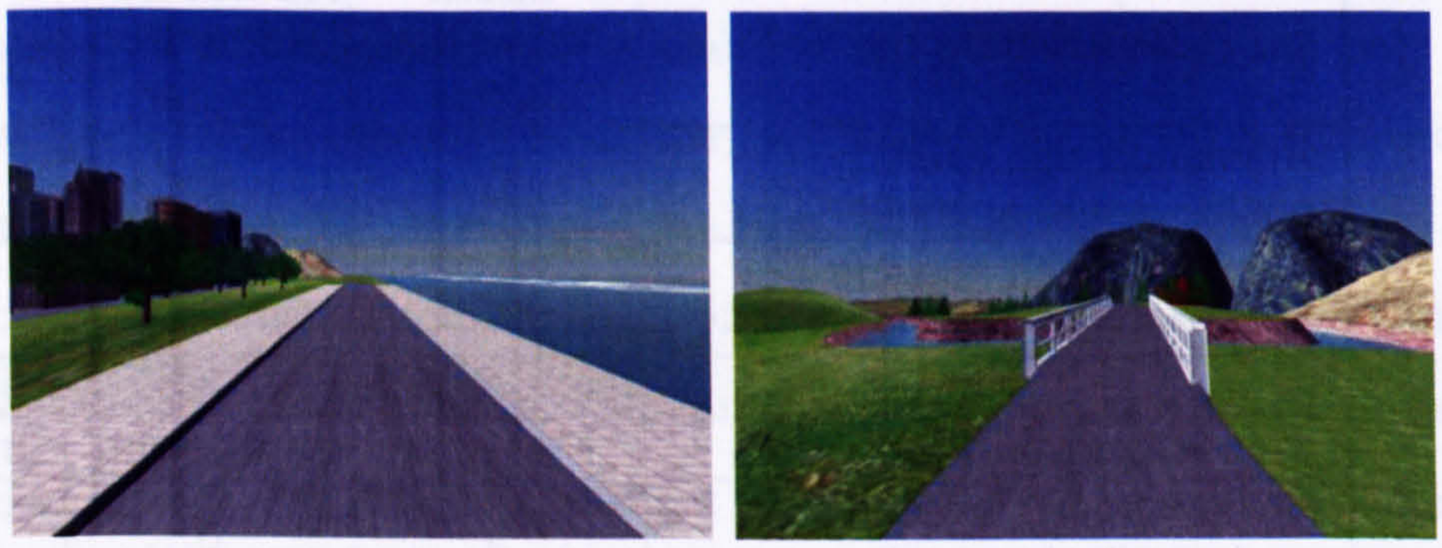


Figure 4.5: Example frames from the Base1 and Base2 animated sequences used for the second pilot study and the main experiment.

In the *Base1* and *Base2* walkthroughs the camera was moving at exactly the same speed (i.e. same scene velocity) and followed exactly the same motion path, so that Base1 and Base2 had exactly the same duration and the same motion characteristics. For the manipulation of the animated image sequences we used Adobe Premier software. Each of the two initial sequences was imported into Premier and its duration was 'stretched' according to the desired percentage, in order to produce The maximum duration of the resulting walkthroughs (at 8 fps) was 1 minute.

The results of the second pilot study are summarised in Figure 4.6. All subjects tested at the '8 fps with music- 12 fps silent' pair of animations easily distinguished that the silent animation had a higher scene velocity. The results from the '12fps with music- 16fps silent' animation pairs were far more promising: Only 2 of the 13 subjects recognised

correctly the animation with the faster motion, i.e. the silent animation.

Perceptual Results	12fps + music vs. 16fps silent	8fps + music vs. 12fps silent
The slower animation perceived as faster	7	0
Faster animation was recognised correctly	2	13
No perceived difference in motion speed	4	0

Figure 4.6: Results of the second pilot study on Temporal Rate Perception. The figures represent numbers of participants who gave the corresponding answers.

A third pilot study with 3 participants was conducted to determine the two music clips that would be used for the main experiment. The candidates for the *fast tempo, exciting* music clip were excerpts from rap musical pieces, which were rated in terms of fast tempo and highly rhythmic quality (in a scale from 1 to 5). "Rap music is a form of pop music based on chanted rhymes accompanied by a thumping rhythmic backbeat" [57]. It was chosen because, especially when played at a very fast tempo, it can induce high arousal (excitement) to people due to its highly rhythmic nature [155]. The rap music clips rated by the 3 subjects were excerpts (without lyrics) from the following songs:

- Ice Ice, Baby Artist: Vanilla Ice, Album: To the Extreme (1990).
- It's a Party, Artist: Vanilla Ice, Album: To the Extreme (1990).
- Get Loose, Artist: Vanilla Ice, Album: Mind Blowin (1994).
- Living, Artist: Vanilla Ice, Album: Hard to Swallow (1998).
- Freestyle, Artist: Vanilla Ice, Album: Hard to Swallow (1998).
- '93 til Infinity, Artist: Souls of Mischief, Album: '93 til Infinity (1993).

The candidates for the *slow tempo, relaxing* music clip were excerpts (without lyrics) from music pieces used for relaxation and meditation and were rated in terms of slow tempo and relaxing quality (in a scale from 1 to 5). More specifically, the relaxing excerpts came from the following tunes:

- Return to the Garden, Artist: Dave Abbott and Dan O'Brien, Album: Sanctuary (2002).
- Beautiful Air, Artist: Dave Abbott and Dan O'Brien, Album: Sanctuary (2002).
- Windmills, Artist: Robert Jennings, Album: Invitation (2000).
- Silhouette, Artist: Robert Jennings, Album: Invitation (2000).
- Lullaby, Artist: Robert Jennings, Album: Invitation (2000).
- Twilight, Artist: Andrew Kim , Album: The Journey (2002).

The music clip which received the highest ratings for its high tempo and rhythmic quality was the excerpt from "Get Loose". The clip that was rated highest for its slow tempo and relaxing quality was the excerpt from "Windmills". These two excerpts were employed in the main experiment.

4.2.4 Equipment and Materials

The test environment comprised a PC placed on a desk in an empty room (so that the subjects would not be distracted by surrounding objects). The subjects watched the animations on the 19" typical desktop monitor of the PC and listened to the music through headphones in order to be isolated from outside noise. For the display of the animations,

Windows Media Player software was used. As in the pilot studies, the animations were displayed in the centre of the screen with a black background.

Two *Base1* and *Base2* animated sequences of images and the two musical clips which resulted from the 3rd pilot study, one relaxing (of very slow tempo) and the other exciting (of very fast tempo and distinctive rhythmic quality), were combined to produce the animation pairs used in the experiment. Again the two animated sequences of images were slowed down accordingly in Adobe Premier, to give the required display rates for our experimental versions (no frames were dropped).

Since music did not affect at all the observers' perception of temporal rate in the "8 vs. 12 fps" condition of the second pilot study, we decided to focus on higher frame rates in our main experiment, for which the frame rates of the animation pairs were decided to be 16 and 12, 20 and 16 fps, respectively.

4.2.5 Procedure

Each participant was tested individually and each experimental session lasted for 10-15 minutes. Participants were informed that they should watch carefully two computer-generated walkthrough animations, one silent and one accompanied by music (except for the control group that watched two silent animations), in random order, and when finished they would have to answer some questions. Subjects had no reason to anticipate time or temporal rate estimation questions.

In the questionnaire the subjects were asked which of the two versions of the animation lasted longer and in which the scene velocity was higher. They also had to comment on the music: whether they liked it and how much, whether they found it relaxing or exciting and whether it was familiar to them. The options they had were: Not at all / A little bit /

Moderately / Very much. The questionnaire also contained a question about the relaxing quality of the overall watching experience. The options ranged from Not at all/ A little bit / Moderately relaxing/ Very relaxing. The full questionnaire given to the participants of this experiment can be found in Appendix A.

4.3 Results

The results of the experiment on Temporal Perception are summarised in Figures 4.7, for the relative perceived duration, and 4.8, for the relative perceived motion velocity. As "wrong" we considered both the answers that incorrectly identified the longer or faster animation within the test pair of animations and also the answers of the participants who found no difference in the duration or motion velocity between the two animations.

The data were initially analysed by carrying out t-tests between the means of our independent groups, pairing every time the control group with the corresponding 'fast tempo-exciting' or 'slow tempo- relaxing' group. The t-test for independent samples is a powerful tool for testing significance between two independent groups when participants are tested only once [23]. It compares the means of two samples in order to determine whether the difference between them is sufficiently great to be unlikely to have occurred by chance (at a chosen level of probability).

Refer to Figures 4.9- 4.12 for the results of the 8 t-tests. In our case, perhaps surprisingly, no statistically significant effect of music (either fast or slow tempo) was revealed at the 0.05 level of risk. The observed values for t were smaller than 1.762 (i.e. the critical value for t in our case), so our initial hypothesis could not be confirmed.

Perceived Duration			
	12 fps vs. 16 fps		
Condition	**CORRECT ANSWERS: 12 fps longer**	**16 fps longer**	**No difference**
12 fps with exciting music + 16 fps silent	3	0	5
12 fps with relaxing music + 16 fps silent	5	2	1
control condition (both animations silent)	6	1	1
	16 fps vs. 20 fps		
Condition	**CORRECT ANSWERS: 16 fps longer**	**20 fps longer**	**No difference**
16 fps with exciting music + 20 fps silent	5	1	2
16 fps with relaxing music + 20fps silent	5	2	1
control condition (both animations silent)	2	4	2

Figure 4.7: Experiment on Temporal Perception - Results for the relative perceived duration across conditions. The figures represent numbers of participants who gave the corresponding answers.

Perceived Camera motion			
	12 fps vs. 16 fps		
Condition	**12 fps faster**	**CORRECT ANSWERS: 16 fps faster**	**No difference**
12 fps with exciting music + 16 fps silent	2	3	3
12 fps with relaxing music + 16 fps silent	0	6	2
control condition (both animations silent)	1	4	3
	16 fps vs. 20 fps		
Condition	**16 fps faster**	**CORRECT ANSWERS: 20 fps faster**	**No difference**
16 fps with exciting music + 20 fps silent	2	4	2
16 fps with relaxing music + 20 fps silent	1	5	2
control condition (both animations silent)	3	2	3

Figure 4.8: Experiment on Temporal Perception - Results for the relative perceived motion velocity across conditions. The figures represent numbers of participants who gave the corresponding answers.

t-test 1 Perceived Duration: EXCITING MUSIC (animations 12 vs 16 fps)		
Values	X_a (both silent animations)	X_b (12 fps exciting music + 16 fps silent)
n	8	8
mean	0.75	0.375
t = +1.53	df = 14	$P_{one\text{-}tailed}$ = 0.0741475
		$P_{two\text{-}tailed}$ = 0.148295

t-test 2 Perceived Duration: EXCITING MUSIC (animations 16 vs 20 fps)		
Values	X_a (both silent animations)	X_b (16 fps exciting music + 20 fps silent)
n	8	8
mean	0.25	0.625
t = -1.53	df = 14	$P_{one\text{-}tailed}$ = 0.0741475
		$P_{two\text{-}tailed}$ = 0.148295

Figure 4.9: Experiment on Temporal Perception - t-test results for the perceived duration regarding the "Exciting music" groups.

t-test 3 Perceived Duration: RELAXING MUSIC (animations 12 vs 16 fps)		
Values	X_a (both silent animations)	X_b (12 fps relaxing music + 16 fps silent)
n	8	8
mean	0.75	0.625
t = +0.51	df = 14	$P_{one\text{-}tailed}$ = 0.3089985
		$P_{two\text{-}tailed}$ = 0.617997

t-test 4 Perceived Duration: RELAXING MUSIC (animations 16 vs 20 fps)		
Values	X_a (both silent animations)	X_b (16 fps relaxing music + 20 fps silent)
n	8	8
mean	0.25	0.625
t = -1.53	df = 14	$P_{one\text{-}tailed}$ = 0.0741475
		$P_{two\text{-}tailed}$ = 0.148295

Figure 4.10: Experiment on Temporal Perception - t-test results for the perceived duration regarding the "Relaxing music" groups.

t-test 5 Perceived Motion: EXCITING MUSIC (animations 12 vs 16 fps)		
Values	X_a (both silent animations)	X_b (12 fps exciting music + 16 fps silent)
n	8	8
mean	0.125	0.25
t = -0.61	df = 14	$P_{one\text{-}tailed}$ = 0.2758175
		$P_{two\text{-}tailed}$ = 0.551635

t-test 6 Perceived Motion: EXCITING MUSIC (animations 16 vs 20 fps)		
Values	X_a (both silent animations)	X_b (16 fps exciting music + 20 fps silent)
n	8	8
mean	0.375	0.25
t = +0.51	df = 14	$P_{one\text{-}tailed}$ = 0.3089985
		$P_{two\text{-}tailed}$ = 0.617997

Figure 4.11: Experiment on Temporal Perception - t-test results for the perceived motion velocity regarding the "Exciting music" groups.

t-test 7 Perceived Motion: RELAXING MUSIC (animations 12 vs 16 fps)		
Values	X_a (both silent animations)	X_b (12 fps relaxing music + 16 fps silent)
n	8	8
mean	0. 5	0.75
t = -1	df = 14	$P_{one\text{-}tailed}$ = 0.167141
		$P_{two\text{-}tailed}$ = 0.334282

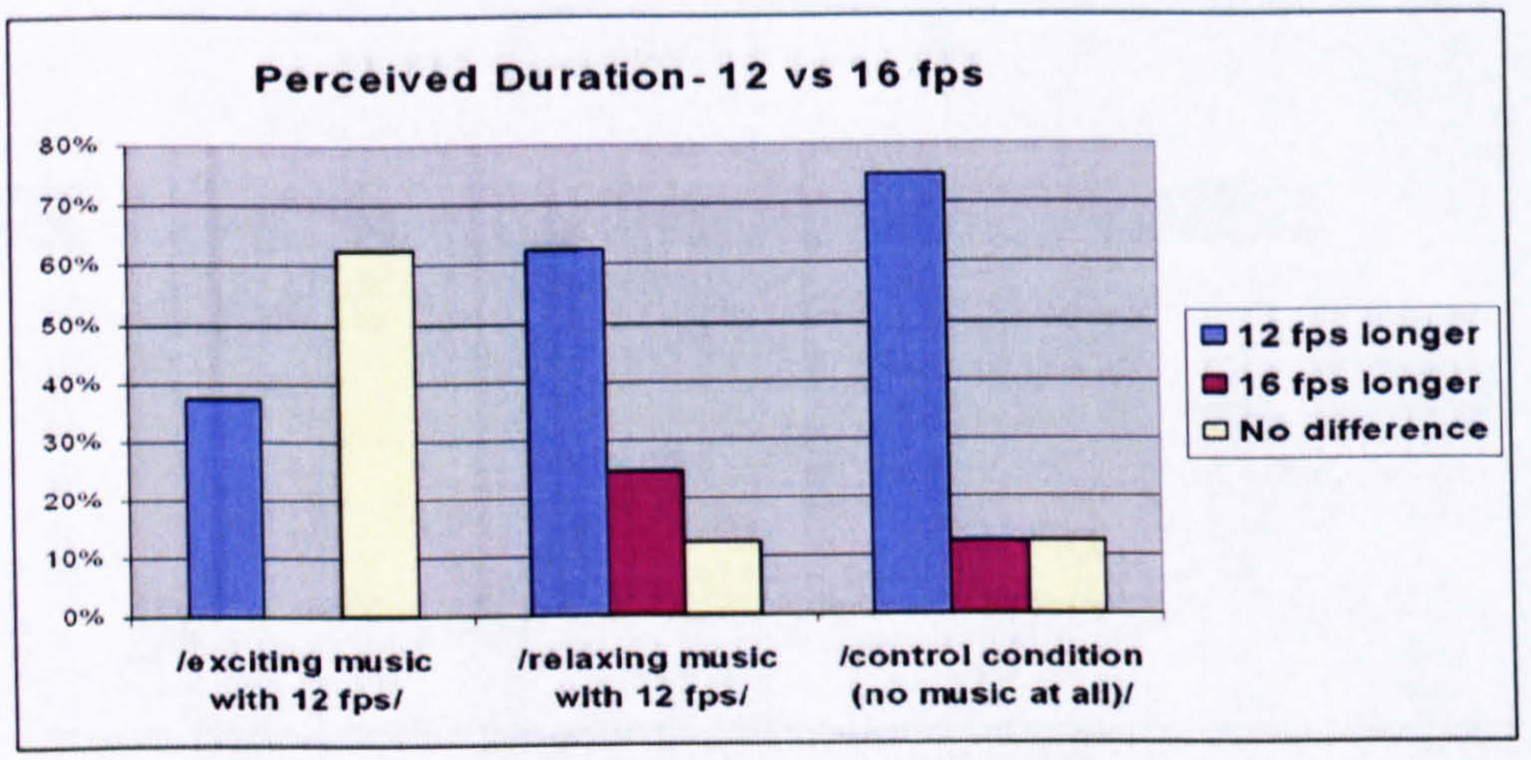
t-test 8 Perceived Motion: RELAXING MUSIC (animations 16 vs 20 fps)		
Values	X_a (both silent animations)	X_b (16 fps relaxing music + 20 fps silent)
n	8	8
mean	0.25	0.625
t = -1.53	df = 14	$P_{one\text{-}tailed}$ = 0.0741475
		$P_{two\text{-}tailed}$ = 0.1482950

Figure 4.12: Experiment on Temporal Perception - t-test results for the perceived motion velocity regarding the "Relaxing music" groups.

12 fps vs. 16 fps			
	Perceived Duration		
Condition	12 fps longer	16 fps longer	No difference
12 fps with exciting music	37,5%	0%	62,5%
12 fps with relaxing music	62,5%	25%	12,5%
control condition (no music at all)	75%	12,5%	12,5%
	Perceived Camera motion		
Condition	12 fps faster	16 fps faster	No difference
12 fps with exciting music	25%	37,5%	37,5%
12 fps with relaxing music	0%	75%	25%
control condition (no music at all)	12,5%	50%	37,5%

16 fps vs. 20 fps			
	Perceived Duration		
Condition	16 fps longer	20 fps longer	No difference
16 fps with exciting music	62,5%	12,5%	25%
16 fps with relaxing music	62,5%	25%	12,5%
control condition (no music at all)	25%	50%	25%
	Perceived Camera motion		
Condition	16 fps faster	20 fps faster	No difference
16 fps with exciting music	25%	50%	25%
16 fps with relaxing music	12,5%	62,5%	25%
control condition (no music at all)	37,5%	25%	37,5%

Figure 4.13: Experiment on Temporal Perception - Results in percentages for the '12 vs. 16 fps' (top) and the '16 vs. 20 fps' (bottom) conditions.



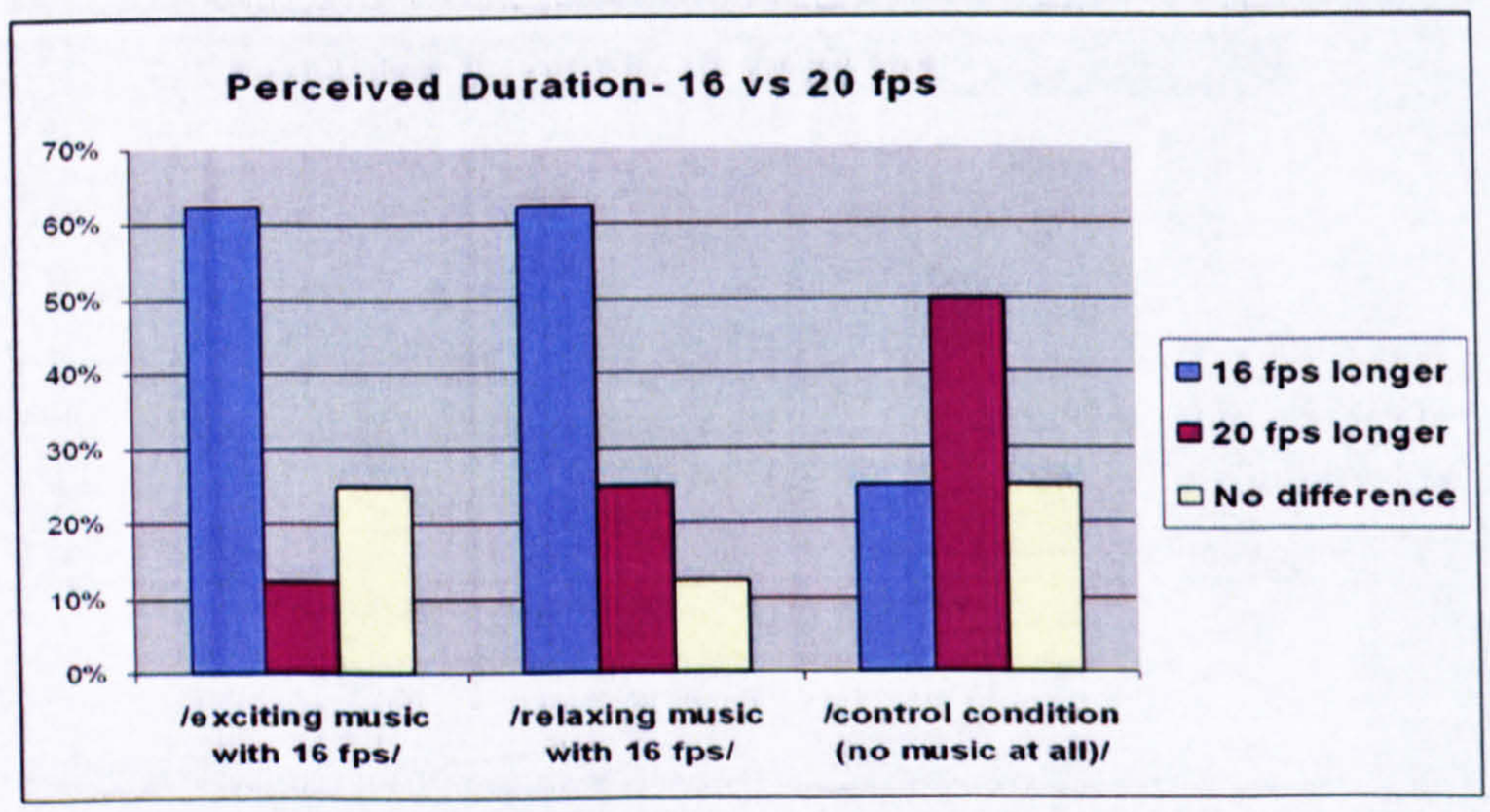
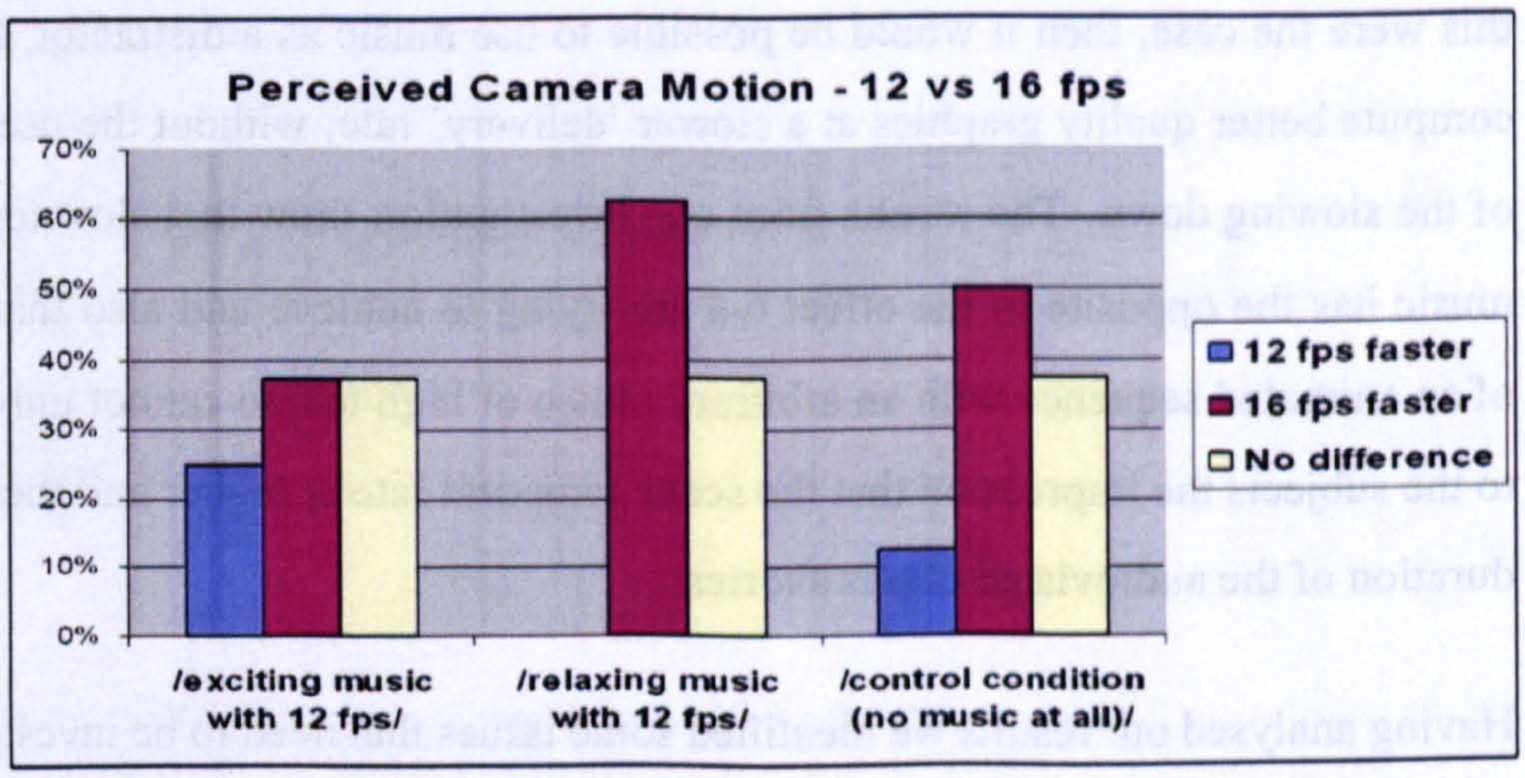


Figure 4.14: Experiment on Temporal Perception - Graphs of the Perceived Duration across conditions



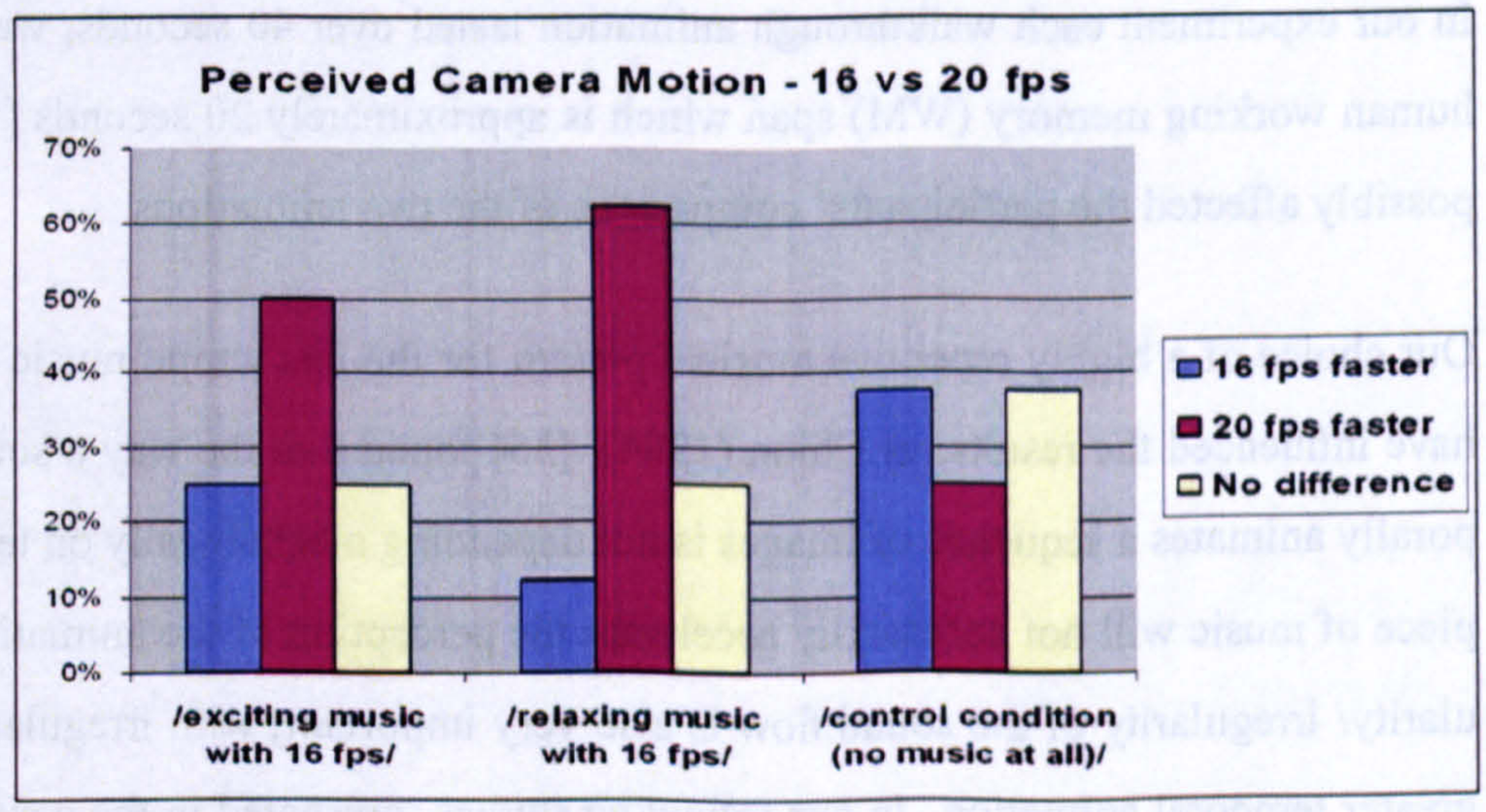


Figure 4.15: Experiment on Temporal Perception - Graphs of the Perceived Camera Motion across conditions

We then went on to analyse our results using frequencies (percentages), because Frequency Analysis can provide informative data on the levels of observed differences between participants. The results in percentages are given in Figure 4.13 and graphically in Figures 4.14 and 4.15. These results indicate the following:

- Relaxing music has the effect of decreasing the perceived scene velocity, but not significantly.

- Even in the control conditions (especially for the '16 fps vs. 20 fps' conditions) participants were not always able to distinguish the longer clip or the clip with the higher scene velocity.

- Regarding the perception of time duration, an unexpected effect of viewing order was found, although the visual parts of the two animations each participant watched were not the same.

The questionnaires confirmed that the music excerpts matched their intended emotional suggestiveness, i.e. all the subjects who listened to the fast tempo musical clip during their experimental task rated it as "Very Exciting" and all the subjects who listened to the slow tempo clip rated it as "Very relaxing".

4.4 Discussion

In this study we investigated whether the combination of tempo and emotional suggestiveness of music would affect the users' perception of temporal rate and duration. If this were the case, then it would be possible to use music as a distractor, allowing us to compute better quality graphics at a slower 'delivery' rate, without the user's perception of the slowing down. The results from our investigation show that slow tempo- relaxing music has the opposite to the effect we are trying to achieve and also that the coupling of an animated sequence with an arbitrary music of high tempo cannot universally create to the subjects the impression that the scene temporal rate is higher and therefore that the duration of the audiovisual clip is shorter.

Having analysed our results we identified some issues that need to be investigated further.

In our experiment each walkthrough animation lasted over 40 seconds, well beyond the human working memory (WM) span which is approximately 20 seconds [165], and this possibly affected the participants' comparison of the two animations.

Our choice of a highly repetitive musical pattern for the fast tempo music clip may also have influenced the results, as Chion (1994) [35] found that the way a soundtrack temporally animates a sequence of images is not depending mechanically on tempo. A rapid piece of music will not necessarily accelerate the perception of the animation as the regularity/ irregularity of the sound flow is also very important, with irregularity favouring greater temporal animation. In our following studies, presented in the next chapters, we experimented with more types of distractors and different characteristics of the distracting stimuli (e.g. congruity and familiarity).

As we mentioned in the introductory paragraph of the chapter, the experimental design employed was not very strict, as this was an initial exploratory study. Our goal was to build on the preliminary study by dealing with its design flaws and by addressing the

issues raised above, in order to develop a formal experimental framework. This framework was subsequently applied to our major studies, presented in the next chapters, which would investigate the perception of both temporal (i.e. frame rate) and visual characteristics (i.e. rendering quality perception) of a computer graphics environment.

Chapter 5

Perceived Frame Rate under the Influence of Music and Sound Effects

5.1 Introduction

The studies described in this chapter were inspired by the research findings on crossmodal perception presented in the background chapters 2 and 3. Reducing the frame rendering rates increases the available computational resources. The most immediate use one can make of these resources is to increase the rendered visual detail. Nevertheless, in our studies frame rate manipulations are examined independently without increased rendering quality, in order to pull apart the effects of reducing frame rates and increasing visual detail.

The results of the experiments that we conducted confirm that in the presence of audio stimuli, and more specifically sound effects, viewers fail to notice variations in the motion smoothness between walkthrough animations displayed at different rates, which are apparent in the absence of sound (Figure 5.1). As discussed in sections 3.6.1 and 3.6.2,

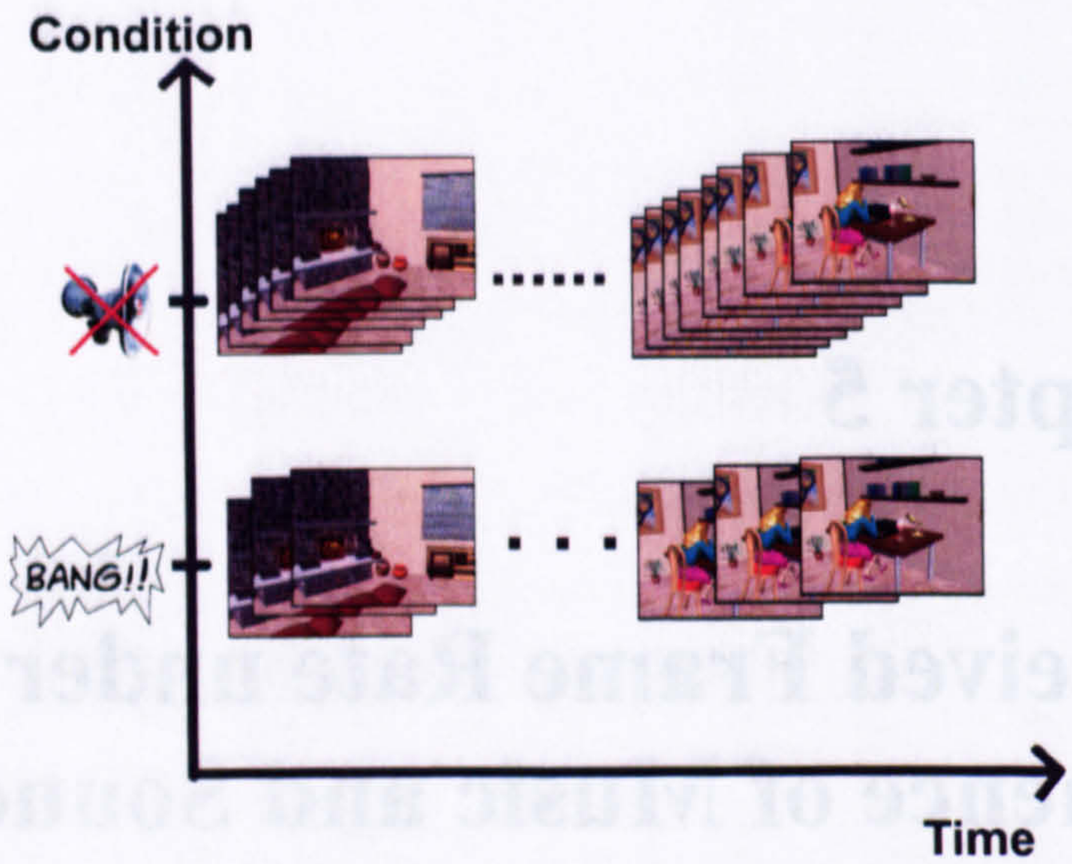


Figure 5.1: With the use of auditory 'distractors', such as sound effects, fewer frames may be displayed per second compared to a silent animation, without any noticeable difference in the motion smoothness.

this is probably due to the fact that the auditory stimuli attract part of viewer's attention to the sound and away from the visual defects, such as jerky motion, which result from low frame rates.

This chapter outlines the experimental methodology employed and the relevant results of the first two of the main studies in this thesis which examined the perceptibility of motion jerkiness resulting from reduced frame rates, under the influence of music and sound effects. Part of the work presented in this chapter has been published in [128]. The preliminary study described in the previous Chapter demonstrated the potential of our methodology, however, it also highlighted the shortcomings of the informal experimental design. The actions taken to remedy those shortcomings are described in this chapter in the context of the larger studies, presented in detail in sections 5.2-5.3 and their subsections. The chapter finishes with a discussion on the findings of the two studies, the conclusions that could be drawn and also suggestions about how this work could be taken further.

5.2 Experiment 1 - The Influence of Sound Effects and Music on the Perceived Smoothness of Rendered Animations

We prepared and conducted a second experiment to investigate the influence music and sound effects have on the perception of frame rate and more specifically on the perceived smoothness of rendered animations.

From the research in multimodal perception presented in previous chapters we can infer that:

1. The redirection of attention and the allocation of cognitive resources to the processing of music or sound effects while watching rendered animations, may reduce the viewer's cognitive resources allocated to the processing of the visual cues employed in judging the motion smoothness / jerkiness of an animation (refer to section 3.6.1).

2. Especially music of high tempo and exciting impression might increase subjects' arousal, which has been found to impose capacity limitations in information processing (see section 3.4.1).

Based on the perception findings we hypothesised that it would be more difficult for subjects to distinguish frame rate differences between audiovisual composites than between silent animations. Three conditions were considered: "Music", "Sound Effects" and the control, "No Sound", group. We also hypothesised that familiarity with animated computer graphics would help the corresponding subjects perform the experimental task more efficiently than participants without any prior experience.

Participants viewed pairs of animations with the same visual content displayed at seven

different frame rates. The animations were either silent or accompanied by music or sound effects. The pairs were displayed in random order and the participant had to decide which of the two animations was displayed at a higher frame rate (i.e. which had a smoother motion).

In the following sections, the experimental methodology of this experiment is described and the results are presented in detail. Experimental design issues such as participants, apparatus and materials are analysed concentrating on extending the preliminary study's (presented in the previous chapter) procedures and methodology.

5.2.1 Participants

Forty eight participants from the undergraduate and postgraduate student population volunteered to participate in this study. Ages ranged from 18 to 41, with an average age of 24. The participants were divided into two groups according to their familiarity with animated computer graphics before performing the experimental tasks, as we also wanted to investigate whether the user's experience would affect the degree of the 'distractive' influence of the sound. The first group consisted of subjects who were moderately/very familiar and the second group included the ones with no or little experience in animated computer graphics. The members of each of these two groups were randomly subdivided across the three conditions. Participants were provided with instructions and were informed that they could withdraw at any time during the experiment and they were naive as to its purpose. They all had either normal or corrected-to-normal vision and they did not report any hearing impairment.

5.2.2 Design

An independent samples design was utilised, in which statistically independent samples are drawn from each population and comparative information about the different populations is derived from the analysis of the samples. The samples are considered independent because each participant is tested separately and contributes data to only one of the conditions [23]. The dependent variables were the perceived relative motion smoothness of the two animated sequences in each trial pair of sequences (i.e. relative frame rate perception). The independent variable was the auditory background of the movies (sound effects, music or silence). A subject watched two animations and then gave his/her judgement. All paired comparisons were randomly generated. Seven different frame rates were considered: 20 fps, 15 fps, 12 fps, 10 fps, 8 fps, 6 fps and 4 fps. In addition, seven of the pairs were at the same frame rate, giving a total of 49 trial pairs. Frame rates of 24 and 30 fps were excluded as previous studies have shown that frame rates from 24 fps and above are indistinguishable for human viewers.

To reduce the degree of boredom which may result from watching the same visual stimulus over and over again, 7 different musical clips (all of high tempo and exciting quality) and 7 different sound effects clips were used, although each pair, of course, had the same auditory stimuli. See next section for more details on the musical excerpts and the sound effects employed in our experiment. Each animation and sound file was 8 seconds in length, so that a paired comparisons task could be completed by subjects within the limits of the human working memory span, which is approximately 20 seconds [165]. The conditions tested are shown in Figure 5.2.

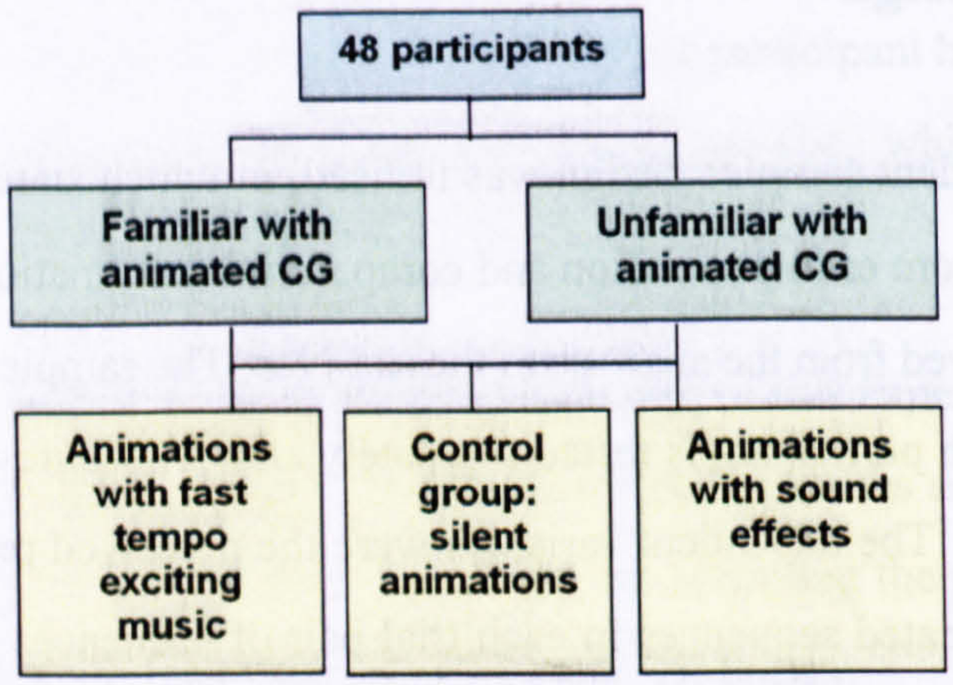


Figure 5.2: Frame Rate Experiment 1 on the influence of sound effects and music on the perceived smoothness of rendered animations - The Conditions tested.

5.2.3 Equipment and materials

The test environment comprised a PC placed on a desk in an empty room (so that the subjects would not be distracted by surrounding objects). The subjects watched the animations on the 19" CRT monitor of the PC and listened to the music/sound effects through quality headphones, with frequency response 18 - 22 KHz, isolated from outside noise. For the display of the animations, Windows Media Player software was used.

The animated sequence of images used to produce the test animations comprised a 3D interior scene: a room with a fireplace, some furniture and a woman sitting at a desk typing using a typewriter. The computer graphics representation of the 3D scene was created using the Maya Alias Wavefront modeling package and was rendered at 640×480 pixels resolution. For example frames of the animation refer to Figure 5.3.

During the animation, the view field of the camera remains static for two seconds (attending to the moving flames in the fireplace), then it "pans" across a part of the room for another four seconds and finally zooms to the sitting woman for the last two seconds. This type of camera motion is used because it accentuates the motion jerkiness caused by

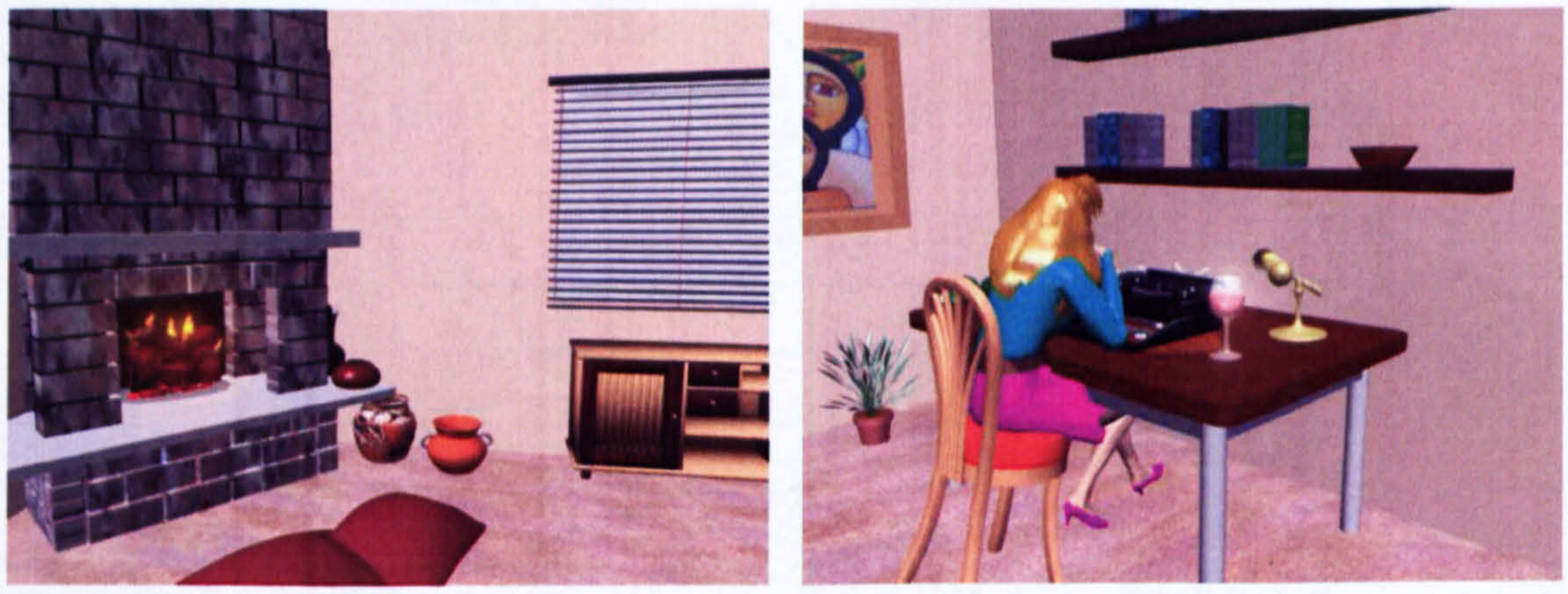


Figure 5.3: Frame Rate Experiment 1 - Example frames from the animated sequence used for the experiment.

reduced frame rates more than a "walk through" type of movement. This was confirmed by a pilot study in which subjects saw pairs of animations rendered at the same frame rate, one of which was a "walk through" and the other a "pan". Participants consistently perceived the motion jerkiness as more intense in the "panned" animations.

No compression was applied to our animated sequences to avoid various visual defects that appear as a result of video compression encoders.

Because of the visual content of our 3D scene, the sound effects for the "Sound Effects" condition were: fire crackling, typewriter, telephone ringing (twice), door bell ringing (twice), door slam, thunder, liquid pouring, cat meowing (twice), female coughing (twice). The fire crackling and typewriter were audible through out each movie (with varying intensity according to the position of the camera in the virtual space), and the other sound effects (only one was included in each animation) lasted approximately for 2 seconds in the middle of the scene traversal, see Figure 5.4.

Fast tempo music was used for the "Music" condition, because, as discussed in the previous Chapter, it can induce high arousal in listeners, which then influences performance on various cognitive skills. No lyrics were included. Some of these songs were at the time of the experiment very popular among this particular age group (according to the British

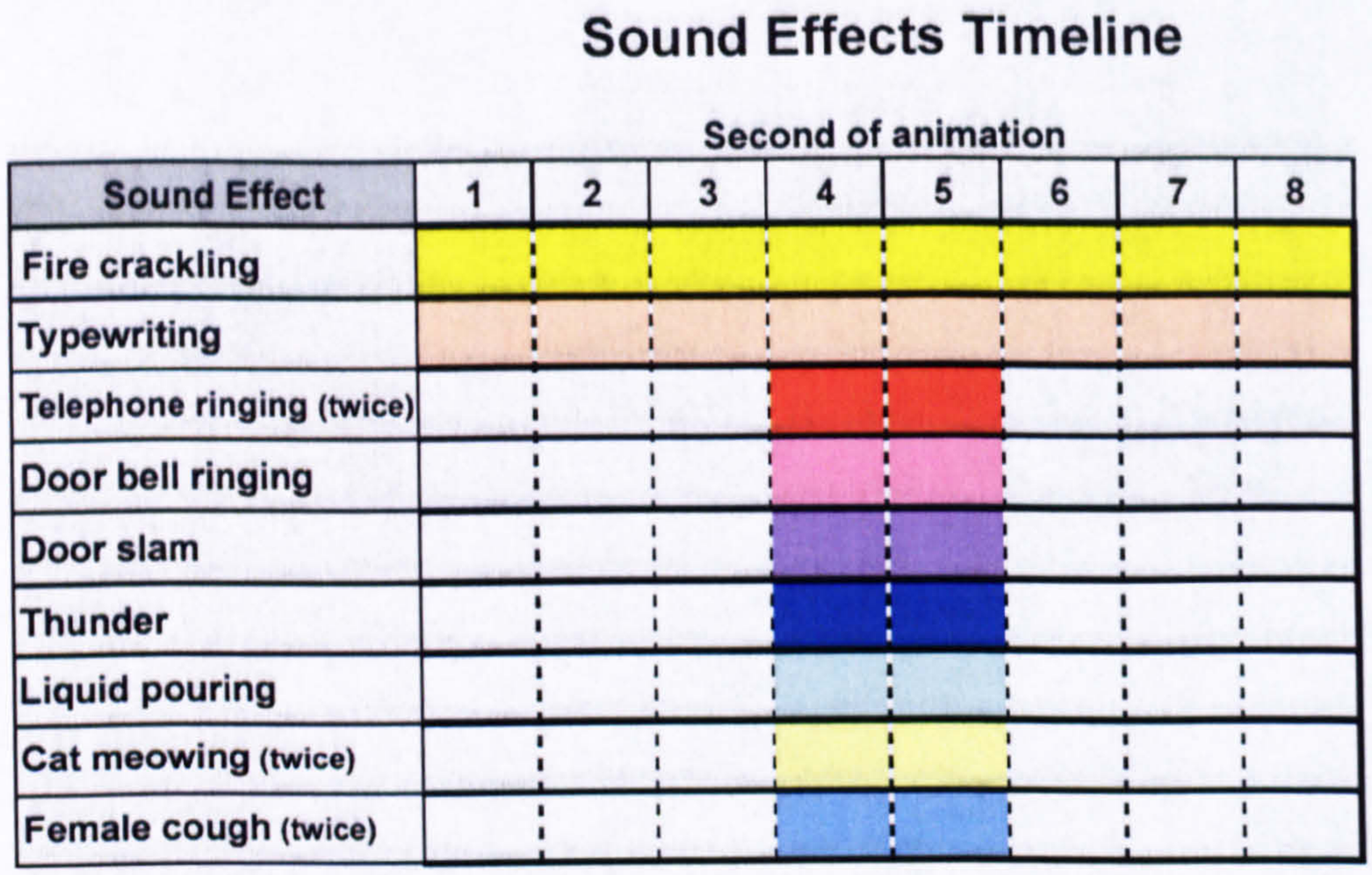


Figure 5.4: Frame Rate Experiment 1 - Timeline depicting the period during the 8-second animation that each sound effect was audible.

Music Charts)- and thus probably familiar to the participants- and the rest were picked in order to represent the unfamiliar music counterpart. This way, any biasing of the results due to music familiarity would be counterbalanced across the various animation pairs. A pre-study with 6 subjects (three of whom were male) confirmed our choice. They all listened to each of the music excerpts and had to answer whether they were "Not at all", "A little bit", "Moderately" or "Very much" familiar with it and also whether they found that particular tune exciting (as opposed to relaxing), using the same scale. They all found four of the music excerpts "Moderately" or "Very much" familiar and the rest three tunes "Not at all" or "A little bit" familiar. All pre-study participants rated the seven tunes as "Very much" exciting. The four familiar music clips were extracted from the following songs:

- Lucky Star, Artist: Basement Jaxx and Dizzee Rascal, Album: Kish Kash (2003).
- Me against the music, Artist: Britney Spears and Madonna, Album: In the zone

(2003).

- Superstar, Artist: Jamelia, Album: Thank You (2003).
- Turn me ..., Artist: Kevin Lyttle, Album: Kevin Lyttle (2004).

The three unfamiliar tunes were extracted from:

- Once in a Lifetime, Artist: Wolfsheim, Album: Spectators (1999).
- Pont des Arts, Artist: St. Germain, Album: Tourist (2000).
- Trust You, Artist: Mesh, Album: Fragmente (1998).

5.2.4 Procedure

Each participant was tested individually. Participants were informed that they should watch carefully 49 pairs of computer-generated animations and that all animations would have the same visual content. When each pair finished they would have six seconds before the next pair loaded to answer the question: "Which of the two movies in the trial pair you just watched do you think had a better visual quality taking into consideration the motion smoothness or on the contrary jerkiness?" For each pair, the participants could select one of: "The first was better", "The second was better", "Cannot tell which was better" and "The two movies were identical". They were instructed that they should pick the third option in the cases where they could perceive some difference between the motion smoothness of the two movies, but they could not decide which was better. A count down was displayed between the showings of the pairs so the participants knew exactly how much time was left for them to give their answer. During this six-second countdown there was silence for all conditions.

The participants who were familiar with animated computer imagery were told that the variable under investigation was the perception of the displayed frame rate and were additionally informed that the rendering quality parameters, such as image resolution, antialiasing, etc. were the same for all movies (so that they did not focus on parameters other than frame rate). The others with little or no experience in animated computer graphics were provided with a very simple introduction to the concept of the frame rate, so they could understand the cause of jerkiness that would be displayed at low frame rates and also to be able to complete their task as efficiently as possible.

The subjects in the "Music group" and the "Sound Effects group" were told that the movies would be accompanied by music and sound effects respectively, and the sound would be delivered to them through headphones. Even participants in the "No Sound" group had the headphones on throughout the experimental session, so as to be better isolated from outside noise.

Before the actual experimental task, all participants received a familiarisation phase, during which they watched a training sample that consisted of sample pairs of silent movies (divided by the countdown periods) of varying frame rate difference within each pair. The training sample was played as many times as each participant wished and during its playback he/she received instructions from the experimenter about the visual cues that would help him/her to distinguish the frame rate differences.

Each experimental session lasted 30-35 minutes. To avoid fatigue or boredom from watching the same animated sequence over and over again, the subjects were instructed that they could pause the display during any countdown interval and continue when they felt ready again. 3 (out of the 48) participants chose to pause the experiment and this happened only once per session.

Each questionnaire concluded with two questions about how relaxed/comfortable the participant felt and how focused he/she was while watching the animations: a) at the begin-

ning and b) towards the end. They could select one of the following options: Not at all / A little bit / Moderately / Very much. These questions were used to check whether any change in the focus or comfort levels would affect the subjects' performance.

5.2.5 Results

Figures 5.5 and 5.6 give each participant's number of correct answers (out of 49), separately for the three conditions and Familiar/Unfamiliar subjects.

	Number of correct answers per subject (in a total of 49 answers)		
Familiar Subject No	**No Sound**	**Sound Effects**	**Music**
1	35	27	35
2	36	31	30
3	33	28	28
4	34	34	35
5	34	39	31
6	39	29	34
7	32	27	42
8	35	31	34

Figure 5.5: Frame Rate Experiment 1 - The results for the Familiar subjects across the 3 conditions, given as the number of correct answers in the total of 49 rate pairs.

	Number of correct answers per subject (in a total of 49 answers)		
Unfamiliar Subject No	**No Sound**	**Sound Effects**	**Music**
1	24	21	17
2	23	13	37
3	28	18	23
4	28	26	28
5	35	16	27
6	23	23	23
7	37	31	34
8	18	22	17

Figure 5.6: Frame Rate Experiment 1 - The results for the Unfamiliar subjects across the 3 conditions, given as the number of correct answers in the total of 49 rate pairs.

Measure of performance in our experimental task was the percentage of times each subject

correctly identified the faster frame rate within a pair of displayed animations. The performance was averaged for each pair of frame rates across all subjects within each group. For example, a performance of 100% for a pair of frame rates within a group indicates that all subjects from this group correctly distinguished the higher frame rate whenever they came across the corresponding pair of rates while performing the experimental task. Since we were considering Familiarity/ Unfamiliarity with animations coupled with the 3 conditions for the auditory background of the animations (No Sound, Music and Sound Effects), we calculated 6 different performance means for each pair of employed frame rates. Figures 5.7- 5.9 illustrate the performance of Familiar versus Unfamiliar participants within each condition for each frame rate pair and Figures 5.10- 5.11 compare the performances measured across the 3 conditions (separately for Familiar and Unfamiliar subjects).

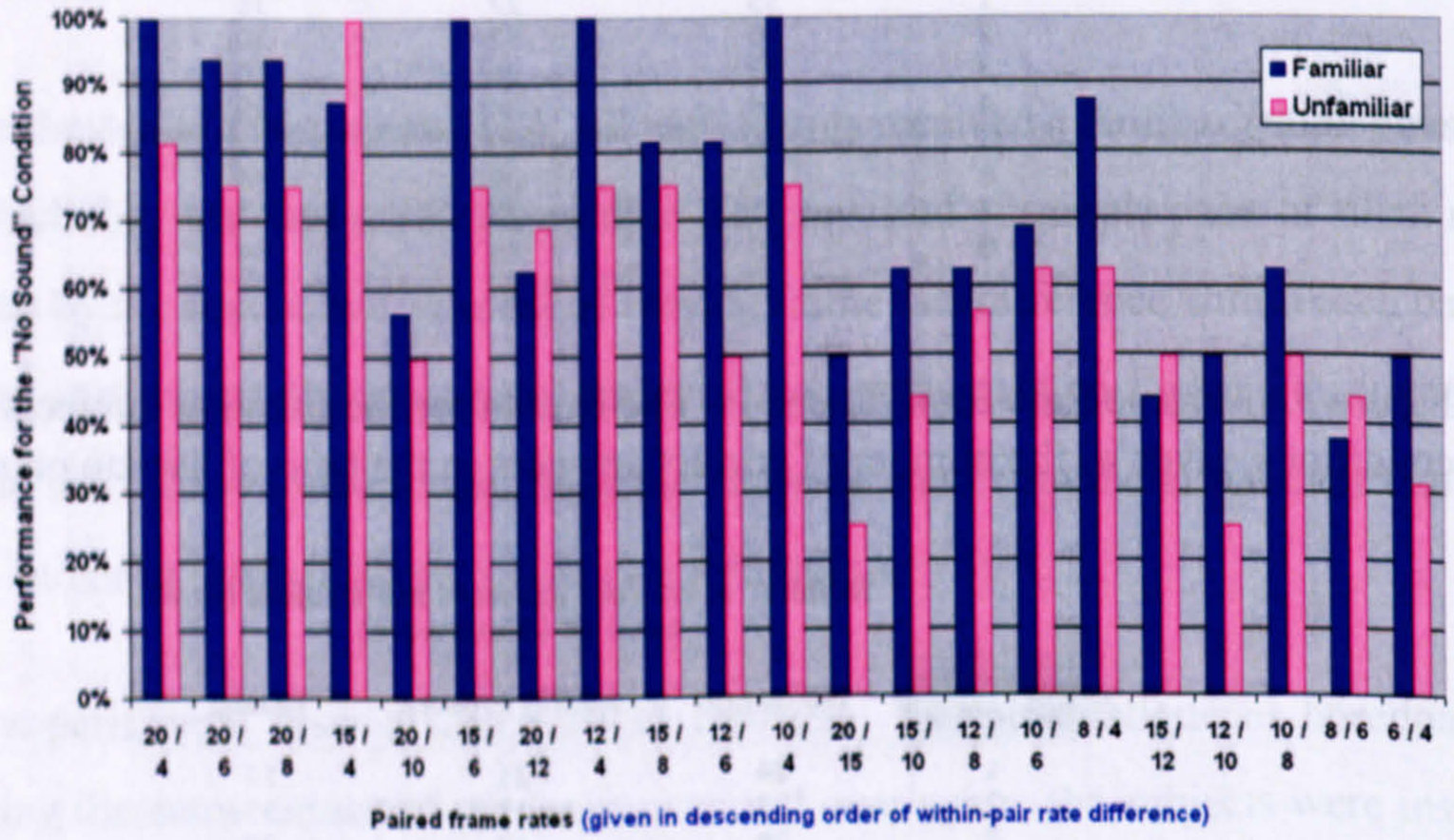


Figure 5.7: Frame Rate Experiment 1 - The performance of Familiar vs. Unfamiliar Subjects for the Control ("No Sound") condition across the trial frame rate pairs, which are ordered in our graph according to frame rate difference within the pair of rates.

From Figures 5.7- 5.9 it is clear that the performance in detecting the animation that was displayed at the higher frame rate in each pair of movies was consistently better for subjects familiar with computer graphics than the performance of the unfamiliar partic-

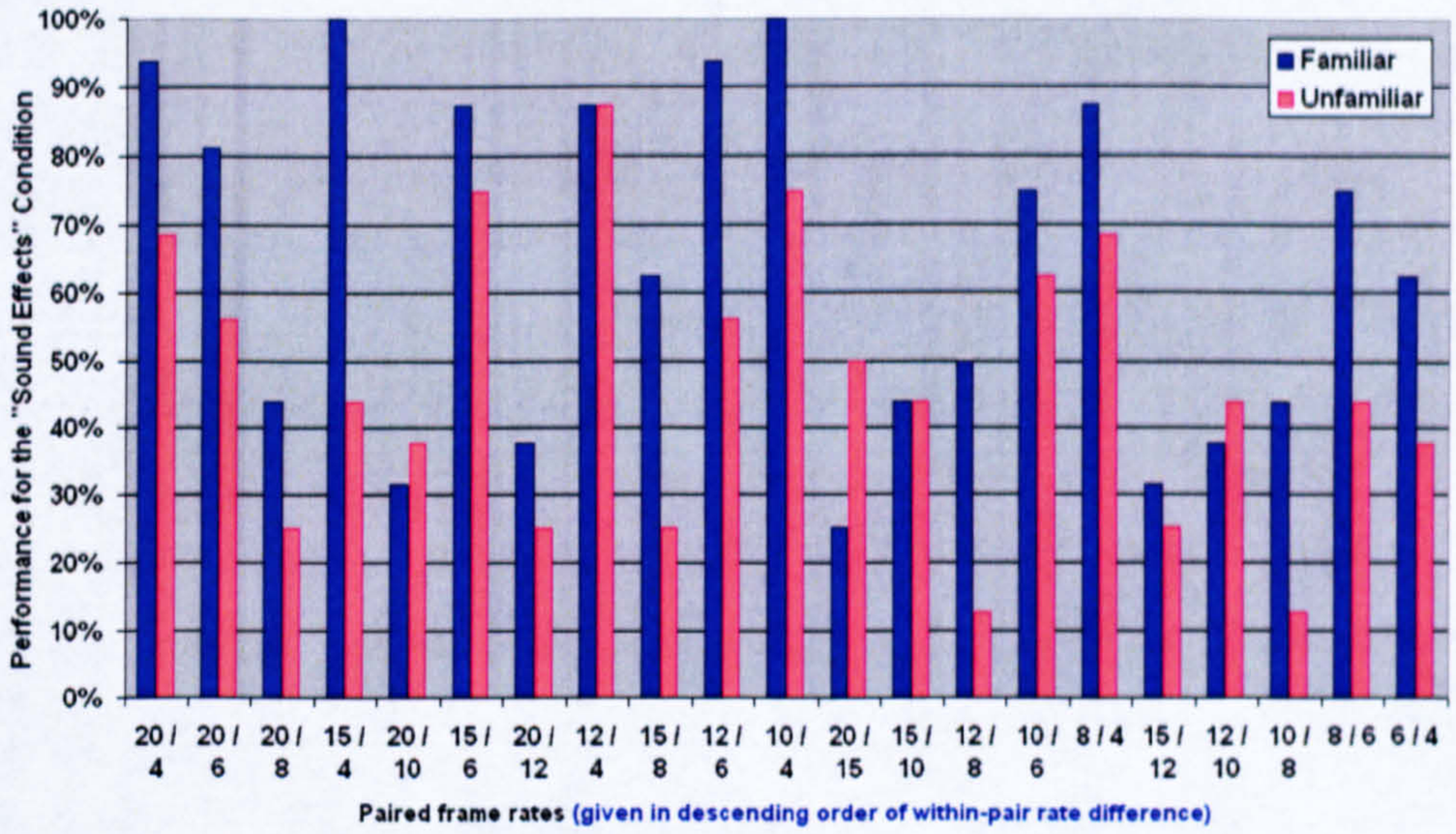


Figure 5.8: Frame Rate Experiment 1 - The performance of Familiar vs. Unfamiliar Subjects for "Sound Effects" condition across the trial frame rate pairs (ordered according to frame rate difference within the pair of rates).

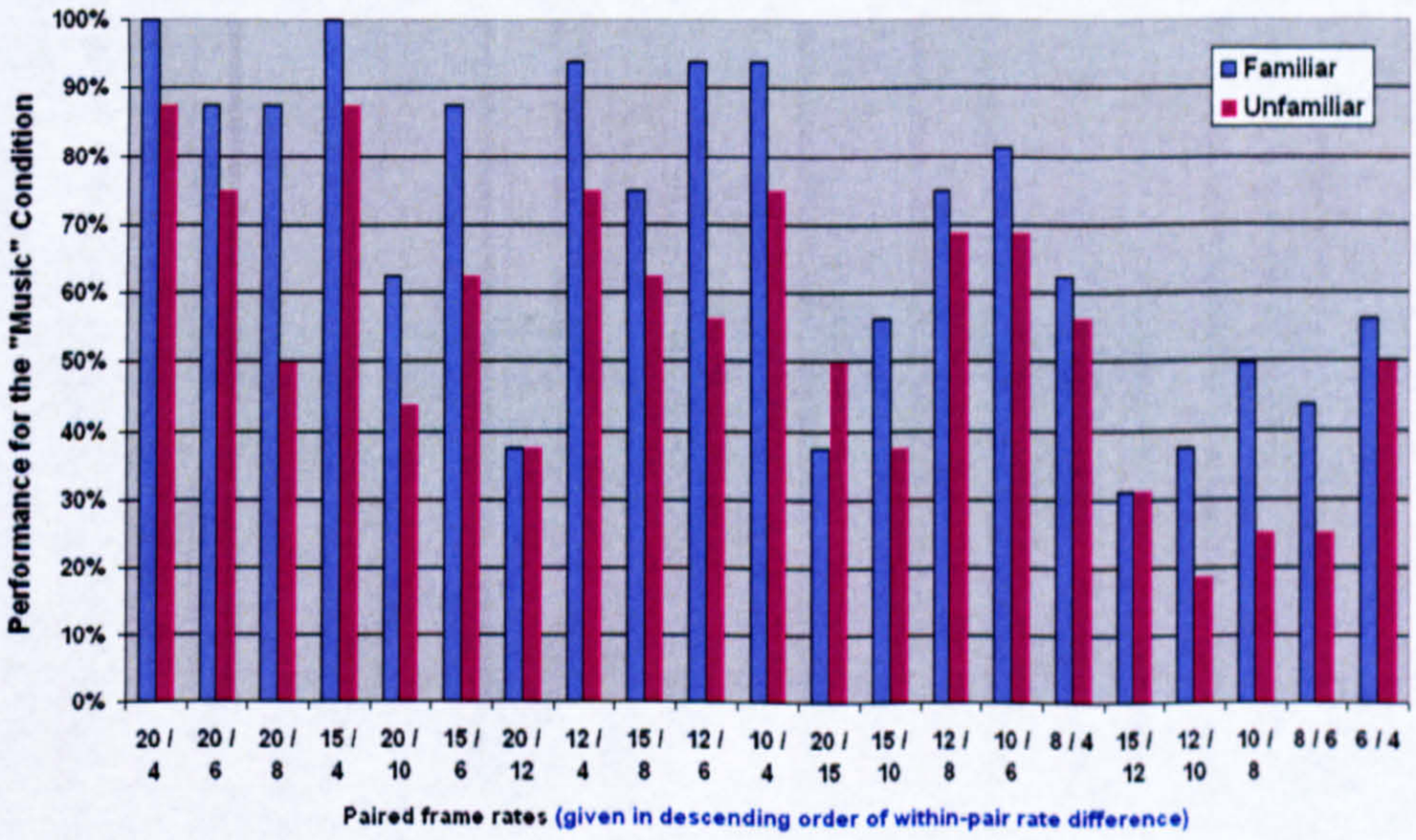


Figure 5.9: Frame Rate Experiment 1 - The performance of Familiar vs. Unfamiliar Subjects for "Music" condition across the trial frame rate pairs (ordered according to frame rate difference within the pair of rates).

ipants. In some cases, the drop in the performance of the unfamiliar subjects compared to the familiar ones, was quite dramatic. For example, familiar subjects in the "Sound Effects" group in all instances correctly identified the faster frame rate within the "15 vs 4 fps" pair, while for the unfamiliar subjects of the same group the performance dropped to 43,75%. This is in accord with the second part of our hypothesis that familiarity would be an affecting factor on the participants' performance for the specific experimental task.

The participants of the control group ("No sound") generally gave more correct answers than the other two groups (see Figures 5.10- 5.11 for the performances of participants across the 3 conditions), but we needed to analyse the results to find out whether this difference in performance was statistically significant and thus whether we should accept or reject our initial hypothesis that the auditory background (music or sound effects) of an animation acts as a distractor, making it more difficult for the human perceiver to detect frame rate variations between animations.

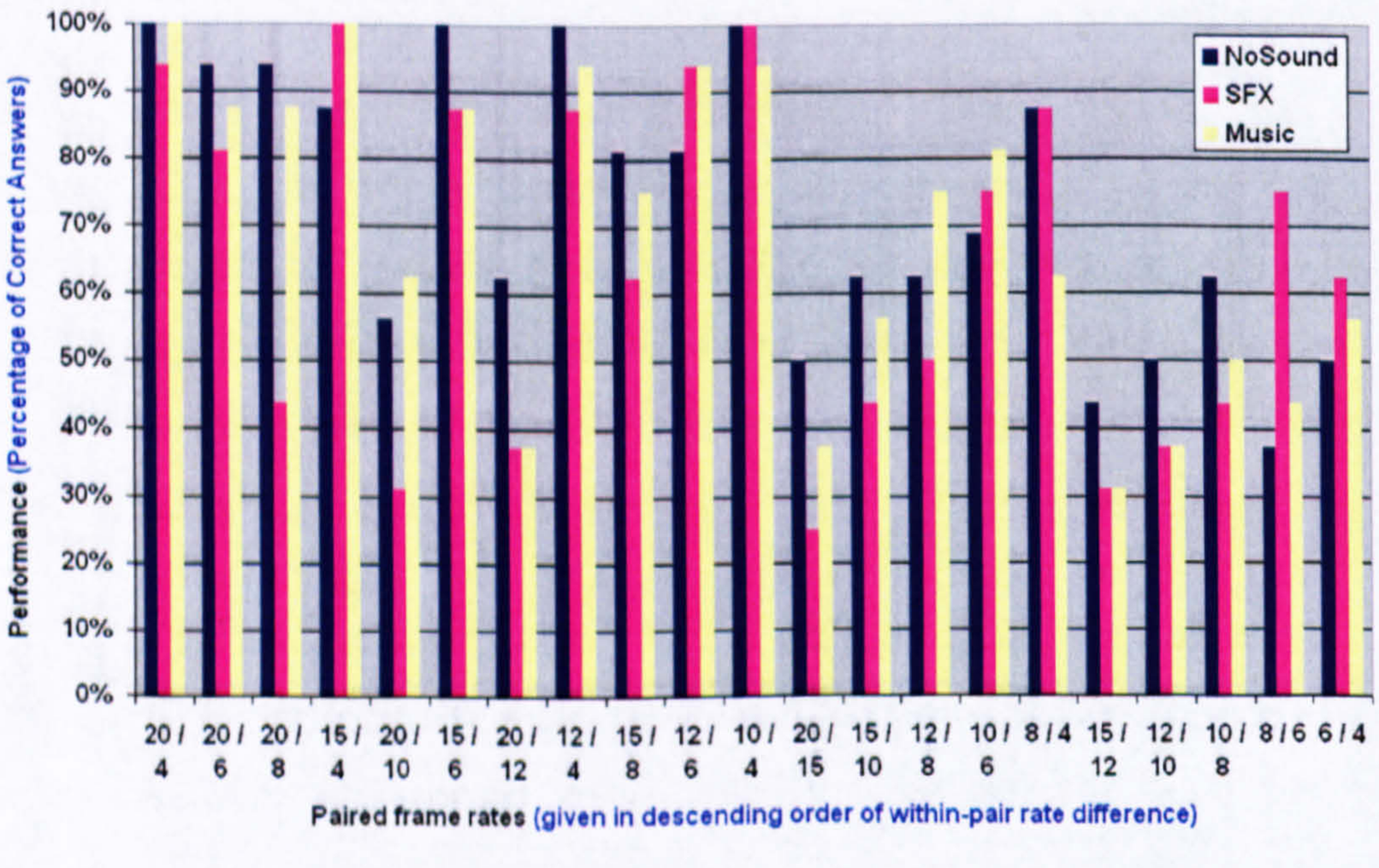


Figure 5.10: Frame Rate Experiment 1 - The performance of Familiar Subjects across all conditions.

A 3×2 factorial analysis of variance (ANOVA) [23] was used as the first method for de-

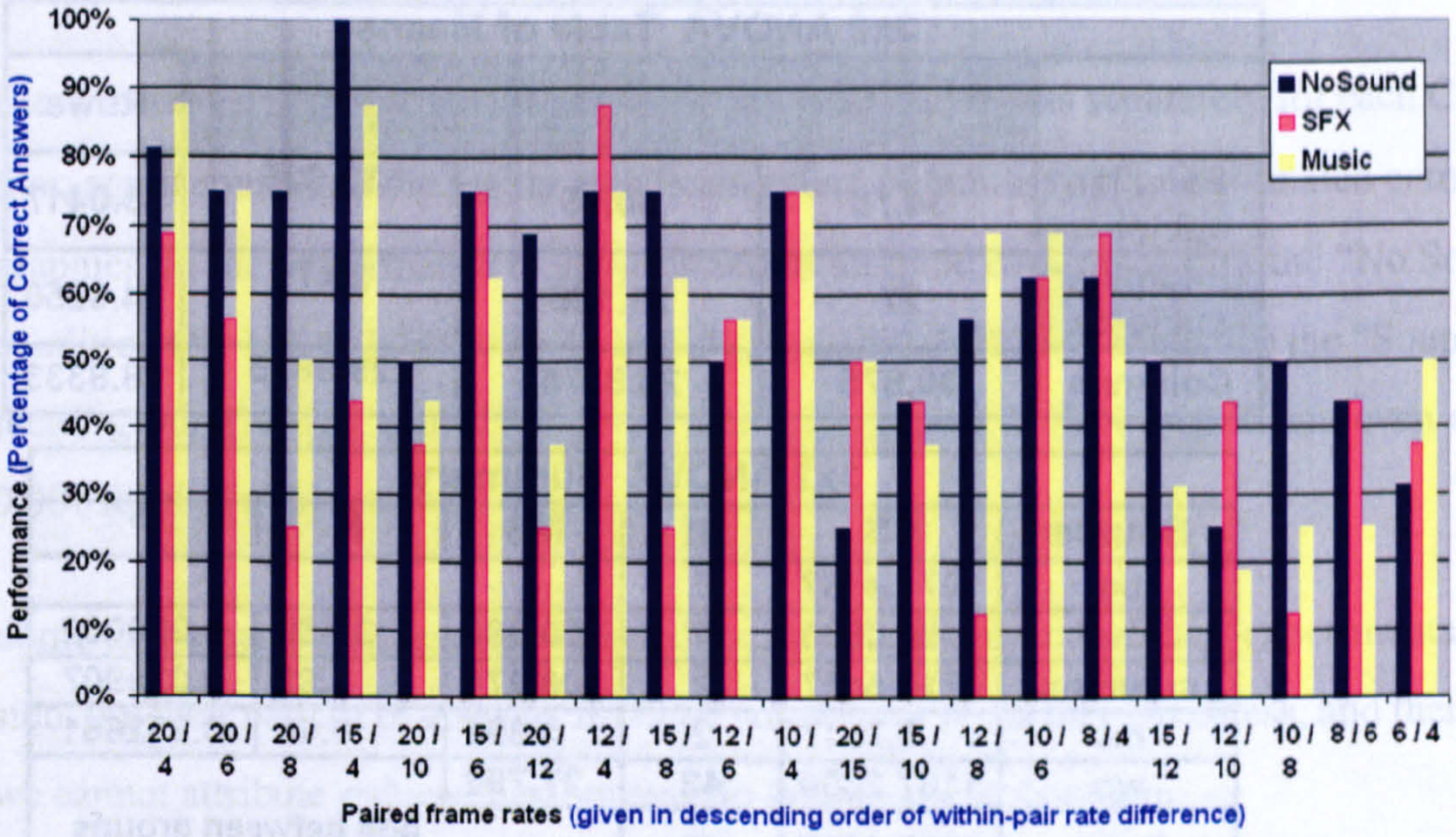


Figure 5.11: Frame Rate Experiment 1 - The performance of Unfamiliar Subjects across all conditions.

termining whether or not there was a significant between-subjects performance difference as a combined function of the auditory background difference and the level of familiarity with animated computer graphics. The 3×2 ANOVA test was applied to the raw data, shown in Figures 5.5 and 5.6. The results of the test, see Figure 5.12, revealed a highly significant statistical interaction between Familiar and Unfamiliar subjects (even at the very low 0.000001 level of risk), a significant interaction between the auditory background conditions (at the 0.05 level of risk), but no overall significant interaction between rows (Familiar/ Unfamiliar) and columns (No Sound/ Sound Effects/Music).

The data were further analysed by carrying out t-tests between the means of our independent groups, pairing every time the control group ("No sound") with the corresponding "Music" or "Sound Effects" group.

The t-test for the pair of the "Sound Effects"- Control Conditions gave a statistically significant result at the 0.05 level of risk for both Familiar and Unfamiliar participants,

3x2 ANOVA Table of Means				
	No Sound Performance	Sound Effects Performance	Music Performance	Rows
Familiar subjects	34.75	30.75	33.625	33.0417
Unfamiliar subjects	27	21.125	25.75	24.6250
Columns	30.875	25.9375	29.6875	28.8333

3x2 ANOVA Summary					
Source	SS	df	MS	F	P
bg	1071.4167	5			
rows	850.0833	1	850.083	30.59	0.000002
columns	212.5417	2	106.271	3.82	0.029907
rxc	8.7917	2	4.396	0.16	0.852661
wg	1167.2500	42	27.792		
Total	2238.6667	47		bg= between groups wg= within groups	

Figure 5.12: Frame Rate Experiment 1 - The results of the 3 × 2 ANOVA, which examined whether there was a significant between-subjects performance difference as a combined function of the auditory background difference and the level of familiarity with animated computer graphics.

confirming our hypothesis that sound effects can indeed alter the users' perception of the frame rate at which an animation is displayed. More specifically, for Familiar subjects t=2.4522, df=14, ($N_{Familiar}$=16) and $P_{one-tailed}$=0.013962. For Unfamiliar subjects ($N_{Unfamiliar}$=16) t=1.8952, df=14 and $P_{one-tailed}$=0.0394545. Therefore, the null hypothesis that there is no difference between silence and sound effects regarding the ability of the viewer to detect frame rate variations, could be rejected.

On the other hand, no statistically significant effect of fast tempo/ exciting music was found. The observed values for t, for both Familiar and Unfamiliar subjects, were smaller than the critical value, so our initial hypothesis about the effect of the chosen background music on the perceived frame rate could not be confirmed (although this does not mean that it should be necessarily rejected). More specifically, for Familiar subjects t=0.6711, df=14, ($N_{Familiar}$=16) and $P_{one-tailed}$=0.256543. For Unfamiliar subjects ($N_{Unfamiliar}$=16)

t=0.3648, df=14 and $P_{one-tailed}$=0.360356.

t-tests between the means of Familiar and Unfamiliar subjects separately for each Condition, again confirmed the highly significant effect of familiarity with animated computer graphics on the performance of subjects across all three conditions. For the "No Sound" condition (N=16), t=3.2446, df=14 and $P_{one-tailed}$=0.002938. Finally, for the "Sound Effects" group (N=16), t=3.8274, df=14 and $P_{one-tailed}$=0.000924 (significant even at the 0.001 level of risk).

Some participants had reported reduced focus towards the end of their experimental session, but their pattern of answers revealed no affect of focus on correctness, and therefore we cannot attribute reduced performance to fatigue and/or boredom.

We then went on to analyse our results using frequencies (percentages), see Tables 5.1, 5.2 and 5.3. The results in percentages indicate, amongst others, the following. Subjects, independently of level of familiarity, across all three conditions had a greater difficulty in distinguishing correctly frame rate differences between pairs of movies when both rates were higher than 8 frames per second (inclusive). This is clearly demonstrated in all 5 graphs (Figures 5.7- 5.11). For the pairs of frame rates which fall into this category the error levels for Familiar subjects of the control group ranged from 37.5% to 62.5%. The corresponding error levels for the Familiar subjects of the "Music" group followed a similar pattern (error levels: 37.5% -62.5%), but exhibited a significant rising trend for frame rates greater or equal to 10 fps in comparison to the "No sound" group. The error levels for Familiar subjects of the "Sound Effect" group were even higher (37.5%-75%). For many frame rate pairs the drop in the performance compared to the control group was greater than 33% and in some cases the performance dropped by 50% or more. Similar results were obtained from the samples of the Unfamiliar subjects, although, as already discussed, for this group the measured performance was in general significantly reduced compared to the Familiar results.

Experimental Results for the "No Sound" condition across rate pairs		
Frame rate pair	**Familiar**	**Unfamiliar**
20 / 4 fps	100%	81.25%
20 / 6 fps	93.75%	75%
20 / 8 fps	93.75%	75%
15 / 4 fps	87.5%	100%
20 / 10 fps	56.25%	50%
15 / 6 fps	100%	75%
20 / 12 fps	62.5%	68.75%
12 / 4 fps	100%	75%
15 / 8 fps	81.25%	75%
12 / 6 fps	81.25%	50%
10 / 4 fps	100%	75%
20 / 15 fps	50%	25%
15 / 10 fps	62.5%	43.75%
12 / 8 fps	62.5%	56.25%
10 / 6 fps	68.75%	62.5%
8 / 4 fps	87.5%	62.5%
15 / 12 fps	43.75%	50%
12 / 10 fps	50%	25%
10 / 8 fps	62.5%	50%
8 / 6 fps	37.5%	43.75%
6 / 4 fps	50%	31.25%

Table 5.1: Frame Rate Experiment 1 - The "No Sound" condition results in percentages, separately for Familiar and Unfamiliar subjects, across the trial frame rate pairs (ordered according to frame rate difference within the pair of rates).

Experimental Results for the "Sound Effects" condition across rate pairs		
Frame rate pair	**Familiar**	**Unfamiliar**
20 / 4 fps	93.75%	68.75%
20 / 6 fps	81.25%	56.25%
20 / 8 fps	43.75%	25.00%
15 / 4 fps	100%	43.75%
20 / 10 fps	31.25%	37.50%
15 / 6 fps	87.5%	75%
20 / 12 fps	37.5%	25%
12 / 4 fps	87.5%	87.5%
15 / 8 fps	62.5%	25%
12 / 6 fps	93.75%	56.25%
10 / 4 fps	100%	75%
20 / 15 fps	25%	50%
15 / 10 fps	43.75%	43.75%
12 / 8 fps	50%	12.5%
10 / 6 fps	75%	62.5%
8 / 4 fps	87.5%	68.75%
15 / 12 fps	31.25%	25%
12 / 10 fps	37.5%	43.75%
10 / 8 fps	43.75%	12.5%
8 / 6 fps	75%	43.75%
6 / 4 fps	62.5%	37.5%

Table 5.2: Frame Rate Experiment 1 - The "Sound Effects" condition results in percentages, separately for Familiar and Unfamiliar subjects, across the trial frame rate pairs (ordered according to frame rate difference within the pair of rates).

Experimental Results for the "Music" condition across rate pairs		
Frame rate pair	**Familiar**	**Unfamiliar**
20 / 4 fps	100%	87.5%
20 / 6 fps	87.5%	75%
20 / 8 fps	87.5%	50%
15 / 4 fps	100%	87.5%
20 / 10 fps	62.5%	43.75%
15 / 6 fps	87.5%	62.5%
20 / 12 fps	37.5%	37.5%
12 / 4 fps	93.75%	75%
15 / 8 fps	75%	62.5%
12 / 6 fps	93.75%	56.25%
10 / 4 fps	93.75%	75%
20 / 15 fps	37.5%	50%
15 / 10 fps	56.25%	37.5%
12 / 8 fps	75%	68.75%
10 / 6 fps	81.25%	68.75%
8 / 4 fps	62.5%	56.25%
15 / 12 fps	31.25%	31.25%
12 / 10 fps	37.5%	18.75%
10 / 8 fps	50%	25%
8 / 6 fps	43.75%	25%
6 / 4 fps	56.25%	50%

Table 5.3: Frame Rate Experiment 1 - The "Music" condition results in percentages, separately for Familiar and Unfamiliar subjects, across the trial frame rate pairs (ordered according to frame rate difference within the pair of rates).

5.2.6 Discussion

The results of this study have shown that the introduction of sound effects affects the perceivers' ability to perceive frame rate differences. For example, the majority of users were unable to distinguish between the animation displayed at 20 fps and that at 8 fps when sound effects were present, whereas without these audio stimuli the difference was almost always noticeable. The results were, however, less conclusive when the subjects were listening to music. The repeated seven musical clips, which had very similar temporal structure (high tempo and repetitive rhythmical pattern), may have had a reduced influence as, according to Iwaki et al. (1998) [94], physiological levels of arousal decrease with repeated listening to music regardless of musical impressions.

This work has also shown if users are unfamiliar with computer graphics then, with the presence of sound effects, the displayed frame rate can be lowered further without the user being aware that this has happened.

5.3 Experiment 2 on The Influence of Sound Effects on the Perceived Smoothness of Rendered Animations

After the promising results of the previous experiment, we prepared and conducted a second large scale experiment about the frame rate perception in VR/ animation rendering. In this experiment [128] we explored further the influence audio has on the perceptibility of motion smoothness in an animation (i.e. on the perception of delivered frame rate). This experiment was a replicate of the previous experiment with a focus on the sound

effects (which gave more promising results during the previous experiment) and using the 2AFC design, which gives more precise results from the statistical analysis point of view.

We again hypothesised that it would be more difficult for subjects to distinguish differences in the motion smoothness between audiovisual composites than between silent animations. Two conditions were considered: "Sound Effect" and the control, "No Sound", condition. We also, once again, hypothesised that familiarity with animated computer graphics would help the corresponding subjects perform the experimental task more efficiently than participants without any prior experience.

Participants viewed pairs of computer-generated walkthrough animations (with the same visual content within the pair) displayed at five different frame rates, in all possible combinations. Both walkthroughs in each test pair were either silent or accompanied by sound effects and the participant had to detect the one that had a smoother motion (i.e. was delivered at higher frame rate). We used two kinds of sound effects, related with the visual content of the scene and unrelated sounds (ambient sound effects), as we wanted also to investigate whether the distracting influence of a sound depends on the visibility of the object emitting it or not.

In the following sections, the experimental methodology of this experiment is described and the results are presented in detail.

5.3.1 Participants

Forty participants, of ages ranging from 21 to 32, from the undergraduate and postgraduate student population of the University of Bristol volunteered to participate in this study. None of them had participated in any of our previous experiments. The participants were initially divided into two groups according to their familiarity with animated computer

graphics. The first group included the subjects who had attended a Computer Graphics course and were moderately/ very familiar and the second group included participants whose studies were not related to Computer Science and had no or little experience in animated computer graphics. The members of each of these two groups were randomly subdivided across the two conditions. Participants were informed that they could withdraw at any time during the experiment and they were naive as to its purpose. They all had either normal or corrected-to-normal vision and they did not report any hearing impairment.

5.3.2 Design

For the second experiment we again used an independent samples design. The dependent variable was the perceived relative motion smoothness of the two animated sequences in each test pair. The independent variable was the auditory background of the movie clips (sound effects or silence). The conditions tested are shown in Figure 5.13.

During each session the subjects watched pairs of animations depicting walkthroughs, one after the other, and they had to judge in which of the two the motion was smoother. The two animations in each test pair had the same visual content, but were rendered at varying frame rates (either different or the same within each pair). For example frames of the animations refer to Figure 5.14.

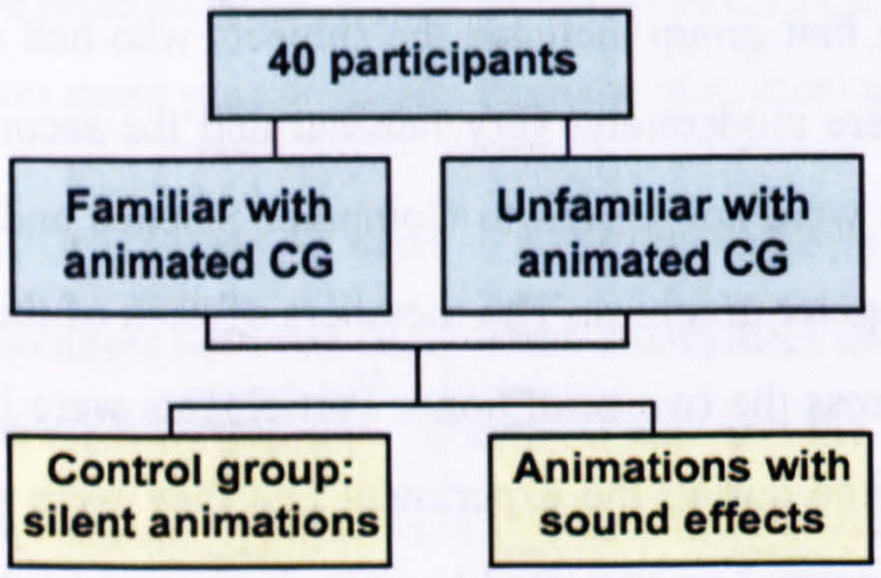


Figure 5.13: Frame Rate Experiment 2 - The Conditions tested

5.3.3 Equipment and materials

We used the same test environment with the previous experiment. The subjects watched the walkthrough animations full-screen on the 19" CRT monitor of the PC (resolution: 1280×1024 pixels). They were seated at normal viewing distance from the monitor (≈ 60 cm). Auditory stimuli were presented through the same headphones, isolated from outside noise. The volume of the sound remained the same for all subjects of the "Sound Effect" condition.

The experiment was preceded by a series of exploratory studies that would help us decide on the selection of the audio and visual stimuli materials used in our main study.

All the clips used in our experiment were based on six animated sequences of images (rendered at 640×480 pixels resolution), depicting 6 distinct parts of a walkthrough in a 3D interior scene (see example frames in Figure 5.14). We used six different sequences in order to reduce the boredom and fatigue which may result from watching the same set of visual stimuli over and over again.

In each of the six animated sequences we used, the camera employs one of the following two kinds of camera motion: a) translation along the x-axis (walk through, forward type of movement) or b) rotation around the y-axis. In half of the sequences we used the

Figure 5.14: Frame Rate Experiment 2 - Example frames from the animated sequences used for the experiment.

first type of camera motion and in the rest the second type of camera movement was employed. We decided to further investigate whether the type of camera motion would make any difference to the results, as we expected the second type of motion to accentuate the motion jerkiness caused by reduced frame rates more than a "walk through" type of motion.

Each animated sequence was rendered at 5 different frame rates to produce the movies used in the experiment: 24 fps, 20 fps, 15 fps, 12 fps and 10 fps, giving 25 frame rate combinations. We did not include more frame rates, as the additional test pairs would prolong significantly each experimental session. The 25 rate pairs coupled with the 6 animated sequences gave a total number of 150 test pairs. In 30 of the pairs the two clips were identical and they were included as a means of gauging each participant's performance. All paired comparisons were randomly generated. Each movie clip lasted for 3 seconds, in order to keep the duration of the experimental session to a minimum. A pilot study confirmed that 3 seconds were enough for a viewer to judge the motion smoothness of a test animation.

Three sound effects were selected to compose the auditory background for the "Sound Effect" condition: phone ringing, cell phone beeping and thunder. The first two sound effects were related with the visual content (phone or cell phone present in the scene) and the third was unrelated (ambient sound effect), as we wanted also to investigate whether the distracting influence of a sound depends on the visibility of the object emitting it or not.

For this experiment we did not implement spatially located 3D sounds, in order to reduce the complexity of the experiment setup. Fortunately, as discussed in section 2.2.3, when visual localisation is good, vision dominates in the spatial domain and can spatially 'capture' sound (the ventriloquist effect) and thus it is not necessary to accurately compute the spatial sound locations.

The sound files were synthesised by manipulating (multiplying and stretching) relevant freeware sounds. The produced sound effects were equated for peak amplitude and their properties were: 44100Hz, 16 bit, Stereo. The movie clips in each test pair had the same auditory background. For the audiovisual composites, each sound effect was assigned to two of the six rendered image sequences: one for each type of camera motion, in order to counterbalance for possible interactions between one of the sound effects and a specific type of camera movement.

Due to the prohibitive memory requirements of loading 150 uncompressed clips (together with the countdown sequences), we had to compress the video clips. It was not possible to avoid completely the blurriness and flickering which results from the video compression, but the same anomalies were present in all conditions and therefore they should not affect our results.

5.3.4 Procedure

Each participant was tested individually. Participants were informed that they should watch carefully pairs of computer-generated walkthroughs with varying audiovisual content. The subjects in the "Sound Effect" group were told that the animations would be accompanied by sound effects and the sound would be delivered to them through headphones. Even the participants of the "No Sound" group had the headphones on during their experimental task, so as to be better isolated from outside noise. When each pair finished they would have five seconds before the next pair loaded to answer the question: "Which of the two movies in the test pair you just watched do you think had a better visual quality taking into consideration only the motion smoothness or on the contrary jerkiness?"

For each pair, the participants could select one of: "The first seemed better", "The second seemed better". They were instructed that they should pick one of the options even when they could not perceive any difference between the motion smoothness of the two movies (2-Alternative Forced-Choice method, 2AFC) [1]. Visual signals, see Figure 5.15, indicated the beginning of the two clips within each test pair and a count down was displayed between consecutive trials so the participants knew exactly how much time was left for them to give their answer. During the five-second countdown there was silence for all conditions.

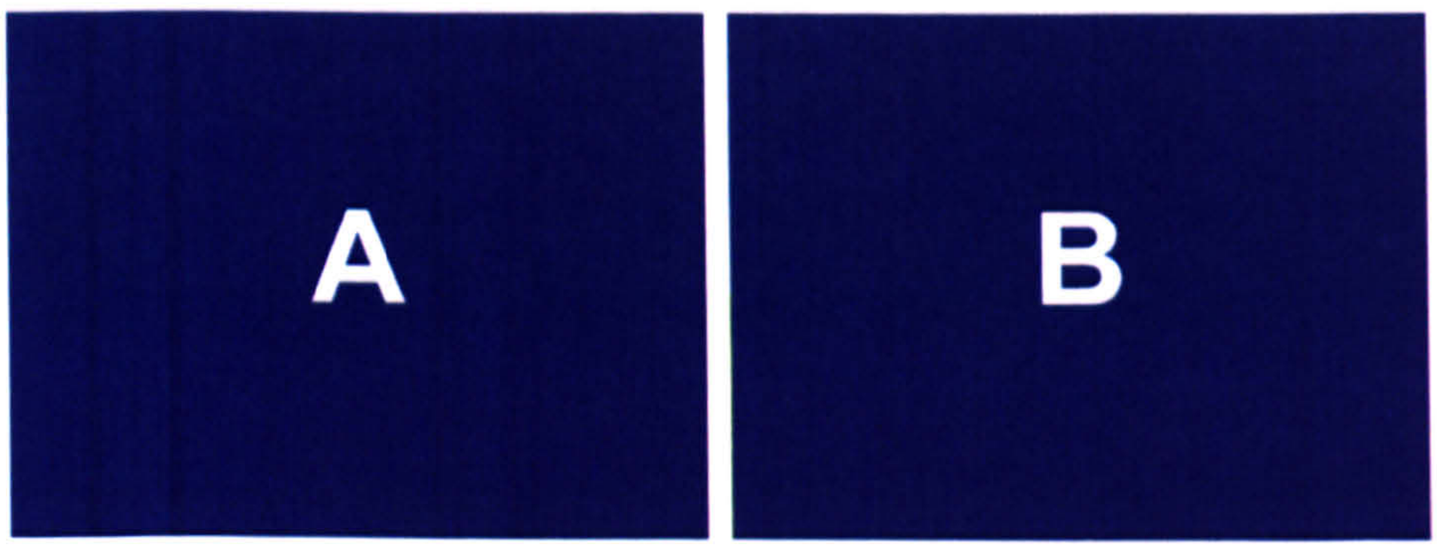


Figure 5.15: Frame Rate Experiment 2 - Visual signals indicating the beginning of the first (left image) and the second clip (right image) within each test pair.

Before the actual experimental task, all participants received a familiarisation phase, during which they watched a training sample that consisted of silent sample pairs of animations (divided by the countdown periods) of varying frame rate difference within pair. The training sample was played as many times as each participant wished and during its playback he/she received instructions from the experimenter about the visual cues that would help him/her perform the task.

Each experimental session lasted 35-40 minutes. To minimise the effect of fatigue/boredom (due to the repeated watching of the same visual stimuli) on the results, the subjects were instructed that they could pause the display during any countdown interval and continue when they felt ready again. Less than five participants chose to pause the experiment, and this happened only once during their experimental task.

Each questionnaire concluded with two questions about how relaxed/comfortable the participant felt and how focused he/she was while watching the animations: a) at the beginning and b) towards the end. They could select one of the following options: Not at all / A little bit / Moderately / Very much. These questions were used to check whether any change in the focus or comfort levels would affect the subjects' performance.

5.3.5 Results

Measure of performance in our experimental task was the percentage of times each subject correctly identified the smoother animation (and thus the higher frame rate) within a pair of displayed animations. The performance was averaged for each pair of frame rates across all subjects within each group. For example, a performance of 100% for a pair of frame rates within a group indicates that all subjects from this group correctly identified the clip with the smoother motion (higher frame rate) whenever they came across the corresponding pair of rates during the experimental task. Figures 5.16 and 5.17 illustrate the performance of Familiar versus Unfamiliar participants within each condition and Figure 5.18 compares the performances measured across the two conditions.

We decided to present and analyse the results for the actual frame rate combinations (e.g 10 vs 15 fps) and not to consider only the difference between a pair of rates (e.g. 5 fps difference), because the perceptibility of visual defects is reduced when the absolute frame rate values are close to 20-24 fps, compared to lower values.

As we have already mentioned, in 30 of the test pairs the two animations were identical and they were included as a means of gauging each participant's performance. If the experiment was designed and conducted properly, the participants should pick one of the two options randomly for these pairs and, therefore, in our results we should get each

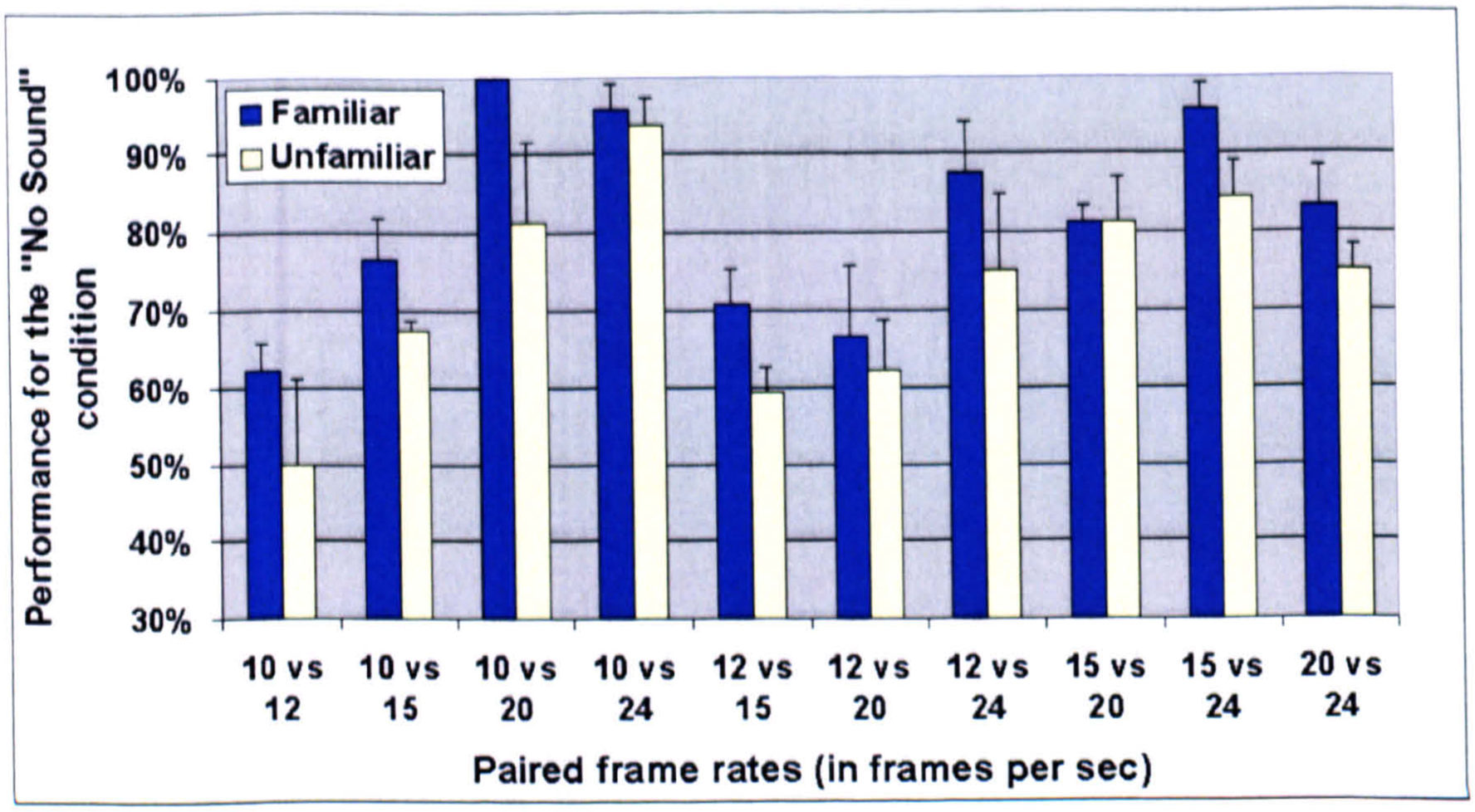


Figure 5.16: Frame Rate Experiment 2 - The performance of Familiar vs. Unfamiliar Subjects for the control ("No Sound") condition across the test frame rate pairs.

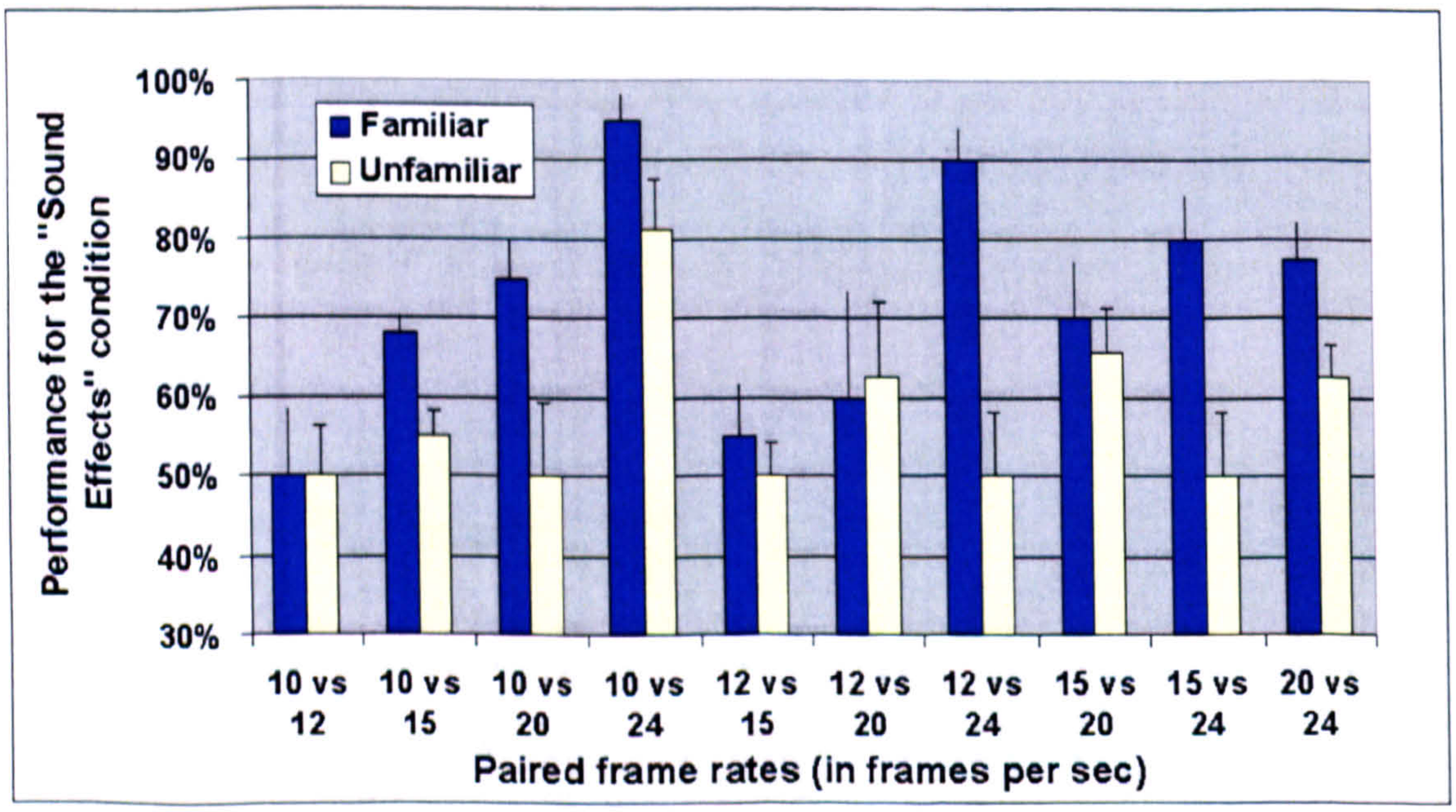


Figure 5.17: Frame Rate Experiment 2 - The performance of Familiar vs. Unfamiliar Subjects for the "Sound Effect" condition across the test frame rate pairs.

of the two possible answers in approximately 50% of the cases. The actual percentage in our results was 50.5%, proving that our experimental results were valid and could be

statistically analysed further.

From Figures 5.16 and 5.17 it is clear that the performance of subjects who were familiar with computer graphics in detecting the animation that was displayed at the higher frame in each pair of animations, was generally better than the performance of the unfamiliar participants across both conditions. This is in accord with the second part of our hypothesis that familiarity would be an affecting factor on the participants' performance for the specific experimental task.

As we can see in Figure 5.18, the control, "No Sound", group consistently gave more correct answers than the other group. Figure 5.19 reveals that both Familiar and Unfamiliar subjects contributed to the performance drop of the "Sound Effect" group. Therefore, not even the viewers who were very familiar with animated computer graphics escaped the influence of the sound effects. Further statistical analysis of the results was nevertheless necessary, in order to find out whether this drop in the performance of the "Sound Effect" group was significant and thus whether we should accept our initial hypothesis that sound effects in an animation make it more difficult for the viewer to detect frame rate variations.

We first had to decide whether we should use parametric or non-parametric tests for the statistical analysis of the results. We were not certain whether the distribution of our population was Gaussian, but according to the *Central Limit Theorem* if the samples are large enough, the distribution of means will follow a Gaussian distribution even if the population is not Gaussian. Assuming the population does not have a really peculiar distribution, a sample size of 10 is generally enough to invoke the Central Limit Theorem. Since most parametric tests, such as the t-test and ANOVA, are concerned only with differences between means, the Central Limit Theorem lets these tests work well even when the populations are not Gaussian.

A two-way ANOVA (ANalysis Of VAriance) for independent samples, see Figure 5.20, was first carried out in order to investigate if the performance in the task was jointly in-

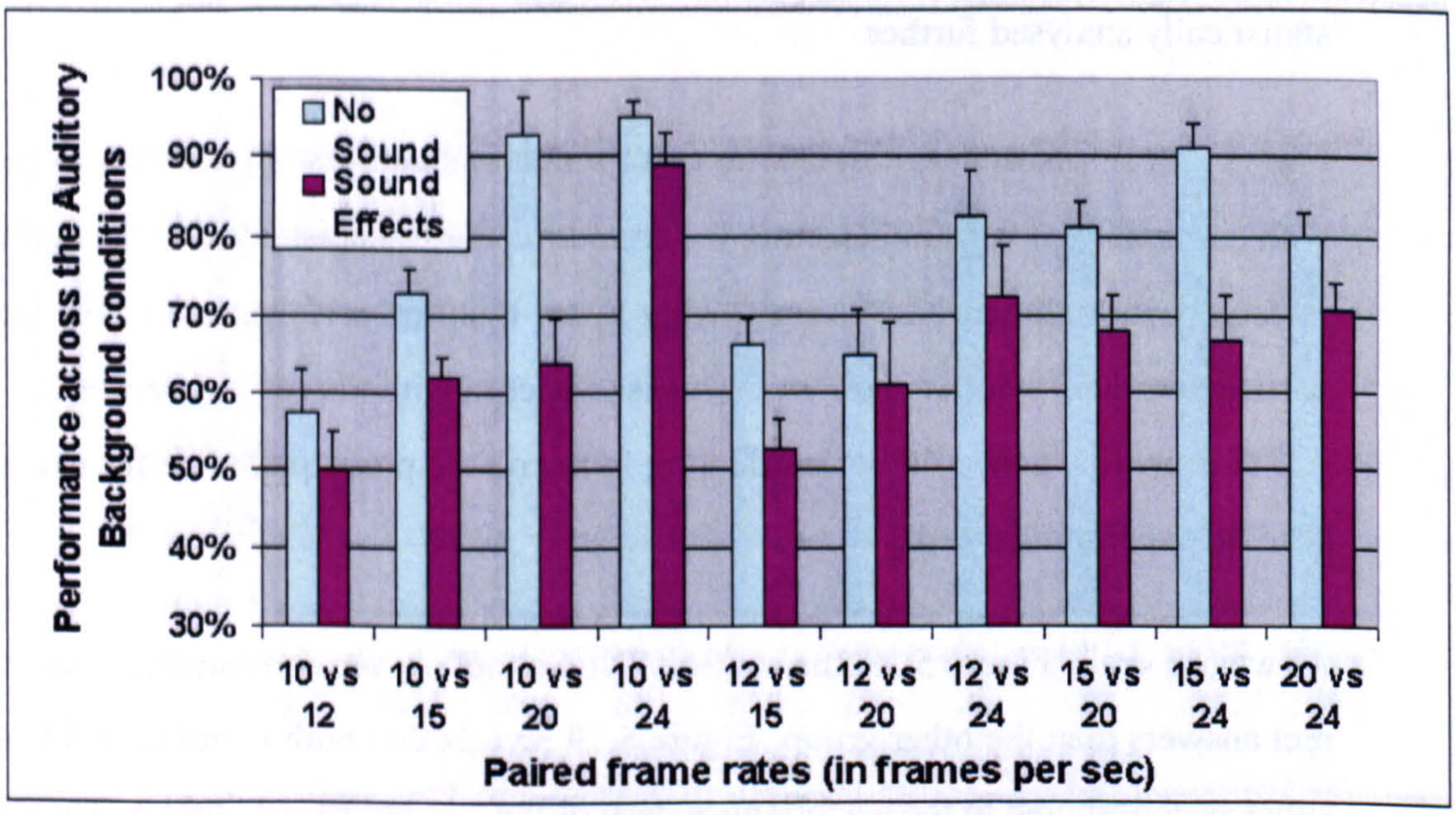


Figure 5.18: Frame Rate Experiment 2 - The performance of all Subjects across the "No Sound" and "Sound Effect" conditions, separately for each frame rate combination.

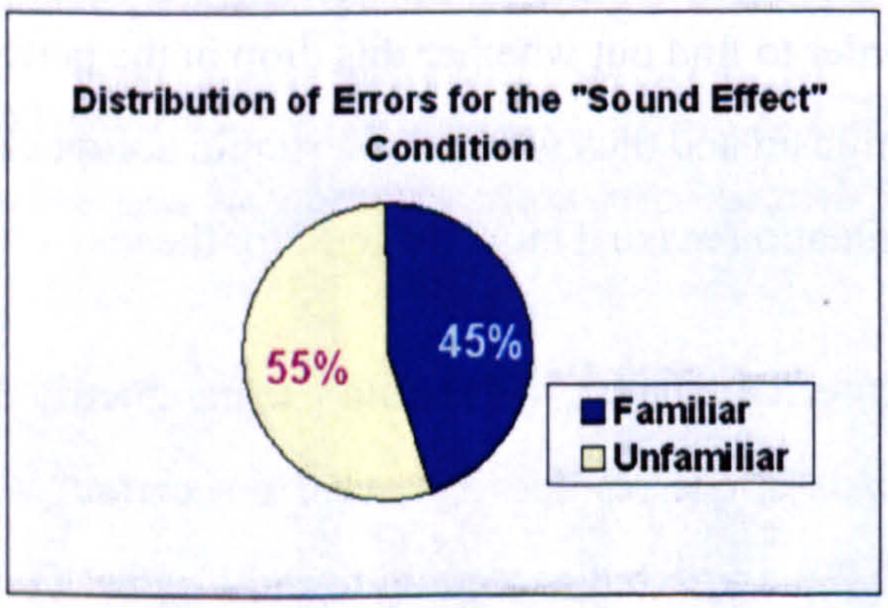


Figure 5.19: Frame Rate Experiment 2 - How errors (i.e. incorrect detections of the higher frame rate within each pair of animations) were distributed between Familiar and Unfamiliar subjects in the "Sound Effect" condition.

fluenced by the auditory background and the familiarity of the subjects with animated computer graphics (i.e. whether there is an interaction between these two independent variables). A two-factor analysis of variance consists of three significance tests: a test of each of the main effects of the two independent variables and a test of the interaction between them. The main effect of the auditory background was found to be very significant, even at the 0.0001 level of risk ($F=26.51$, $df=1$ and thus $P=0.00001 < 0.0001$).

The main effect of the familiarity with animated CG was also very significant, even at the 0.00001 level of risk (F=29.41 df=1, P=0.000004 <0.00001), revealing that prior experience with computer graphics has a major influence on the ability of the viewer to perceive temporal defects which result from low frame rates. Furthermore, a significant interaction between these two independent variables (auditory background and degree of familiarity) was revealed at the 0.05 level of risk (F=5.13 df=1, P_{rxc}=0.0296 <0.05).

2x2 ANOVA Table of Means			
	No Sound Performance	Sound Effects Performance	Rows
Familiar subjects	50.8	46.6	48.7
Unfamiliar subjects	46.2	35.4	40.8
Columns	48.5	41	44.75

2x2 ANOVA Summary					
Source	SS	df	MS	F	P
bg	1295.5	3			
rows	624.1	1	624.1	29.41	0.000004
columns	562.5	1	562.5	26.51	0.000010
rxc	108.9	1	108.9	5.13	0.029633
wg	764	36	21.222		
Total	2059.5	39		bg= between groups wg= within groups	

Figure 5.20: Frame Rate Experiment 2 - The results of the two-way ANOVA, which examined whether the performance in the experimental task was jointly influenced by the auditory background and the familiarity of the subjects with animated computer graphics (i.e. whether there is an interaction between these two independent variables).

The data were then analysed by carrying out an unpaired t-test for 2 Independent Samples between the means of our independent "No Sound" and "Sound Effect" conditions, separately for Familiar and Unfamiliar subjects, in order to determine whether or not there was a significant between-subjects performance difference as a function of the auditory background only.

The t-test gave a statistically significant result for both Unfamiliar and Familiar partici-

pants, see Figure 5.21, at the 0.001 and 0.01 levels of risk, respectively. More specifically, for Familiar subjects t=2.9469 df=18 ($N_{Familiar}$=20) and $P_{one-tailed}$=0.004312. For Unfamiliar subjects ($N_{Unfamiliar}$=20) t=4.4764, df=18 and $P_{one-tailed}$=0.000146. Therefore, the null hypothesis that there is no difference between silence and sound effects regarding the ability of the viewer to detect smoothness/jerkiness variations, could be rejected.

t-test for Unfamiliar subjects		
Values	**NO SOUND (a)**	**SOUND EFFECT (b)**
n	10	10
mean	74.518	57.096
variance	103.4111	48.0654
t = +4.4764 **df** = 18	$P_{two-tailed}$ = 0.000292	
	$P_{one-tailed}$ = 0.000146	

t-test for Familiar subjects		
Values	**NO SOUND (a)**	**SOUND EFFECT (b)**
n	10	10
mean	81.936	71.292
variance	50.1692	80.2886
t = +2.9469 **df** = 18	$P_{two-tailed}$ = 0.004312	
	$P_{one-tailed}$ = 0.008624	

Figure 5.21: Frame Rate Experiment 2 - The results of the unpaired t-test between the means of the independent "No Sound" and "Sound Effect" conditions, separately for Unfamiliar subjects (left) and Familiar subjects (right).

We further analysed the results from the "Sound Effect" group to examine whether the distracting influence of a sound depends on the presence in the scene of the object emitting it, but no statistically significant effect of the type of sound, scene-related or unrelated, on our subjects' performance was found.

In Figure 5.22, we can see the performance of all subjects from both conditions for the two types of camera motion, translation and rotation. Another two-way ANOVA for independent samples was carried out, see Figure 5.23, in order to investigate the main effect of the type of camera movement on the subjects' in the task performance and also if there was a significant interaction between the type of camera motion and the auditory background. The main effect of type of camera movement was not significant (F=0.51 df=1, P=0.4743), and therefore the null hypothesis, that there is no difference between the "translation" and "rotation" types of movement in the scene regarding the ability of the viewer to detect smoothness/jerkiness variations, could not be rejected. This was a bit surprising as we had expected that participants would detect temporal artefacts more

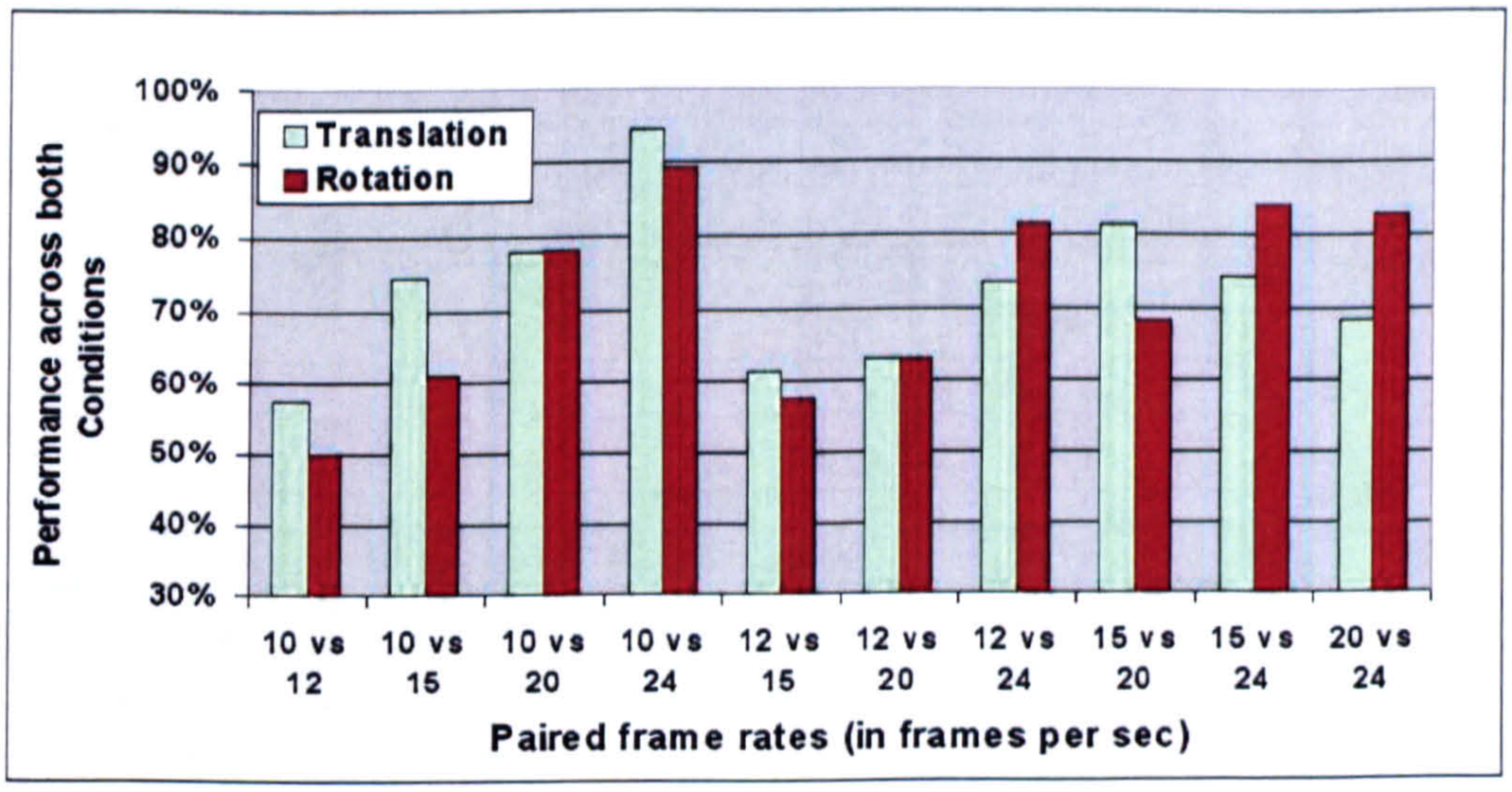


Figure 5.22: Frame Rate Experiment 2 - The performance of all subjects across the 2 types of camera motion (translation and rotation).

easily for the animations where the camera was rotating. Furthermore, no interaction between the auditory background of the scene and the type of camera movement was revealed (F=1.52 df=1, P_{rxc}=0.2214). The main effect of the auditory background was again found to be very significant, even at the 0.0001 level of risk (F=23.71, df=1 and thus P=0.000006 <0.0001).

During the design of our experiment we were concerned about the influence of the familiarisation to the three sound effects on their distracting influence, and more specifically that after a number of animation pairs the viewers would get used to them and the sounds would not 'attract' their attention any longer. An important ability of the auditory system is habituation to a continuous sound, which can fade into the 'background' of consciousness after a short period of time. Nevertheless, the frequency of the correct answers did not increase during the course of the experimental task for the vast majority of the participants (38 out of 40), revealing that the influence of the sound effects remained significant throughout the task, despite the fact that they were no longer novel to the viewers. This may have to do with the fact that if the sound changes or stops, as in our case, then it would come to the foreground of attention, because of the sensitivity of the auditory system to

2x2 ANOVA between AUDITORY BACKGROUND and TYPE OF CAMERA MOTION

2x2 ANOVA Table of Means			
	NO SOUND	SOUND EFFECT	Rows
Translation	25	20.3	22.65
Rotation	23.5	20.7	22.10
Columns	24.25	20.5	22.375

2x2 ANOVA Summary					
Source	SS	df	MS	F	P
bg	305.35	3			
rows	6.05	1	6.05	0.51	0.477326
columns	281.25	1	281.25	23.71	0.000006
rxc	18.05	1	18.05	1.52	0.221420
wg	901.4	76	11.861		
Total	1206.75	79		bg= between groups wg= within groups	

Figure 5.23: Frame Rate Experiment 2 - The results of the two-way ANOVA, which examined whether the performance in the experimental task was jointly influenced by the auditory background and the type of camera motion in the 3D scene.

change [29].

Some participants (7 out of 40) had reported reduced focus towards the end of their experimental session, but the pattern of their answers revealed no affect of focus on correctness, and therefore we cannot attribute reduced performance to fatigue and/or boredom.

We also analysed our results using frequencies (percentages), see Figure 5.24. The results in percentages indicate, amongst others, the following.

Unfamiliar subjects in the "Sound Effect" condition could identify with ease only the difference between 10 and 24 frames per second, in 81.25% of the cases. Their performance for almost all the other frame rate pairs (10 vs. 12 fps, 12 vs. 15 fps, 10 vs. 15 fps, 15 vs. 24 fps, 10 vs. 20 fps, 12 vs. 24 fps) was just 50%-55%, which indicates that they gave random answers when they encountered those pairs while performing their experimental

Frame rate pair	"Sound Effect" condition Performance		"No Sound" condition Performance	
	Familiar	Unfamiliar	Familiar	Unfamiliar
10 vs 12 fps	50%	50%	62,5%	50%
12 vs 15 fps	55%	50%	70,83%	59,38%
20 vs 24 fps	77,5%	62,5%	83,33%	75%
10 vs 15 fps	68%	55%	76,67%	67,5%
15 vs 20 fps	70%	65,63%	81,25%	81,25%
12 vs 20 fps	60%	62,5%	66,67%	62,5%
15 vs 24 fps	80%	50%	95,83%	84,38%
10 vs 20 fps	75%	50%	100%	81,25%
12 vs 24 fps	90%	50%	87,5%	75%
10 vs 24 fps	95%	81,25%	95,83%	93,75%

Figure 5.24: Frame Rate Experiment 2 - The results in percentages for the two auditory background conditions, separately for Familiar and Unfamiliar subjects, across the trial frame rate pairs (ordered according to frame rate difference within the pair of rates).

task. On the other hand, Unfamiliar subjects in the "No Sound" condition had difficulty in "identifying" the frame rate difference only for the pairs "10 vs. 12 fps" and "12 vs. 15 fps".

Familiar subjects in the "Sound Effect" group could not distinguish between animations displayed at frames rates with difference 2-5 frames per second, while the Familiar subjects in the "No sound" group could generally distinguish between clips with difference greater than 2 frames per second.

The results presented above confirm our hypothesis that sound effects can significantly affect the viewers' ability to detect motion smoothness/jerkiness variations which result from changes to the frame rate, regardless of the type of camera movement in the scene (translation or rotation). The observed influence of the sound does not depend on whether the sound affects are ambient sounds (unrelated to the scene content) or they are related to specific objects in the 3D scene. Users who are unfamiliar with computer graphics are

affected more than the familiar viewers by the presence of sound effects and they do not detect smoothness/jerkiness variations which are obvious when there is no sound.

5.4 Summary

The results of the experiments that we conducted on the influence of sound on the perceived smoothness (i.e. perceived delivery frame rate) of rendered animations confirm that in the presence of audio stimuli, and more specifically sound effects, viewers fail to notice variations in the motion smoothness between walkthrough animations displayed at different rates, which are apparent in the absence of sound. According to psychophysical findings discussed in previous chapters, this may be due to the fact that the auditory stimuli attract part of viewer's attention to the sound and away from the visual defects, such as jerky motion, which result from low frame rates.

We also demonstrated that viewers who are not familiar with animated computer graphics, can much harder notice variations in the motion smoothness between two audiovisual animations, compared to people with prior experience.

The effect of the type of camera movement in the scene (translation or rotation) on the viewers' perception of the motion smoothness/jerkiness was also investigated, but no significant association between them was found.

Chapter 6

Experiments on the Perceived Rendering Quality under the influence of Sound Effects

6.1 Introduction

As we mentioned in section 3.2.1, according to the well-documented phenomenon of "inattentional blindness", there is no conscious perception without attention. Our attentional system has a limited capacity and attention is shared among stimuli from the various sensory modalities and our thoughts. According to [178], usually only one 'object' is the centre of attentional focus and in order to handle tasks involving more objects, humans have to rapidly switch attention between them.

As discussed in section 3.4, attention can be voluntarily allocated, but sometimes an intense novel stimulus may capture it. What is very important for our research is that attentional capture is subject to intersensory interactions and more specifically, an auditory event also attracts visual attention to the perceptual origin of the sound [48], regardless of

the visual perceptual load [219]. Despite the growing body of studies on how visual stimuli can be manipulated to speed up rendering by employing perceptual criteria, researchers in the rendering field have not as yet considered the findings regarding crossmodal interactions on the spatial allocation of visual attention when designing perceptually adaptive rendering algorithms.

In this study we investigated whether auditory stimuli, and more specifically sound effects with abrupt onsets, affect a viewer's perceived quality of rendered images while watching computer generated walkthroughs. This work was inspired by the phenomenon of inattentional blindness [122] and the evidence from the field of crossmodal perception that, under certain conditions, visual attention can be mislocated by auditory distractors.

From the perception findings mentioned above, we can infer that unexpected sounds- auditory distractors- in CG animations or 3D environments may attract the viewer's attention to the perceptual origin of the sound in the scene, allowing us to selectively render only the sound emitting objects (SEOs) and their surrounding area at the highest quality while the rest of the scene objects can be rendered at a much lower quality and thus with less computational effort, without any noticeable difference to the observer. If so, computation time could be greatly decreased by only rendering the pixels that are being attended to.

To test our hypothesis we developed a suitable selective renderer. A psychophysical experiment with 120 participants was run which revealed a significant effect of sound effects on the perceived rendering quality. This experiment is described in detail in section 6.2 and its sub-sections. Although our selective renderer calculates an approximation of the full global illumination computation which is currently prohibitive for interactive scenarios, as it still takes minutes to compute, our approach could also be applied to gaze-contingent multi-resolution displays (GCMRDs), which put high resolution only at the center of visual attention to greatly reduce the bandwidth required by interactive single-user image display applications. This work has been published in [129].

An eye tracking experiment was also conducted to analyze the pattern of saccades and fixations during a sound effect while a viewer is attending a rendered scene. This experiment is presented in detail in section 6.3 and its sub-sections.

The chapter concludes with a discussion on the findings of the aforementioned experiments.

Figure 6.1: Our Selective Rendering approach. It renders at higher quality the sound emitting objects (SEOs) and the surrounding pixels, while reducing the rendering quality of the rest of the pixels, by shooting to each a lower number of rays (what we call reduced rendering quality, RQL). When there are no SEOs on screen all the pixels are rendered at the predefined high quality.

6.2 Experiment on the Perceived Rendering Quality under the Influence of Sound Effects

Our next large-scale experiment investigated whether auditory stimuli, and more specifically sound effects with abrupt onsets, affect a viewer's perceived quality of rendered images while watching computer generated animations [129]. Based on the previous work on audio-visual perception we hypothesised that it would be more difficult for subjects to notice quality variations (degradations) in audio-visual composites than in silent animations. Two conditions were considered: "Sound Effect" and the control, "No Sound", group.

6.2.1 The Selective Renderer

In order to test our hypothesis we developed a selective renderer, which is an extended version of the rendering engine of the Radiance Lighting Simulation package [242]. Our renderer is given as input a list of frames to render, the default, high, rendering quality (HQL) to be maintained in the absence of sound effects-in rays per pixel- and for each frame the sound-emitting objects (SEOs) which are active within that frame. When there is an active SEO on screen only this and its surrounding pixels, are rendered at the default quality (HQL), and all the other pixels are rendered at a lower level (reduced rendering quality level, RQL), as seen in Figure 6.1.

The pixels around the SEO are rendered at a quality proportional to their distance from its boundary. This causes a quality gradient as can be seen in Figure 6.3. The size of the area to be rendered at higher quality is also one of the user-adjustable rendering parameters. For our main experiment it was 350×350. The resolution of the rendered images was 720×540.

The way our renderer was designed, it can smoothly change the rendering quality between consecutive frames, in order to facilitate for possible habituation to the sound and also for the fact that there is a delay between the onset of the sound effect and the fixation of the observer to the SEO. This delay depends on the saliency of that object and also on the auditory cues that help the localisation of the sound source.

When there are more than one visible sources of sound in the renderable part of the scene, we render the pixels associated with all of them at high quality, as it would be very complicated to try to determine which of the sound-emitting objects would actually capture attention. Which of the areas competing for attention might stand out depends on many factors, such as the saliency of the object and the distracting effect of each sound. The latter depends, among others, on the loudness of the sound and on its information content.

Selective Rendering Pipeline

Our selective rendering pipeline, illustrated in Figure 6.2, has two phases (see also [40]). The first phase, consisting of the pre-selective rendering and selective guidance stages, is responsible for rendering the pixels of every frame to the base (low) quality level (RQL) and at the same time 'locating' sound-emitting objects present in each frame, with the help of the primary ray or certain secondary rays (*detector rays*). A data structure termed *Sound-Emitting Object List* (SEOL), is a queue responsible for storing the SEO data and the corresponding pixel locations in each frame.

The second phase renders each frame selectively, by additionally applying a gradient decrease in quality around the 2D projection of the SEOs according to the user-defined size of high-quality area, as described later in the chapter.

For our selective renderer quality is a function of rays per pixel, according to user-defined

the quality is updated only for the pixels corresponding to the SEOs and their surrounding area, by shooting extra rays, as described above.

Please note that a frame may be tagged as selective even when it does not contain any SEOs. This is particularly useful for the case of frames that have a rich auditory background, such as attention-grabbing sounds coming from sources outside the viewer's field of view. These sounds, according to already presented psychophysics findings, will capture the attention of the viewer and will limit the available perceptual resources for the processing of the visual stimuli.

The Quality Buffer structure

The second phase of the selective rendering pipeline introduces another data structure, the *quality buffer*, or *q-buffer* for short, which ensures that the correct number of rays are shot for each pixel and that computing more than the desired quality for neighbouring pixels is avoided. The q-buffer functions in a similar way to a z-buffer [32], which is used for hidden surface removal in renderings by solving depth issues and deciding which elements of a rendered scene are visible, and which are hidden.

The q-buffer has the size of the image to be rendered and stores the number of rays shot up to that point for each pixel in the corresponding entry. At the beginning of the second phase all entries in the q-buffer are initialised to the value of the user-defined base quality. In the second phase, the SEO list is parsed and for each entry a number of rays intended to be shot is calculated. It is also tested whether or not the pixel is a boundary pixel of the projected image of the distractor object.

Before shooting rays for a given pixel coordinate, the q-buffer is consulted and if the entry for that co-ordinate is less than the number of rays to be shot, D rays are shot, where D is the difference between these two values. The q-buffer entry at this pixel is then

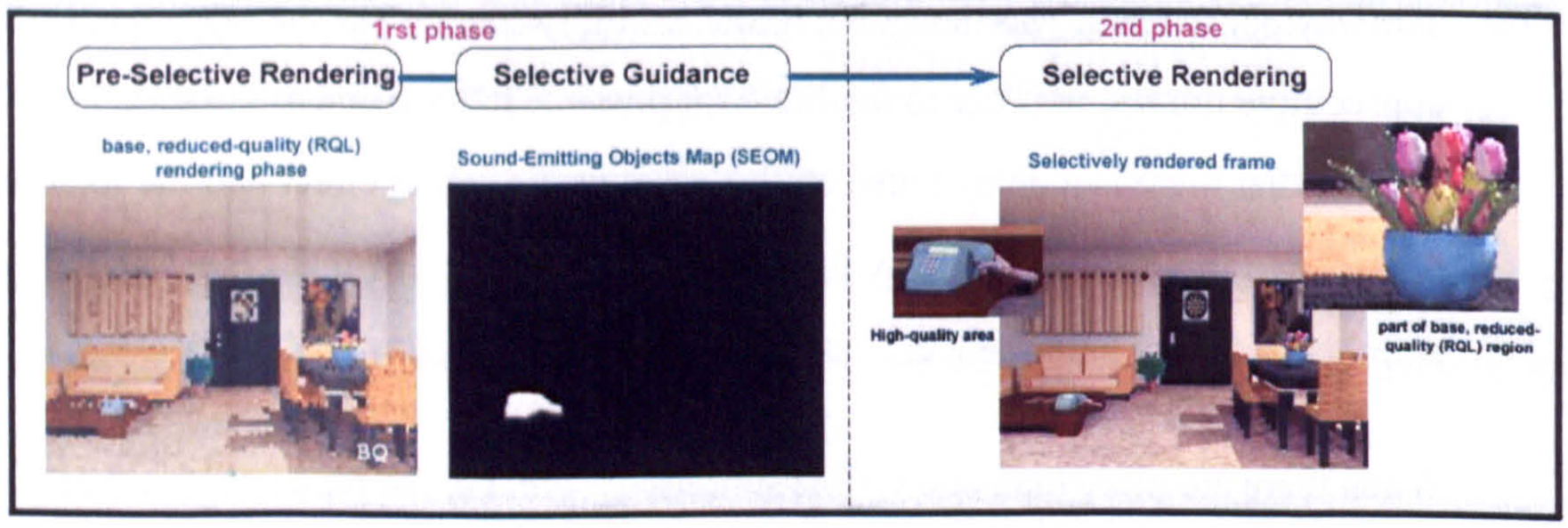


Figure 6.2: Our Selective Rendering Pipeline.

levels. As we mentioned above, our selective renderer is an extended version of the rendering engine of the Radiance Lighting Simulation package [242] to which a number of user-adjustable parameters were defined:

- a base, reduced, quality level (RQL) and a high quality level (HQL) for each frame
- the size of the area around a SEO to be rendered at the gradient quality (quality proportional to a pixel's distance from the boundary of the SEO)
- a tag to characterise each frame as High Quality (HQ), for the case that there are no active SEO, or Selective Quality (SQ): base (low) quality images with high quality rendering around the sound emitting objects.
- for frames tagged as selective, a list of SEO objects for each of these frames is additionally given by the user.

Since our selective renderer is primarily created for animations, an global animation file maintains a list of all frames, the RQL and HQL for each frame, the frame tag (selective or high quality), accompanied by the list of active SEOs if the frame is tagged as selective.

In the case of a frame tagged as high quality, all the pixels of that frame are rendered at high quality (HQL). For frames that are tagged as selective, during the first phase the number of rays corresponding to the base quality (RQL) are shot, while in the second phase

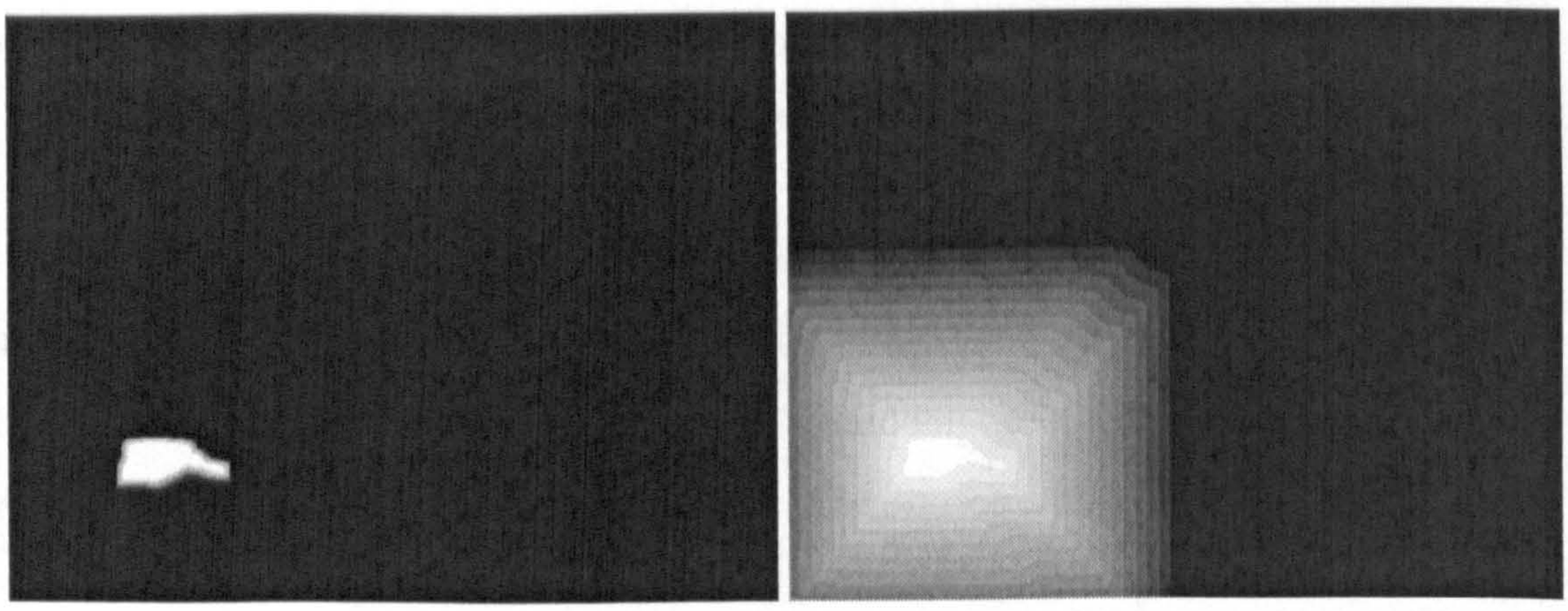


Figure 6.3: Example visualisations of the q-buffer, as maps depicting the quality gradient around the sound emitting object (phone), for a frame selectively rendered using high detail insets of size 150×150 (left) and 350×350 (right). In the maps, white represents the highest quality and black the base (reduced) quality, 1 ray per pixel.

incremented to the new value to reflect the new quality of the pixel. If there are no active SEOs, and thus the difference D is negative or zero, no further action is taken. Figure 6.3 depicts example visualisations of the q-buffer as grey-scale maps.

For border pixels, the renderer degrades the quality around the border of the distractor objects within the user-defined radius. This option provides a method for rendering at high quality an area larger that the size of an object. For each pixel within this radius, the desired quality in rays per pixel is calculated and the q-buffer is consulted. If the calculated quality is greater than the corresponding q-buffer entry, the difference in rays is shot and the q-buffer entry is updated, in the same way as described above. This method ensures that the renderer will not shoot more rays than necessary to pixels lying within the degradation radius of more than one SEOs.

Implementation issues

Our selective renderer is simple to implement and also introduces a number of novel concepts, such as a) the selective rendering of an area larger than the traditional foveal angle, in order to include whole SEOs and their surrounding pixels, b) the quality gradient for the surrounding pixels of a SEO and also c) the ability to render a frame at the reduced base quality, even when there are no active SEOs on screen, but the viewer may be distracted by peripheral auditory stimuli.

Furthermore, our approach improves the traditional map-based selective rendering methods, which require a pre-processing step to calculate the map (identify the important pixels), as our selective renderer is just fed with the list of sound emitting objects per frame and it identifies the corresponding pixels in image space, in order to render them at the predefined high quality.

6.2.2 Participants

120 participants from the undergraduate and postgraduate student population volunteered to participate in this study. The vast majority of the participants had attended a Computer Graphics course and were very familiar with concepts such as rendering quality, artifacts due to reduced quality (aliasing, flickering etc.). Ages ranged from 20 to 33, with an average age of 25. The participants were randomly divided across four test groups, as described below. They all had either normal or corrected-to-normal vision and they did not report any hearing impairment. The participants had to be totally naive as to the purpose of the experiment and therefore each contributed to only one of the conditions.

6.2.3 Design

For our experiment we used an independent samples design. Dependent variable was the perceived rendering quality and independent variables were a) the actual quality at which the animation was rendered, either high quality or selective quality, and b) the auditory background of the animation. Regarding the latter, two conditions were considered: "Sound Effect"and the control, "No Sound" condition. These conditions coupled with the two levels of rendering quality gave four independent test groups.

6.2.4 Equipment and materials

The test environment was the same as the one used in the previous two experiments, described in section 5.2.3. The subjects were seated at normal viewing distance from the monitor, approximately 60 cm, and had their head rested on a chin rest.

The visual stimuli used in the study were based on the walk-through of a 3D interior scene (see Figure 6.4), rendered with our Renderer at 720×540 pixels resolution at two different qualities:

High Quality (HQ): Entire image rendered at high quality (16 rays per pixel)

Selective Quality (SQ) for the frames with sound effects: Reduced quality (RQL) images with High Quality rendering around the sound emitting object(the phone).

In the SQ case, 16 rays per pixel were shot to the area which included the sound emitting object (phone) and its neighbouring pixels, while for the rest of the scene objects only 1 ray was shot per pixel. Figure 6.5 shows the pixels of one of the selectively rendered frames that were rendered at higher quality. Figure 6.6 shows the obvious quality

difference between scene details rendered at high and low quality, respectively.

A pilot study with six subjects was conducted before the main experiment to determine whether a viewer would notice the difference between an animation rendered at HQ and another animation entirely rendered at the lowest quality. The visual defects in the latter, and especially the visual flicker due to the low rendering quality, were apparent to all the people who participated in the pilot study.

All the frames of the first animation used in the main experiment, were rendered at HQ (*HQ_Anim*). For the second animation we reused the HQ frames of the former for the "silent" parts (frames without any SEOs) and rendered the frames with sound at SQ (*SQ_Anim*).

No compression was applied to our animated sequences to avoid various visual defects that appear as a result of video compression encoders. Each animation was 20 seconds long, within the limits of the human working memory span which is approximately 20 seconds [165], and was displayed at 24 frames per second.

Because of the visual content of our 3D scene, the sound effect used for the audiovisual animations was the sound of a ringing telephone, lasting for 3 seconds. We manipulated the left and right channels of the headphones in order to help the viewers locate the phone faster and associate it with the ringing sound they heard. As the phone in our scene was located on the left of the walkthough path, the volume of the left headphone channel was higher compared to the volume of the right channel.

To minimise the visual saliency of the phone, we made it a pale greyish blue. The right hand side of the scene contained more highly salient objects than the area around the phone, and more specifically, a bowl with highly salient brightly coloured flowers, a painting depicting a landscape, a dining table and chairs casting strong shadows on the floor. In the central part of the scene a dartboard of high spatial frequency and a plant with

Figure 6.4: Rendering Quality Experiment - Single frame from the walkthrough in the 3D scene which represented the visual stimuli for our experiment.

polished leaves stood out. This way, if the spatial locations in the scene competed for saliency, the area of the telephone should not stand out and persist over the others. Figure 6.7 shows the saliency map for one of the frames that the sound distractor was active, generated using the Itti & Koch method [93]. The brightest areas in the map represent the areas of greatest saliency.

6.2.5 Procedure

Every experimental session lasted for approximately 5-10 minutes. Each participant was tested individually and was allocated to one of the conditions randomly. All subjects were instructed that they would have to watch carefully an animation depicting a walk-through in a 3D scene and when it finished they would be handed a questionnaire. No hints were

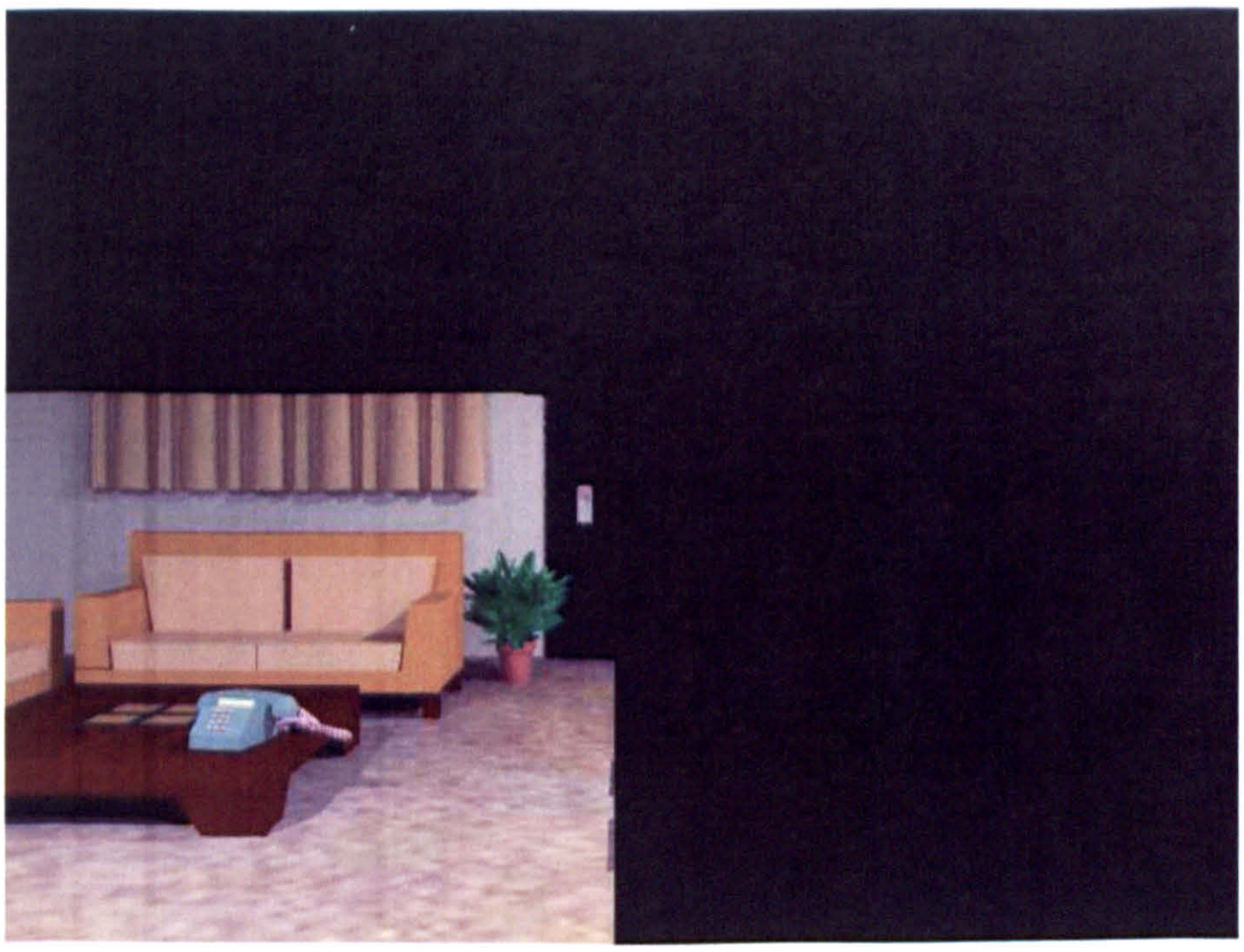


Figure 6.5: The area, 350x350, rendered at higher quality in one of the selectively rendered frames. In this area, the sound emitting object is rendered at the highest quality, 16 rays per pixel, and the quality of the surrounding pixels is gradually reduced as their distance from the boundary of the object increases.

Figure 6.6: Rendering Quality Experiment - Close up of scene details rendered at high quality (16 rays/pixel) and reduced quality (1 ray/pixel), respectively.

given to them regarding the areas in the 3D scene or the features of the animation they should focus on. They were also told that the animation might have sound which would

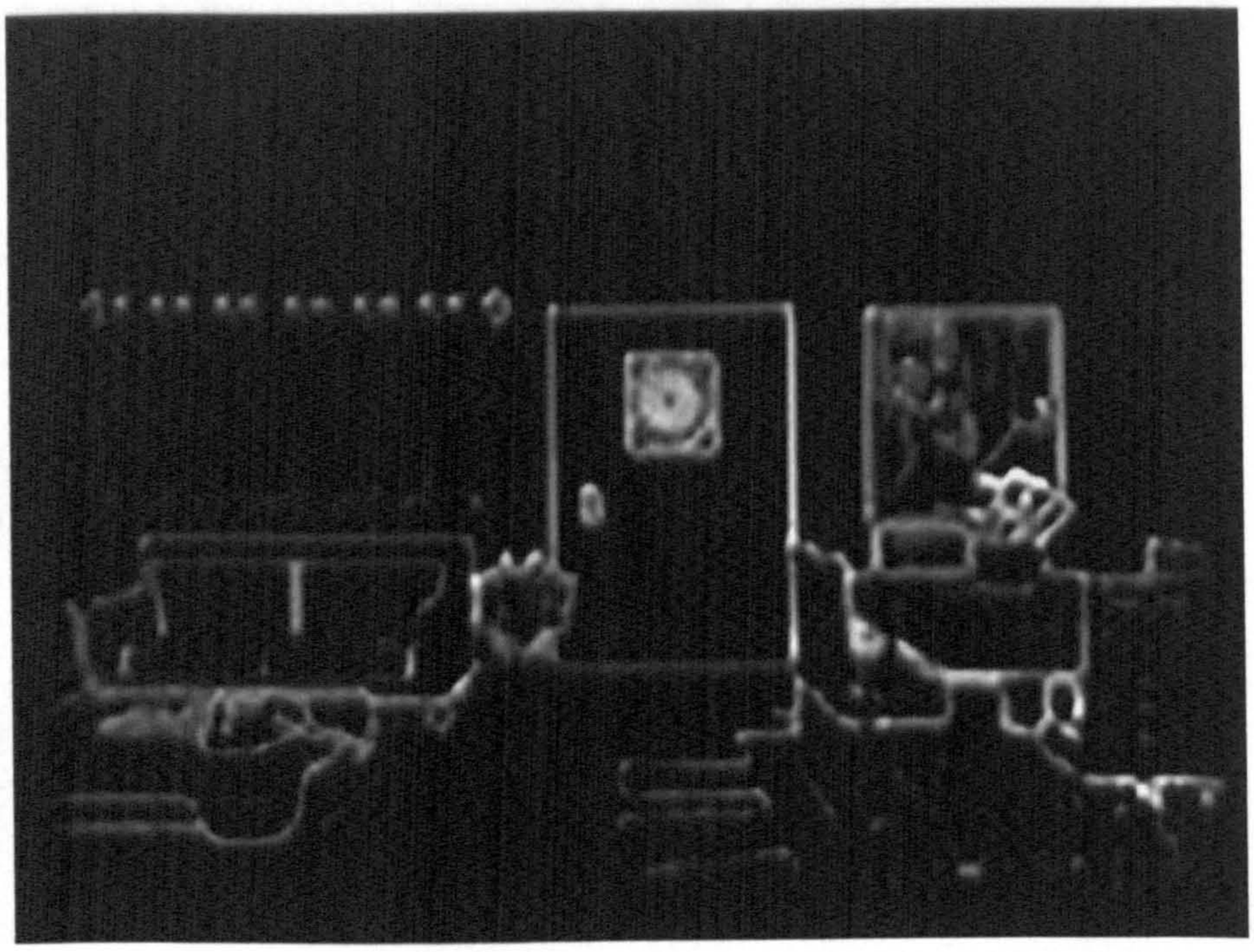


Figure 6.7: Saliency map of the example frame depicted in Figure 6.4. The brightest areas in the map represent the areas of greatest saliency.

be delivered to them through headphones, so they should have the headphones on while watching the animation.

Each subject watched either the *HQ_Anim* or the *SQ_Anim* animation, which was silent or contained the sound effect of the ringing phone. When the animation finished each viewer was shown two still images on the computer screen. The two images represented the same frame of the animation (Figure 6.4). They differed only in the rendering quality: one was rendered at the high quality (HQ) and the other was rendered at the SQ (high quality around the phone). Since the animation was displayed full-screen, the images were also displayed full screen and the participants had to switch from one image to the other to be able to examine them both. The questionnaire included the question: "Which of the given two images do you think was a frame of the animation you have just watched? To decide, think about the rendering quality (possible aliasing effects, jagged lines etc.) at the specific point of the animation that this frame was displayed". To avoid experimental bias, the order in which the two frames were shown to each viewer was randomized.

Experimental Results			
Rendering Quality		**No Sound**	**Sound Effect**
High Quality Animation	Correct:	30	30
	Incorrect:	0	0
Selectively Rendered Animation	Correct:	20	8
	Incorrect:	10	22

Table 6.1: The results from our experiment summarised across the conditions.

6.2.6 Results

Each high-quality frame, rendered by shooting 16 rays per pixel for every pixel, took on average 18.22 minutes to render on an Intel Pentium 4, 2.4 GHz Processor, while the selectively rendered frames were each rendered on average in only 4.75 minutes (rendering speedup: $\approx 383\%$). In each selectively rendered frame, the sound 'distractor' was rendered at 16 rays per pixel and for the surrounding area the rendering quality degraded gradually up to a base quality of 1 ray per pixel. Had we chosen a smaller high-quality area, the speedup would have been larger. For instance, the time for a selectively rendered frame when using a 250×250 area, which is also quite large, was on average 3.5 minutes. The corresponding average time for 150×150 was 2.37 minutes, just 13% of the time a high quality frame would take to render, accomplishing a speedup of approximately 769%.

Table 6.1 summarises the results from our experiment. The initial examination of these results, reveals that all 60 participants who watched the animation that was entirely rendered at the default high quality (*HQ_Anim*) correctly identified the rendered quality, regardless of the auditory background of the animation. On the other hand, only 26.6% (8 out of 30) of the participants who watched the selectively rendered animation in the presence of

the ringing sound, noticed the degradation in the quality. The performance increased to 66.6% for the participants who watched the silent *HQ_Anim* animation. The other 33.3% did not notice any degradation in the quality, despite the fact that there was no sound that could potentially attract their gaze toward the sound source. Maybe this happened because during the 3 seconds that this degradation lasted they could, by chance, be looking toward the part of the scene that was rendered at higher quality.

Since for the high quality animation the performance was 100%, for both the "No Sound" and "Sound Effect" conditions, no further statistical analysis was needed to confirm that in this case the presence or not of the sound made no difference to the viewer.

For the SQ case, the participants of the "No Sound" condition gave more correct answers that the participants of the "Sound Effect" group, but the results needed to be further analysed in order to determine whether there was a significant performance difference as a function of the auditory background, that is whether the difference between the two groups was sufficiently great to be unlikely to have occurred by chance.

Since the response of the subjects was binary, the appropriate method for the statistical analysis of the results was the Chi-square test (X^2) for comparing frequencies of categorical events between multiple groups. The use of parametric tests, such as Analysis of Variance (ANOVA) and t-test, was not appropriate since these tests require data sampled from normally distributed populations.

The Chi-Square test (for 2 Independent Samples) for the unpaired comparison between the two groups gave a very significant result even at the 0.005 level of risk ($X^2 = 8.103 > P_{a=0.005} = 7.8794$ for df=1, therefore $p < 0.005$), revealing a strong interaction between the auditory background of the selectively rendered animation and the perceived rendering quality. Therefore, our initial hypothesis that it would be more difficult for subjects to notice rendering quality variations in the presence of sound than while watching silent animations, was confirmed.

6.2.7 Discussion

Our approach works by identifying the area(s) where the user will probably fixate, determined by the presence of sources of unexpected intense sound effects, rendering the corresponding pixels to high quality and significantly dropping the quality for the rest of the scene. Our experimental results show that observers generally fail to notice the quality variations within the scene while the sound lasts, whereas without the sound the change from the high to the low quality is by far more noticeable [129].

Preliminary tests using an eye tracker indicated that, although the scan path for each viewer is different, the eye scans corresponding to the time interval that the ringing sound is audible fall onto and closely around the phone. Refer to Figure 6.8 for example scan paths and fixations on the phone from our preliminary eye tracking tests. This finding supports our hypothesis that the sound captivates a viewer's visual attention toward the origin of the sound. Nevertheless, a more formally designed eye tracking experiment was necessary in order to validate our findings and further investigate the strength of the phenomenon against task objects and competing highly-salient objects present in the scene.

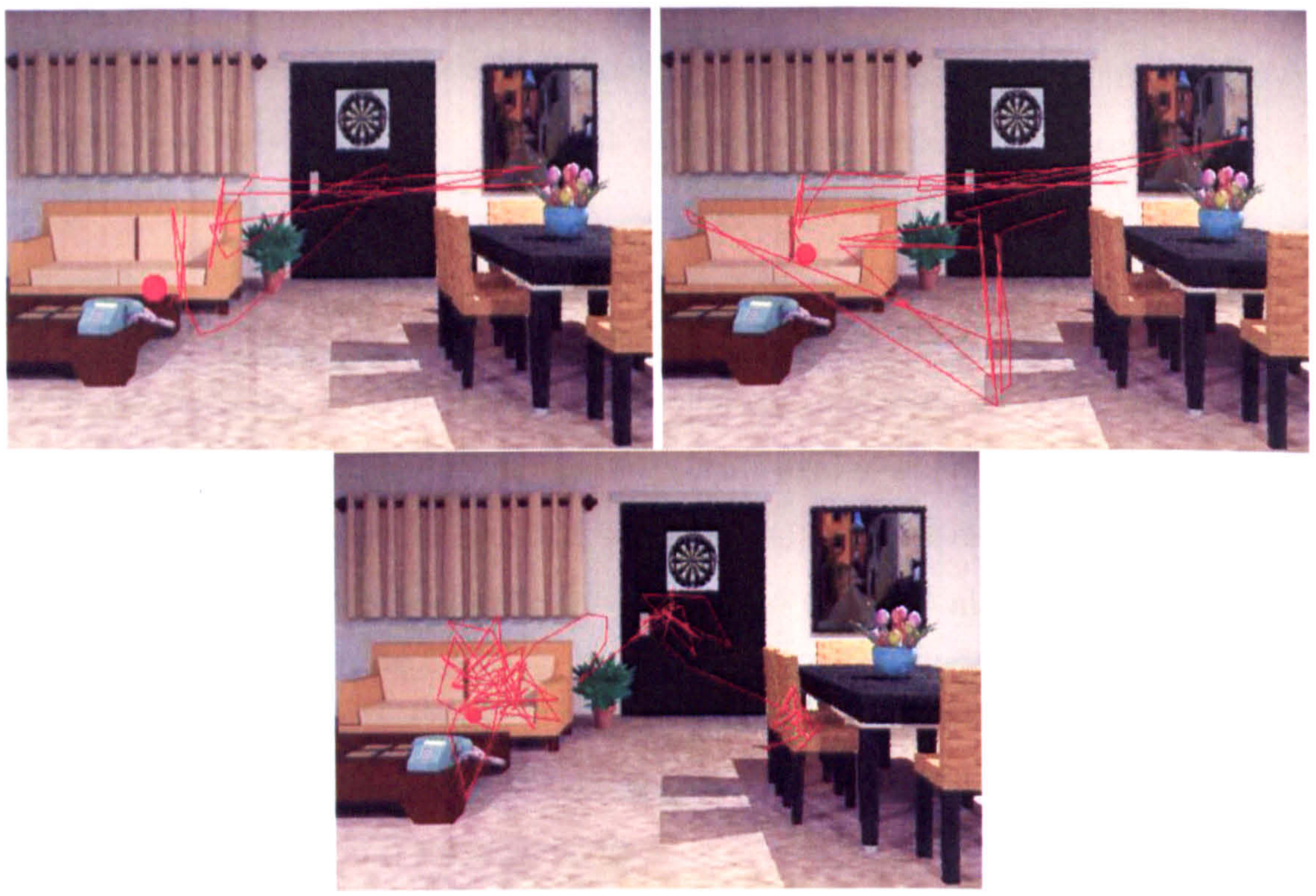


Figure 6.8: Preliminary eye tracking tests: Example saccades and fixations on the phone while the ringing sound was audible.

6.3 Eye tracking Experiment

Based on the very promising preliminary eye tracking results we performed a formally designed eye tracking experiment in order to validate our findings and further investigate the strength of the phenomenon against task objects and competing highly-salient objects present in the scene.

6.3.1 Experimental Set Up

The set up being used in our experiment involves the use of an eye tracker to track the eye's point of focus. In our experiment, we used an a Tobii x50 desk mounted eye tracker

(Figure 6.9), which allows subjects to move their head while they view stimuli, an animation in our case, on a computer monitor screen. A high-resolution camera with a large field-of-view is used to capture images of the subject's eyes. Near infra-red light-emitting diodes (NIR-LEDs) are used to generate even lighting of the viewer and reflection patterns in the eyes of the subject. From the data received the eye tracker computes and records the positions subjects watch on the screen.

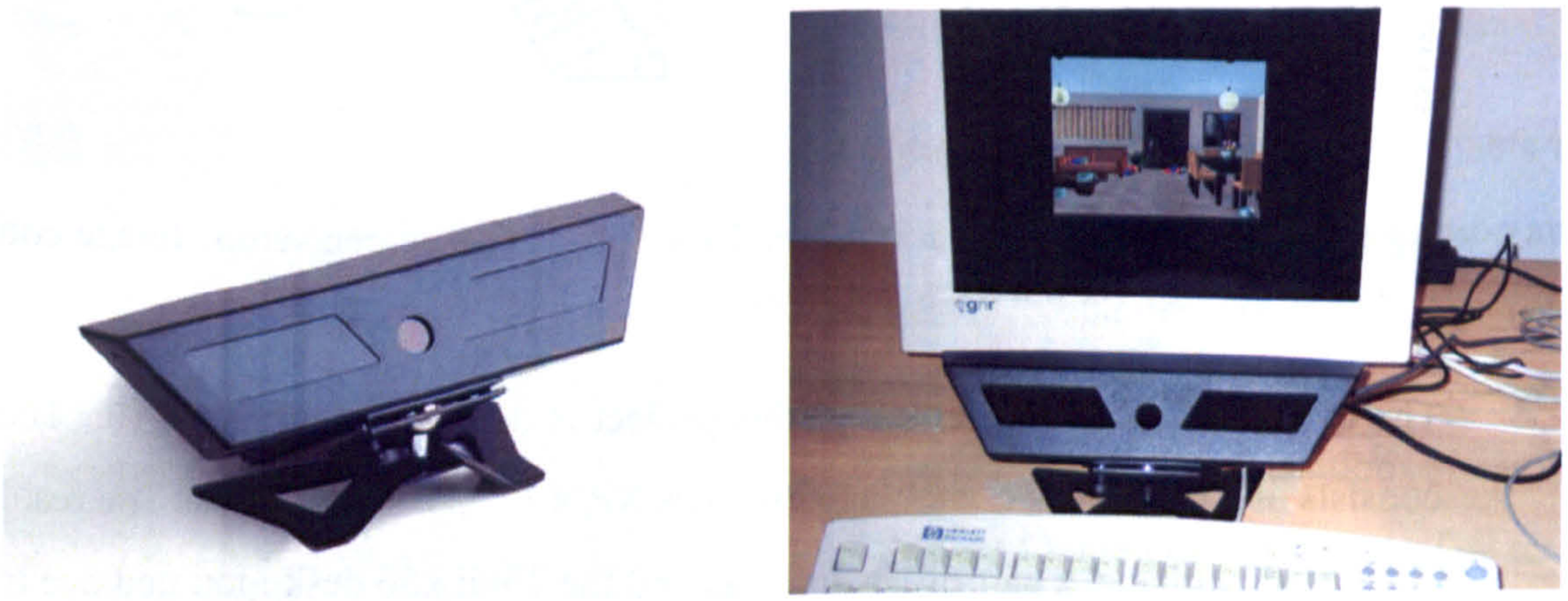


Figure 6.9: The Tobii x50 desk mounted eye tracker.

Figure 6.10: The overall hardware experimental setup for the eye tracking experiment.

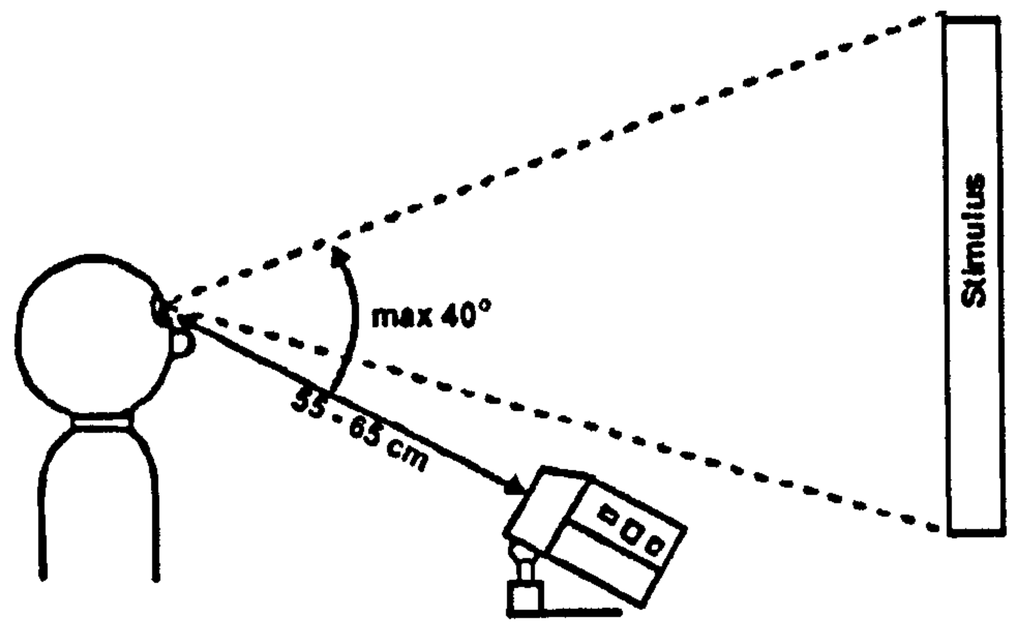


Figure 6.11: Tobii x50 and a monitor, TV or projection screen setup. Image courtesy of *Tobii Technology* (http://www.tobii.com)

The overall setup for the model of this project is depicted in Figure 6.10. The system consists of a powerful PC running Windows XP, a 17" TFT monitor set at a resolution of 1280 x 1024 pixels, a pair of headphones and the Tobii x50 desk mounted eye tracker to track the points where the eye was fixating. The eye tracker shall be positioned straight in front of the stimuli and at a particular angle below the user, see Figure 6.11. Tobii x50 provides binocular tracking at a very high spatial accuracy: 0.5 degrees. The term "accuracy" refers to the deviation between the measured and actual gaze point of the user. One degree of accuracy corresponds to an average error of about 1 cm between the measured and actual gaze point at 50 cm distance between the user and the object viewed.

The field of view of the camera is 20 × 15 × 20 cm (width × height × depth) at 60 cm from the screen. It is enough that one of the eyes is within the field of view, which gives an effective tolerance to head-motion of about 30 x 15 x 20 cm. This is enough to compensate for head positions which normally occur when sitting in front of a computer screen [225]. It has a stable frame-rate of 50 Hz, i.e. 50 gaze data points per second are collected for each eye. Each gaze data point is provided with a time stamp in milliseconds, which describes when each camera image of the eyes was taken. The time-stamp is accurate to about +/- 5 ms [225].

Before starting an eye tracking experiment the eye tracker has to be calibrated in order to learn the characteristics of the eyes of each user. A dot is presented successively and randomly at different positions on the monitor screen of the subject PC. The subject is asked to fix his gaze on this dot every time it appears on the screen (see Figure 6.12). The calibration procedure is fully automatic, lasts 30-60 seconds only, filters out bad calibration points and also provides indications of the resulting calibration quality. Only one calibration per subject is required.

Tobii x50 eye tracker is almost completely non-intrusive and therefore no helmets or markers of any kind are required. Moreover, its very good tolerance to head-motion and its excellent drift reduction removes all need for chin-rests and other restraints. For Tobii x50, head-motion compensation is very accurate, with an error that is less than 1 degree across the entire field of view of the camera. This includes head translations sideways and up or down as well as movement back and forth and large head rotations. The term "drift" describes the deterioration of a calibration that occurs over time. This is caused by changes in characteristics of the eyes that are caused by change of pupil size (for example, because of changes in surrounding light conditions or screen illumination levels) or if the eyes become dry. For Tobii x50 eye tracker, drift over long time periods and great differences in light conditions range from 0 to 2 degrees for each eye individually. By using "binocular averaging" a large portion of horizontal drift effects can be removed and drift effects for most users are reduced to less than 0.5 degrees [225].

The monitor in our experiment was placed at a distance of around 60 cm from the view-point of the subject. The experimental stimuli and the stimuli for calibration were displayed on the monitor. The same PC was responsible for recording the data received by eye-cameras, and for calculating the gaze positions, saccade lengths, fixations etc. After each experimental session, the eye movement data would be analyzed to determine the impact of the task and the sound on saccade and fixation durations and saccade lengths.

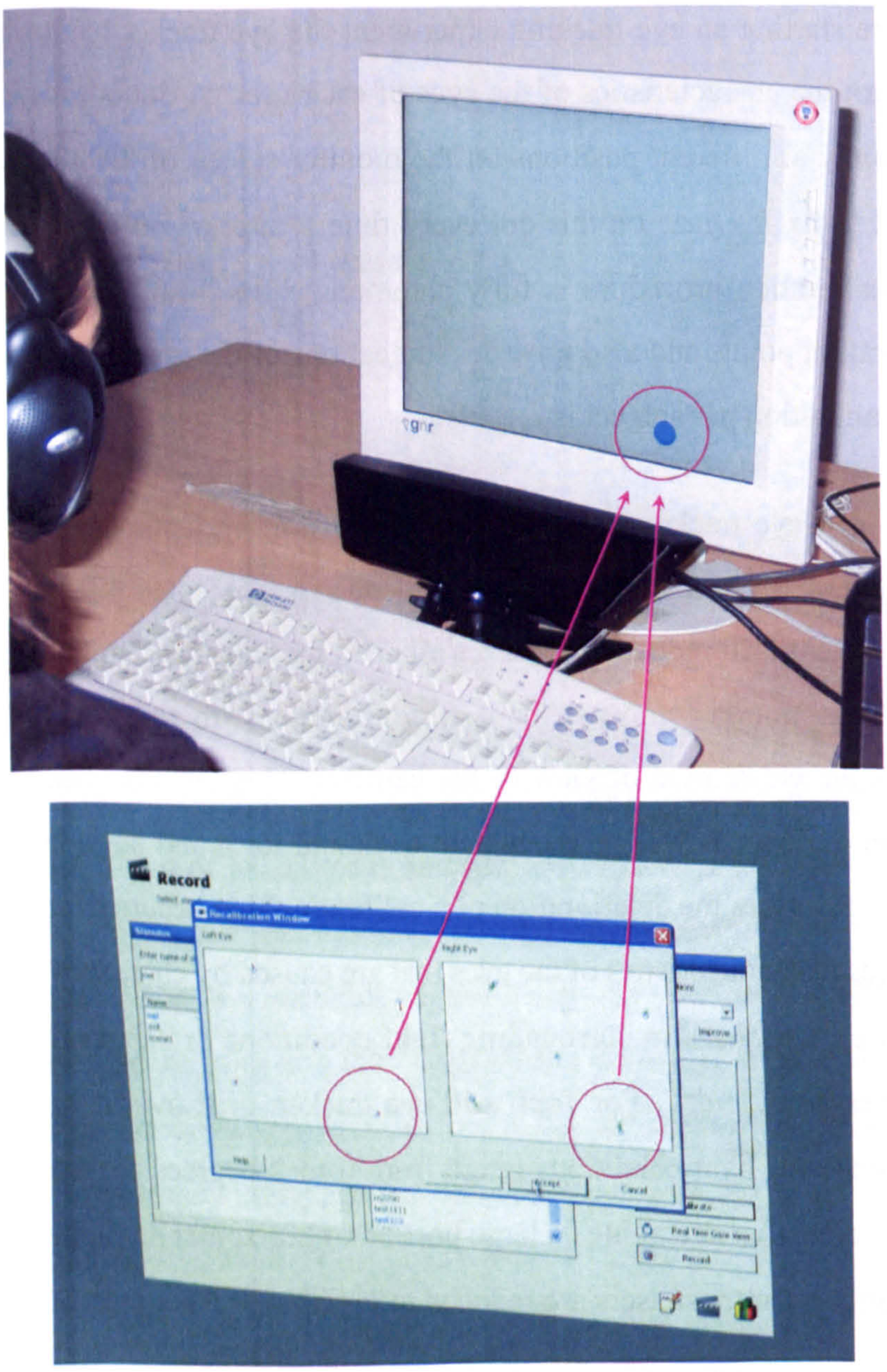


Figure 6.12: Snapshots of the eye tracker calibration procedure. The subject is asked to fix his gaze on the dot, which is presented successively and randomly at different positions on the monitor screen, every time it appears. The bottom picture is an example of a very good calibration quality. The left panel includes the calibration points for the left eye and the right panel the calibration points for the right eye. Where the calibration quality is very good, the green and red marks coincide.

6.3.2 Participants

36 participants from the undergraduate and postgraduate student population volunteered to participate in this study. Ages ranged from 20 to 30, with an average age of 22.8. The

participants were randomly divided across four test groups, as described below. They all had either normal or corrected-to-normal vision and they did not report any hearing impairment. The participants had to be totally naive as to the purpose of the experiment and therefore each contributed to only one of the conditions shown in Figure 6.13.

6.3.3 Design

In this experiment the participants' eye movements (fixations and saccades) were measured as they performed (a) free viewing of a rendered animation or (b) a search task while watching the same animation, according to the condition they contributed to.

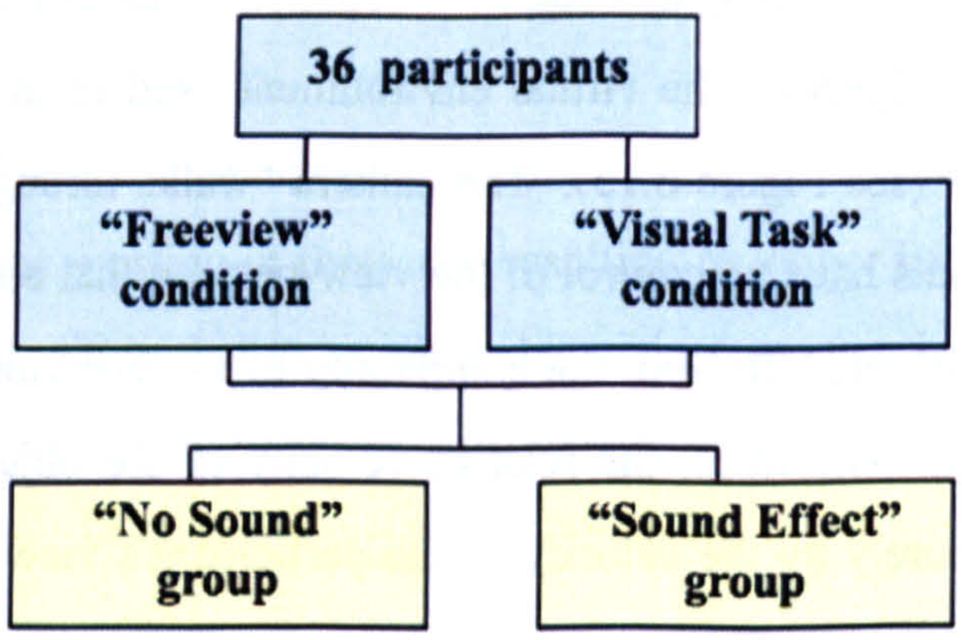


Figure 6.13: The Conditions tested in the Eye tracking Experiment.

Each condition included 18 subjects. The first condition consisted of subjects who had no task ("Freeview" condition), they just had to free view the animation, and the second group included the participants who had to perform a "virtual visual search" task ("Task" condition). In the search task they had to count separately the number of instances of the given target objects (task objects) existing in the 3D scene. The target objects were either an orange cylinder and a red ball or a red cylinder and a blue ball. There were 2 orange cylinders, 5 red balls, 3 red cylinders and 4 blue balls in the two rooms comprising the

scene (see Figure 6.15).

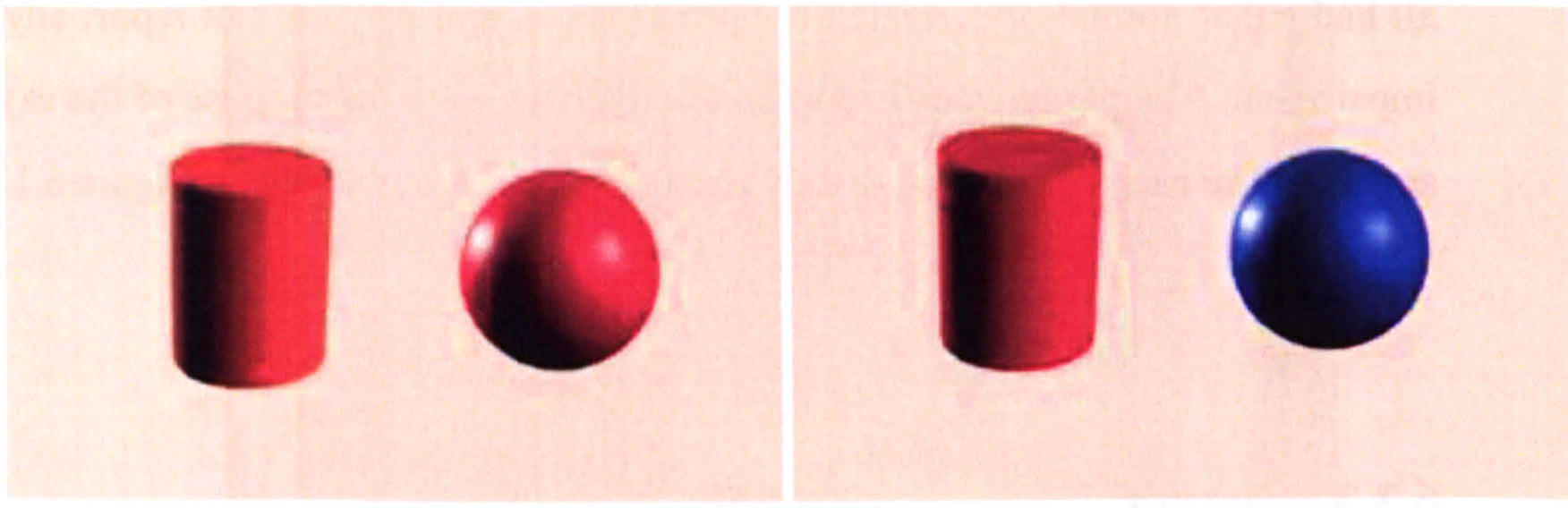


Figure 6.14: The task objects combinations used for the "Task" condition in our experiment.

Our paradigm is similar to the traditional visual search paradigm in that participants are required to search for the specified target items in a display. The virtual component of the task requires that participants search a complex three dimensional virtual environment for the target objects. The virtual environment used in the experiment is a two-room home interior (see Figure 6.15). The camera "walks through" the virtual environment. The participants have no control of the view-port, so that all participants see exactly the same visual stimuli. To limit the complexity of the experiment, all objects in the scene are stationary. Therefore, the relative velocity of the object across the visual field is determined purely by the velocity of the participant's view-port (i.e the velocity of the camera movement around the scene).

We chose to use this natural task with rich and complex stimuli rather than a more simplistic virtual search task, so that the results obtained in our experiment will be more likely to generalise to applied settings, for example, architectural walk-through and video games.

The participants of the two conditions were further divided into two subgroups, according to the auditory background of the animation they watched during the experimental session. The first subgroup of each condition consisted of 12 participants who watched the audiovisual version of the animation (it included the sound effect of a ringing phone lasting for 3 seconds, as described in the following paragraph) and the remaining 6 partic-

ipants were assigned to the second subgroup, which was the control, "No Sound" group and included the participants who watched the silent version of the animation.

The recorded eye movement data would be analyzed to determine the impact of sound on saccade and fixation durations and saccade lengths while the sound was audible. In the "Freeview" condition, the scanpaths would provide information on the impact of the sound versus the highly salient objects in the scene, especially the geometric objects which were irrelevant to the rest of the scene contents. In the case of the condition involving the search task, the impact of sound vs the task would be examined.

Based on previous research it was predicted that the viewers performing the search task would fixate longer on the geometrical objects in the scene- red, orange, blue and light green balls, cylinders and cubes- in an attempt to identify among them the target objects they had to count.

It was also predicted that the sound effect of the ringing sound would attract the viewers' gaze toward the phone existing in the scene resulting in longer fixations on it, not only in the "Freeview" condition, but even when the subject was participating in the "Task" condition.

6.3.4 Stimuli

The visual stimuli used in the study were based on the walk-through of the 3D interior scene we used in the previous experiment, to which we added the task-related geometrical objects. You can see example frames of the visual stimuli in Figure 6.15. These objects were geometrical objects- cubes, spheres and cylinders- scattered around the two rooms of the 3D scene. The colours of the task objects were bright red, orange, blue and light turquoise. The scene was rendered at 720 × 540 pixels resolution at high quality for all

pixels.

No compression was applied to our animated sequence. The rendered animation was 27.63 seconds long and was displayed at 24 frames per second.

Again, the sound effect used for the audiovisual animation was the sound of a ringing telephone- two rings lasting for 3 seconds in total- heard from the 20.20 till the 23.21 seconds of the animation. We manipulated the left and right channels of the headphones in order to help the viewers locate the phone faster and associate it with the ringing sound they heard.

Figure 6.15: Example frames from the walkthrough in the 3D scene which represented the visual stimuli for our eye tracking experiment.

6.3.5 Procedure

Every experimental session lasted for approximately 10-15 minutes. Each participant was tested individually and was allocated to one of the conditions randomly.

At the beginning of each session, the eye tracker was calibrated, in order to learn the characteristics of the eyes of each participant. The latter was seated in front of the eye tracker and the subject PC, at a distance of 60 cm from the monitor screen, and was instructed to fix his gaze on the dot that would appear successively on the screen at random positions. In the case of bad calibration points, the subject had to repeat the corresponding fixations on the dot.

When the calibration procedure was completed successfully, the subject was instructed that he would have to watch carefully an animation depicting a walk-through in a 3D scene and when it finished he would be handed a questionnaire. In the case of the "Freeview" condition, no hints were given regarding the areas or objects in the 3D scene he should focus on. In the case of the "Task" condition, he was shown an image of the two target objects he should be looking for in the scene and was asked to count separately the number of instances of these objects during the walk-through. As mentioned above, the target objects were either an orange cylinder and a red ball or a red cylinder and a blue ball. Half of the subjects in the "Task" condition were given the first combination of target objects and the second combination was allocated to the rest.

The participants were also told that the animation might have sound which would be delivered to them through headphones, so they should have the headphones on while watching the animation. Each subject watched the test animation, which was silent or contained the sound effect of the ringing phone. Figure 6.16 shows a participant watching

the test animation. After the end of the test animation, the subjects in the "Freeview-No Sound" group were dismissed without having to answer any questions. The rest of the participants were handed a questionnaire.

Figure 6.16: A participant seated at the eye tracker during the experimental session.

The subjects who belonged to the two groups of the "Task" condition, were asked to fill in the two target objects they were given and the number of instances of each object they had detected in the test scene. The participants who had watched the audiovisual version of the animation (both from the "Freeview" and the "Task" conditions) were asked to identify the direction from where the ringing sound came. Our intention was to find out whether the participants associated the sound effect of the ringing phone with the telephone present in the scene.

6.3.6 Results

One participant from the "Task-Sound Effect" group did not give any fixations during the 3-second period that the sound effect was audible and was excluded from the study. The scan paths for each of the remaining participants during these 3 seconds are given separately in Appendix B.

Even for the "Freeview" condition the geometrical objects attracted a lot of fixations due to their saliency and the fact that they were not related to the rest of the scene objects. The participants tried to look at as many objects and grab as many scene details as possible, in case they were asked memory questions about the scene contents after the end of the animation. Example fixations of of participants assigned to the "Freeview" condition are shown in Figure 6.17.

Figure 6.17: Eye tracking experiment - Example saccades and fixations from the "Free-view" condition.

Participants of the "Task" group, as expected, focussed mainly on the geometrical objects in order to count the instances of the target objects they were given. Figure 6.18 shows example fixations of participants from the "Task" condition.

Figure 6.18: Eye tracking experiment - Example saccades and fixations from the "Task" condition.

A summary of the results across the participants of each group is given in Figures 6.19-6.22. More specifically, Figures 6.19 and 6.20 show all the scan paths for the "Freeview" group overlaid on top of each other (left image) and the fixation hotspots (right), where the deeper the colour of the hotspot is, the more fixations the corresponding area has

attracted. The corresponding Figures for the "Task" group are 6.21 and 6.22.

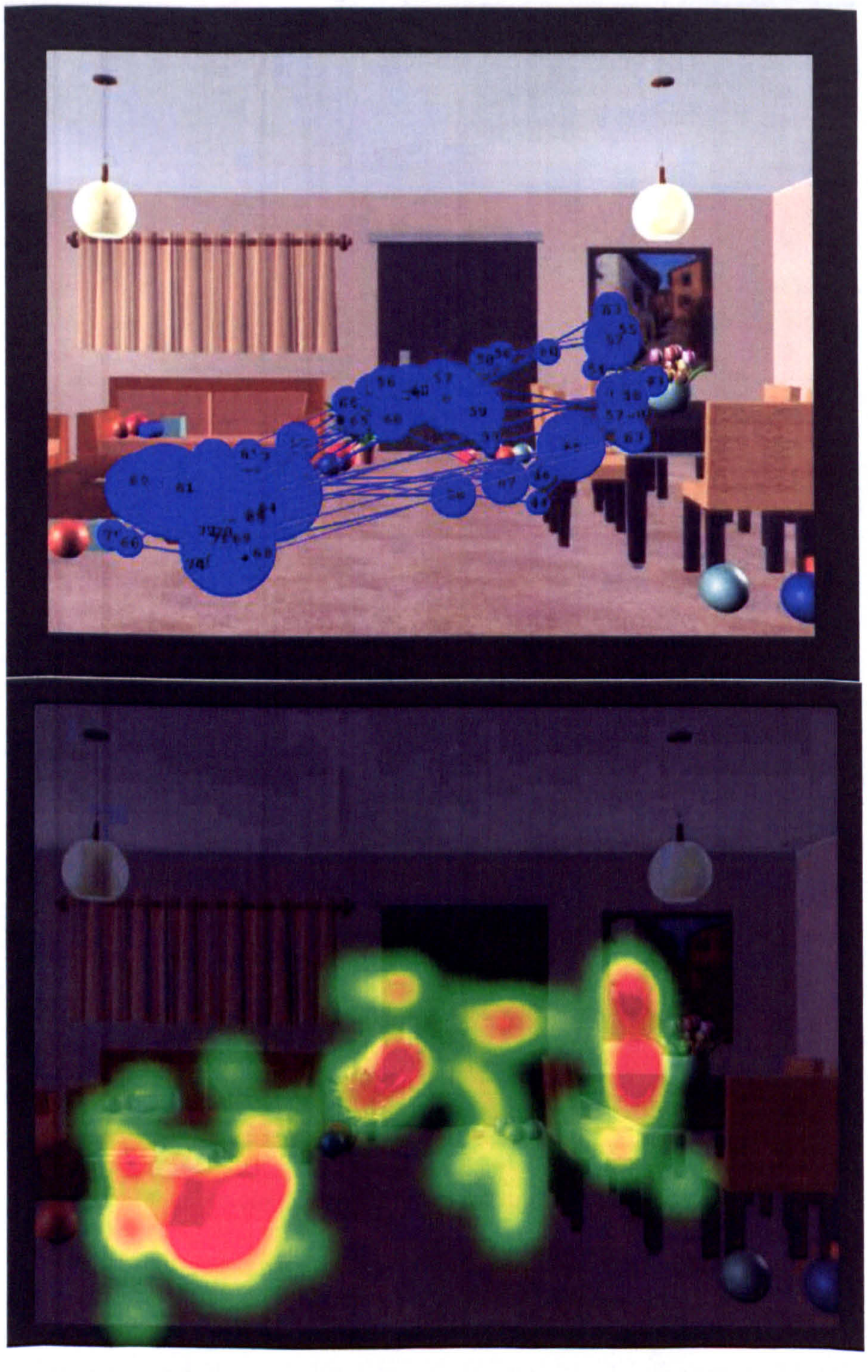


Figure 6.19: Top: Combined scan paths for the "Freeview-Sound Effect" group, regarding the 3-second period that the sound effect of the ringing phone was audible. Bottom: Fixation hotspots for the same time interval. The deeper the colour of the hotspot is, the more fixations the corresponding area has attracted.

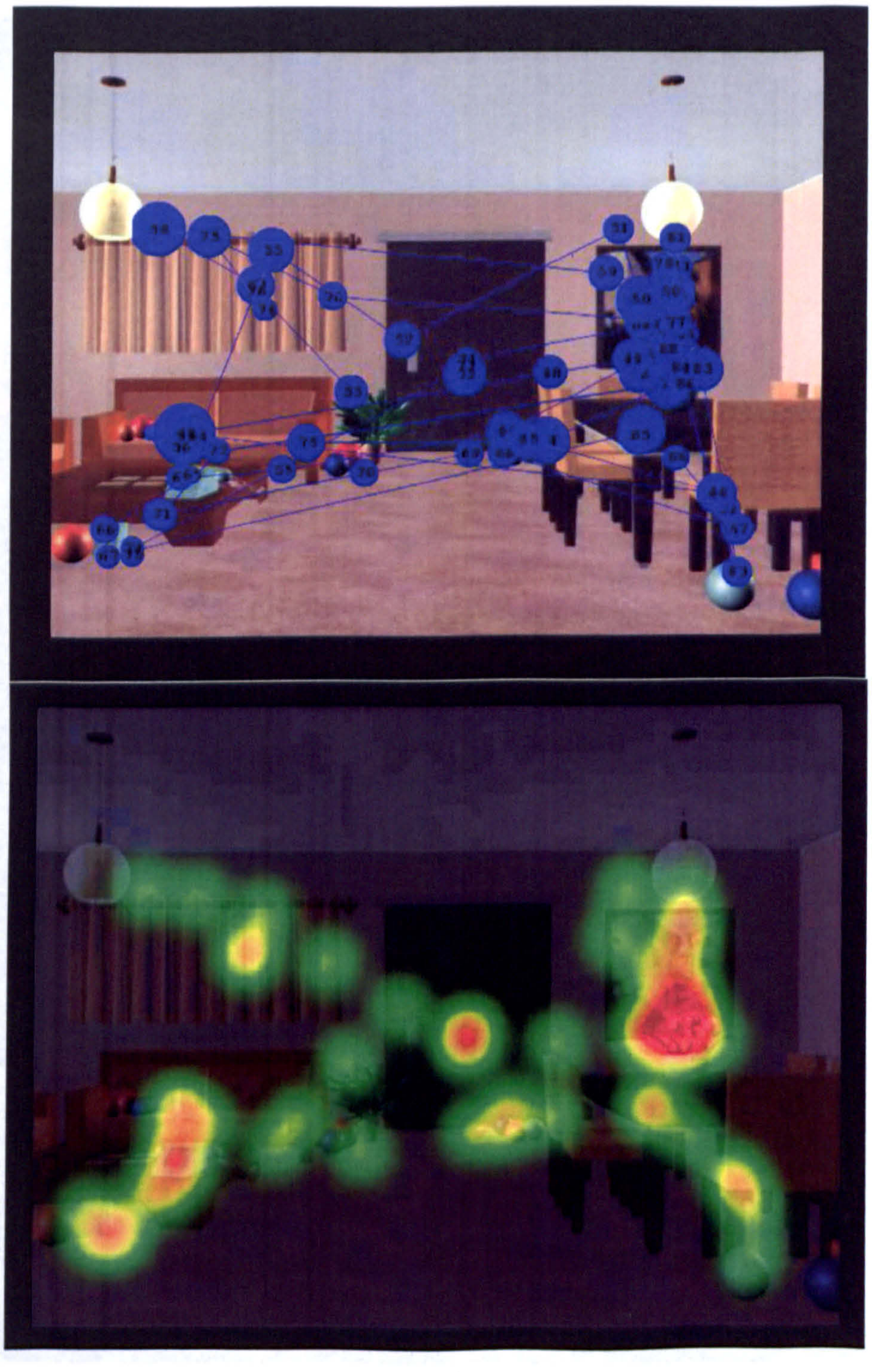


Figure 6.20: Top: Combined scan paths for the "Freeview-No Sound" group. Bottom: Corresponding fixation hotspots.

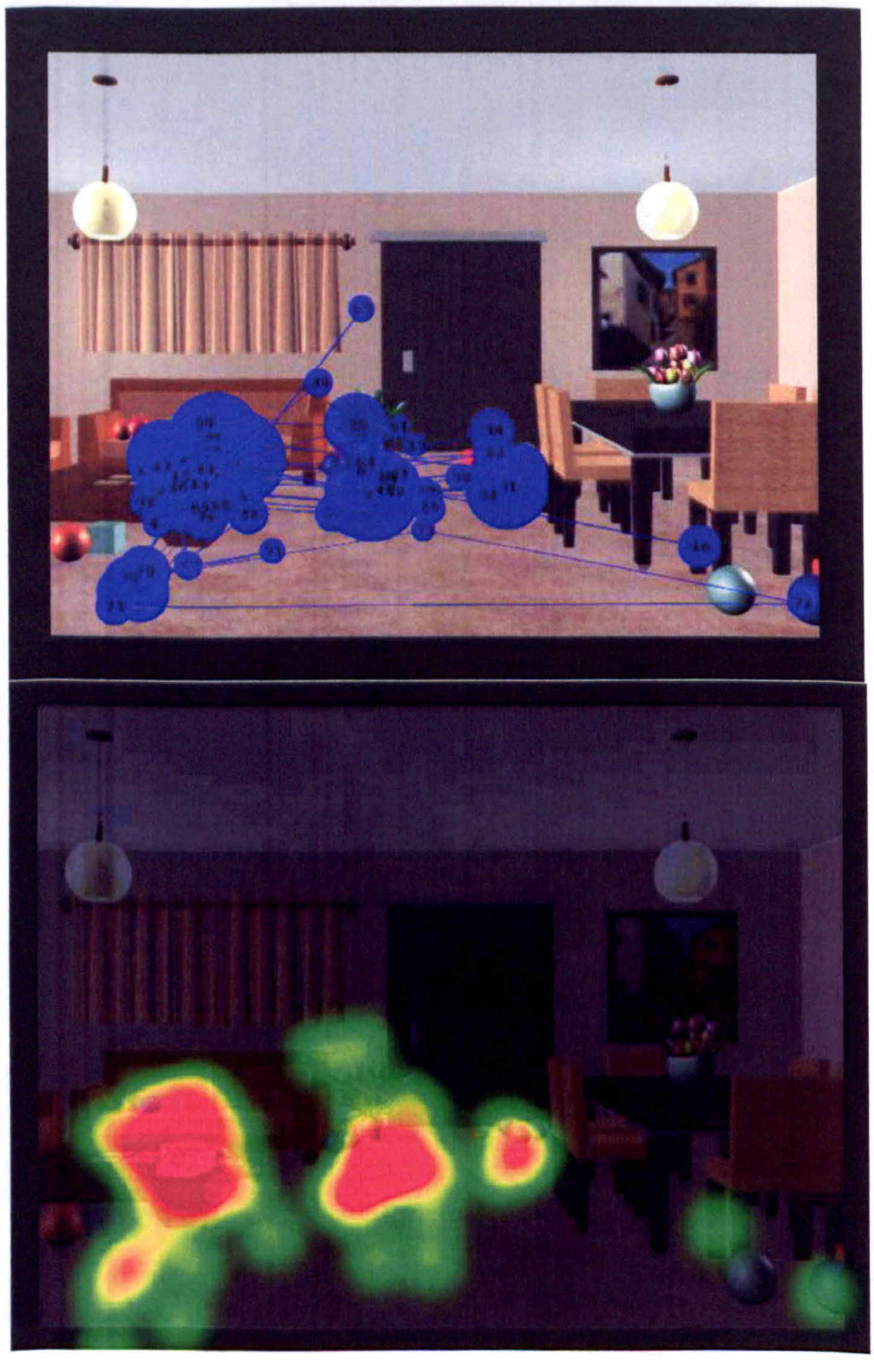


Figure 6.21: Top: Combined scan paths for the "Task-Sound Effect" group. Bottom: Corresponding fixation hotspots.

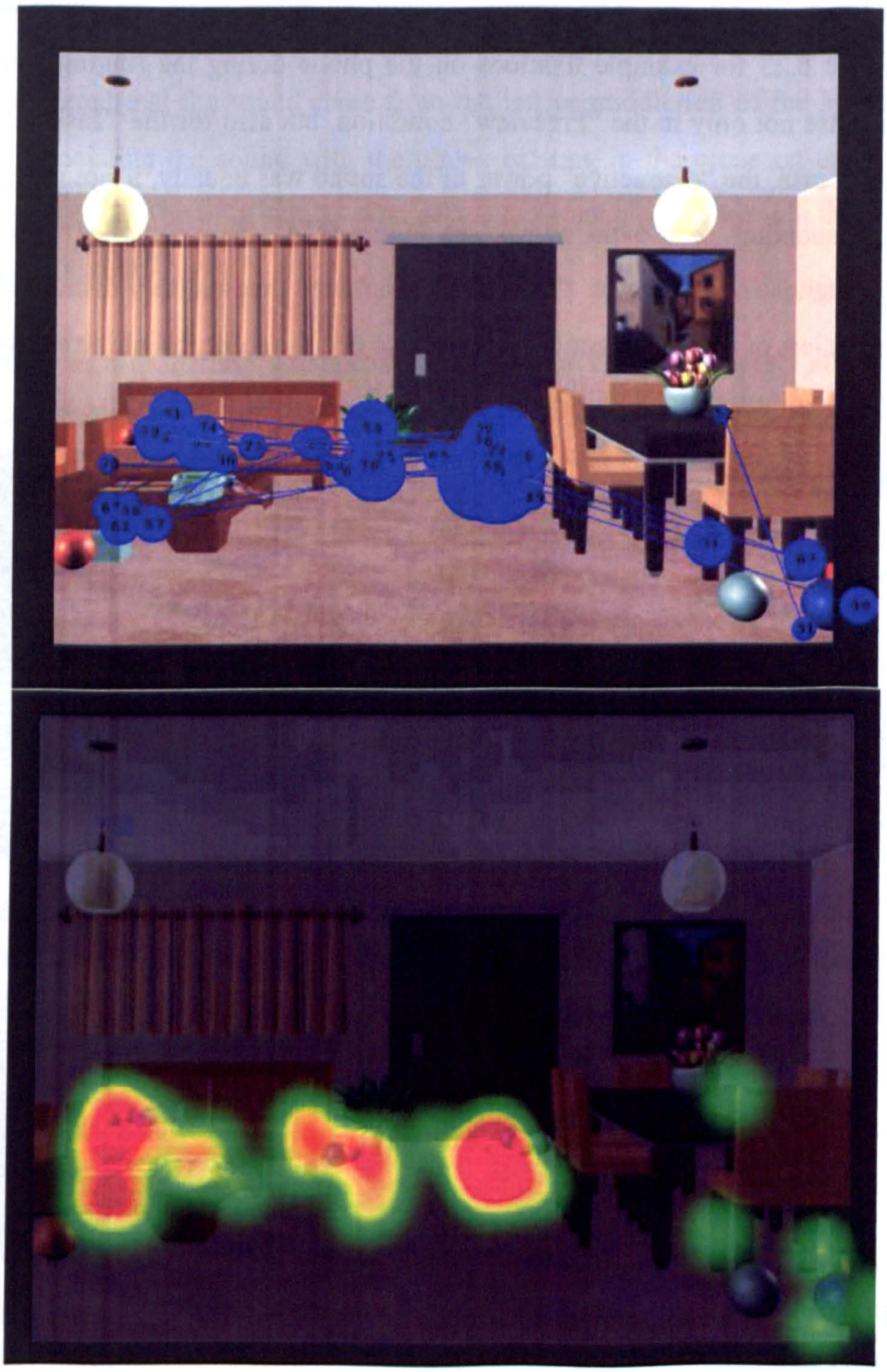


Figure 6.22: Top: Combined scan paths for the "Task-No Sound" group. Bottom: Corresponding fixation hotspots.

The attention of the subjects who watched the audiovisual animation containing the sound of the ringing phone, was attracted toward the phone while the sound was audible (see Figure 6.23 for example fixations on the phone during the ringing sound). This was the case not only in the "Freeview" condition, but also for the "Task" condition. In the latter case, the "distractive" power of the sound was equally, if not more, evident as the corresponding scan paths reveal, see Figure 6.24. This complies with the findings of Tellinghuisen and Nowak (2003) who concluded that auditory distractors are processed regardless of visual perceptual load and also that the ability to inhibit crossmodal influence from auditory distractors is reduced under high visual load [219].

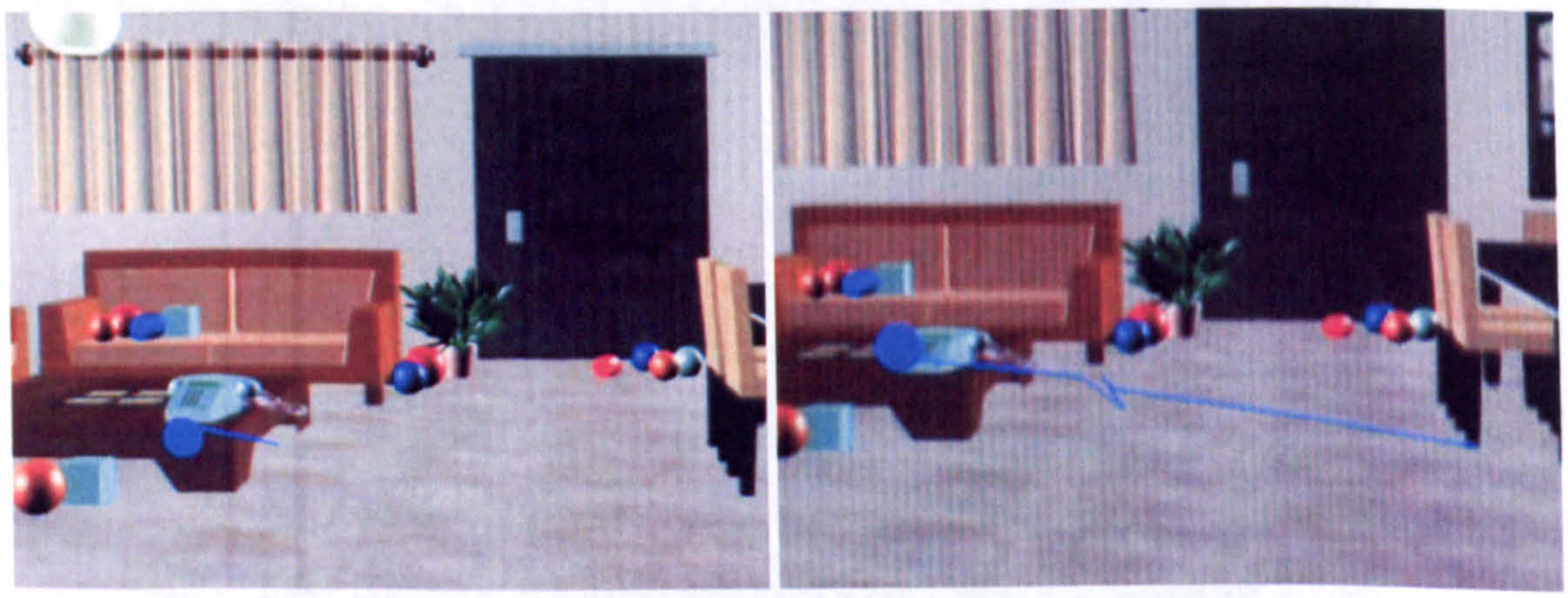


Figure 6.23: Example saccades and fixations on the phone while the ringing sound was audible.

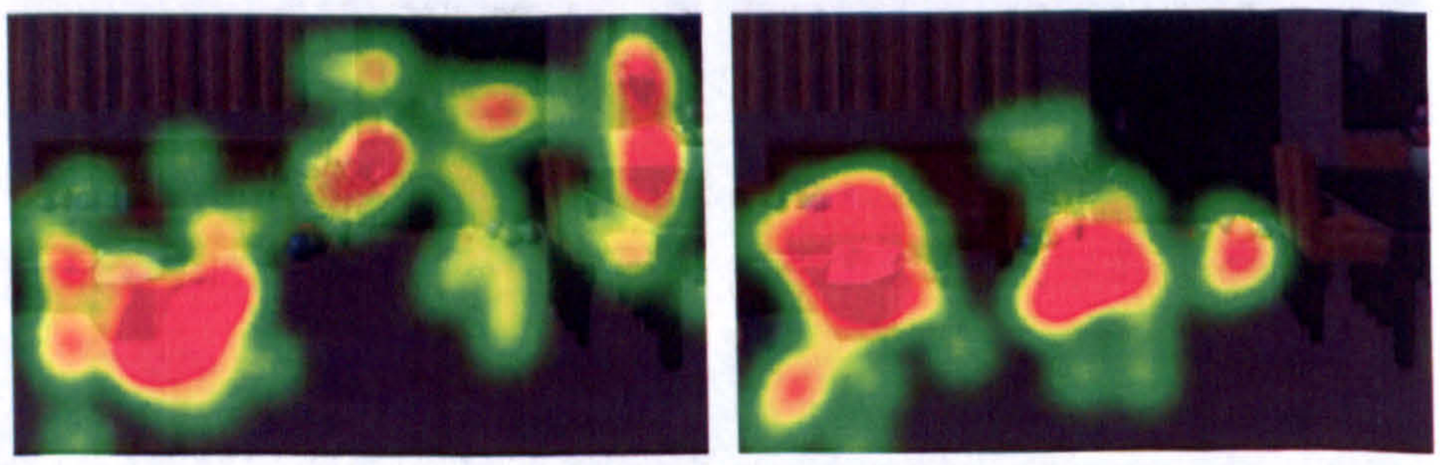


Figure 6.24: Closeup of the fixation hotspots for the "Freeview" (left image) and "Task" (right) conditions, around the time that the ringing sound was audible. The deeper the colour of the hotspot is, the more fixations the corresponding area has attracted

As we mentioned above, the sound effect consisted of two distinct rings and in some cases the gaze of the participants drifted away from the phone when the first ring was over, but

they fixated to the phone again when the second ring was heard.

All participants who watched the audiovisual animation, except for one from the "Task" group, answered that the sound came from the left or middle-left of the 3D scene, apparently associating the sound with the phone existing in the scene on the left of the walkthrough path. The only subject who gave a different answer, thought that the sound came from the right.

The performance in the task was measured as the percentage of correct answers in the search task ("accuracy"). For a subject's answer to be considered correct, both of the target objects should be counted correctly. Out of the 17 participants who performed the task, only three- representing the 17.6%- counted correctly the instances of their target objects in the 3D scene. This low percentage verifies that the task was indeed difficult and required the full attention of the participants.

The *accuracy* as a function of the auditory background is shown as bars in Figure 6.25, as the mean taken across participants. None of the participants in the "Task-No Sound" group correctly counted their target objects, while the "accuracy" for the "Task-Sound Effect" group was 37.5%. Figure 6.26 gives the accuracy separately for the two combinations of task objects. We had expected the participants who watched the silent animation to perform better than the others who heard the ringing sound, but we have to take into consideration that the duration of the sound was just 3 seconds, while the duration of the whole animation, during which the subjects had to count the number of the target objects, was 27.63 seconds. Moreover, the subjects in the "Task-No Sound" group were almost half the number of the participants in the 'Task-Sound Effect" group and therefore we cannot reach any safe conclusion about how performance in the search task varied with the auditory background.

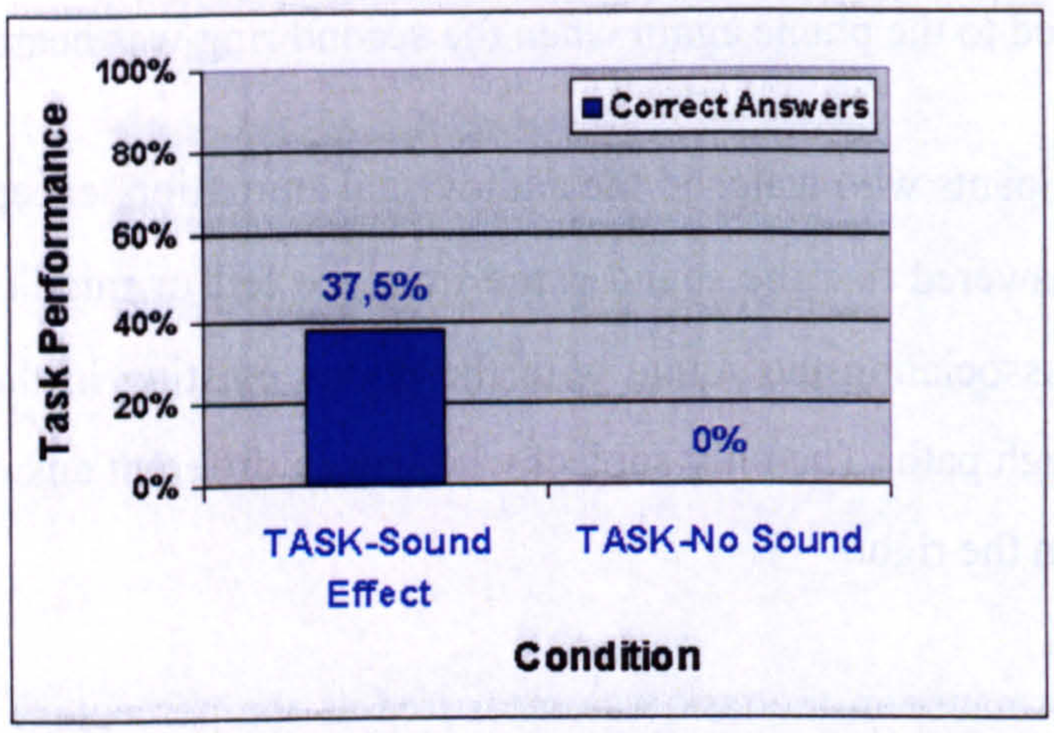


Figure 6.25: Eye tracking experiment - Performance (accuracy) in the search and memory task (percentage of correct counts of the target objects across participants) as a function of the auditory background.

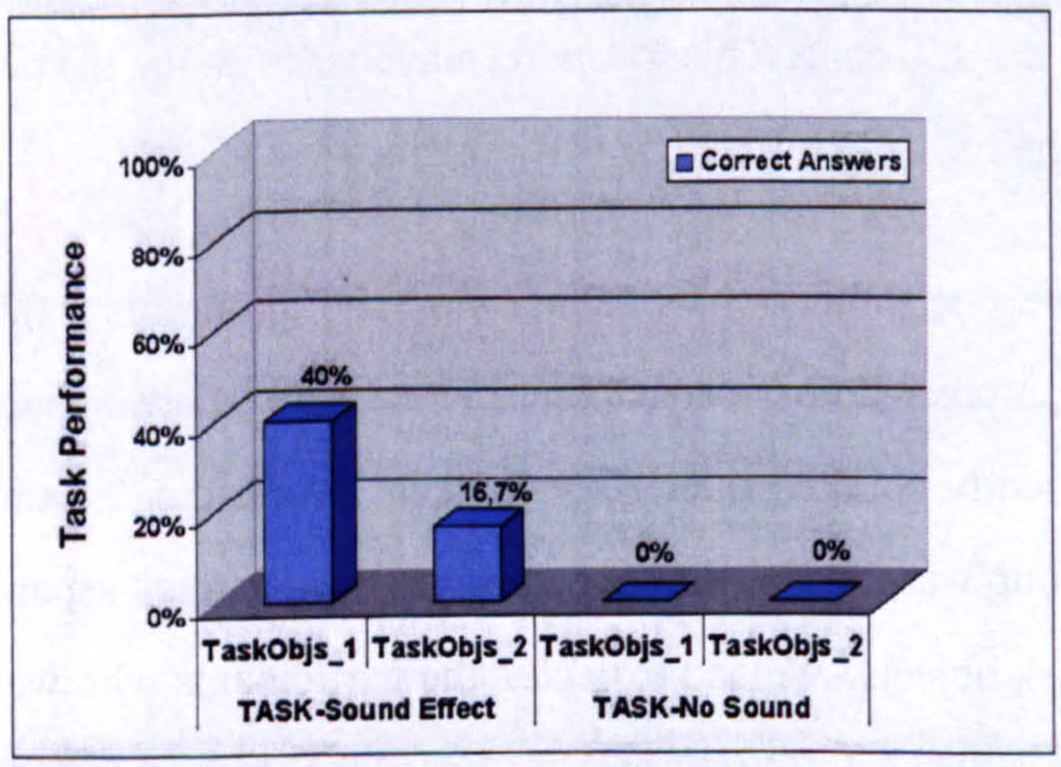


Figure 6.26: Eye tracking experiment - Performance (accuracy) in the search and memory task given separately for the two combinations of task objects. *TaskObjs_1* represents the "Orange cylinder-red ball" target objects, while *TaskObjs_2* are the "Red cylinder-blue ball" target object combination.

6.4 Summary

The results of the experiment we conducted on the influence of sound on the perceived rendering quality of rendered animations confirm that in the presence of audio stimuli, and more specifically sound effects with abrupt onsets, viewers fail to notice significant

quality degradations in the image regions that are not related to the sound. According to psychophysical findings discussed in previous chapters, this effect could be attributed to the fact that an auditory stimulus attracts a viewer's visual attention toward the origin of the sound, leaving the rest of the scene unattended.

An eye tracking experiment was designed and conducted in order to confirm that a sound-emitting object in a scene attracts a viewer's gaze and also to further investigate the strength of the phenomenon against task-relevant objects and competing highly-salient objects present in the scene.

The scan paths for the participants of our eye tracking experiment revealed that the gaze of the subjects who watched the audiovisual animation containing the sound of the ringing phone (lasting for approx. 3 seconds), was attracted toward the phone while the sound was audible. This was the case not only for the "Freeview" condition, where the sound-emitting object had to compete for attention with the highly salient objects present in the scene, but also for the "Task" condition, where the subjects were engaged in a demanding visual and memory task. These findings support our initial hypothesis about the captivation of a viewer's visual attention toward a sound-emitting object in his visual field.

To summarise, our experimental results demonstrate that sound effects may in fact be exploited to significantly speed up rendering, by reducing the quality of a large portion of the rendered scene without any noticeable difference to the viewer. These findings may have major implications for the developers of applications that make extensive use of sound effects, such as VEs and games, as the computational savings resulting from the exploitation of this observed auditory bias of visual attention can be dramatic.

Chapter 7

Conclusions and Future Work

The delivery of high-fidelity graphics for interactive graphics and multimedia applications, such as simulations, 3D games and VR environments, remains one of the major challenges for computer graphics practitioners, despite the huge progress in the related hardware and software during the past few years. For such applications, limitations of the human visual perception and cross-modal interactions on visual attention can be exploited in order to reduce the computational complexity and make more efficient use of the available resources, while trying to minimise the perceptibility of the resulting visual defects.

It is well known in the human perception community that many factors, including auditory stimuli, reduce a human's cognitive resources available to perform a visual task. In addition, spatial sound may attract attention to specific parts of a 3D scene, leaving the rest of the scene practically unattended. As a consequence, the exploitation of auditory 'distractors' in order to vary (degrade) the delivered frame rate and/or the quality of the visuals in animated scenarios could help towards more realistic graphics at interactive (or almost interactive) rates.

The main goal of our research efforts is to aid the development of high fidelity interactive scenarios, such as VR systems, by taking advantage of auditory-visual intersensory

interactions. More specifically, the main scope of this thesis was to investigate how degradations along different dimensions of rendering and delivery approximations (i.e. frame rate and rendering quality) when auditory stimuli are present can influence the perceived visual quality of a 3D scene.

This chapter discusses how the goals defined in introduction have been achieved, enumerates contributions of the thesis to the field and discusses ideas for future research on this topic.

7.1 Achievement of Goals

To gain a better understanding of the crossmodal interactions between the auditory and visual sensory modalities and identify whether such interactions could lead to a new generation of perceptually-adaptive graphics techniques, that would take into account not only the visual stimuli, but also the auditory background of a 3D scene, 292 subjects participated in five experiments. Temporal and 'spatial' visual display quality perceptions were investigated by manipulating the frame rate and the rendering quality (number of rays shot per pixel of the image), separately, and by considering different auditory backgrounds.

Our experimental studies verified that we can affect the viewer's perception of delivered frame rate with the use of auditory stimuli. More specifically, statistically significant results indicate that we can render less frames per second- without any cost in the user's perception of delivery rate- and thus pursue a higher image quality per frame. Another set of our results showed that the viewers do fixate to sources of sound effects in a scene-even when engaged in a demanding visual task- allowing us to render the corresponding pixels to high quality and significantly drop the quality for the rest of the scene, without any noticeable difference to the observer, and therefore we can save time from the calculation of the rendering solution for increasing the frame rate of the delivered graphics.

The main findings of our experimental studies, which were designed and conducted in order to pursue the goals of this thesis, are summarised here.

After completing all the theoretical groundwork, we designed and conducted a preliminary study, described in Chapter 4, on the effect of audio- and more specifically of musical stimuli- on temporal perception. The foundations for this study were the well demonstrated phenomena of auditory capture of vision in the temporal domain, such as the 'auditory driving' phenomenon presented in section 2.3.2, and also the reported effects of musical stimuli on human performance in tasks requiring cognitive activity and resources ("Arousal-Mood" hypothesis- discussed in section 3.4.1- according to which, music of high tempo and exciting impression might increase subjects' arousal and as a consequence impose capacity limitations in sensory information absorption and processing). This study investigated the influence of musical stimuli on the human perception of temporal rate and duration during the rendering and display of 3D animations, and in particular whether music can perceptually 'drive' the presentation rate of animated scenarios and consequently affect their temporal length. One of the most interesting results of the preliminary study revealed that relaxing music has the effect of decreasing the perceived scene velocity. Also, exciting music of fast tempo was found to create the impression that the scene temporal rate is higher. However, these results were not statistically significant to reach safe conclusions.

The preliminary study demonstrated the potential of our methodology, however, it also highlighted the shortcomings of its informal experimental design. We built on this design, addressed its flaws and developed a more formal experimental framework that was employed in the main experimental studies, presented in Chapters 5 and 6, which investigated the perception of both temporal and visual characteristics of a 3D graphics environment.

Chapter 5 presented in detail the experimental methodology employed and the relevant results of the first two of the main studies in this thesis, which examined the perceived

smoothness of rendered animations under the influence of music and sound effects. The cornerstone for these two studies was the well demonstrated perception finding that, in multisensory environments, auditory stimuli attract part of the observers' attention, 'occupy' a portion of their cognitive resources and therefore reduce the total available resources allocated to the visual information processing. In addition, we again took into account the Arousal-Mood hypothesis. Based on these findings we hypothesised that it would be more difficult for subjects to distinguish frame rate differences between audiovisual composites, accompanied by sound effects or music, than between silent animations.

The results of these experiments confirm that in the presence of audio stimuli, and in particular sound effects, viewers fail to notice variations in the motion smoothness between walkthrough animations displayed at different rates, which are apparent in the absence of sound. This may be attributed to the fact that the auditory stimuli attract part of viewer's attention to the sound and away from the visual anomalies, such as jerky motion, which result from low frame rates. In addition, habituation to an auditory stimulus was not found to decrease its 'distractive' ability. Furthermore, no association between the type of camera movement in the scene (translation or rotation) and the viewers' perception of the motion smoothness/jerkiness was found. It was also demonstrated that viewers who are not familiar with animated computer graphics, find it much harder to notice variations in the motion smoothness between two animations, compared to people with prior experience.

The main objective of Chapter 6 was to investigate the influence of sound on the perceived rendering quality of rendered animations ('spatial' dimension of our research). This work was based on the phenomenon of inattentional blindness, according to which there's no perception without attention, and demonstrated auditory-visual links in spatial attention allocation and visual orienting. Regarding the latter, the finding that was particularly relevant to our work revealed that, although visual attention can be voluntarily allocated according to a viewer's intentions or task, auditory events may attract spatial visual atten-

tion to the perceptual origin of a sound, regardless of the visual perceptual load (refer to sections 3.2.1 and 3.4 for details).

From the perception findings mentioned above, we hypothesised that an auditory 'distractor' in 3D graphics scenarios may attract the viewer's gaze and focal attention to the perceptual origin of the sound in a scene, allowing us to selectively render only the sound emitting object to high quality and keep the quality of the rest of the scene objects much lower, without any noticeable difference to the observer. To test our hypothesis an appropriate selective renderer was implemented and an experiment with 120 participants was conducted. The results of this experiment confirm that in the presence of audio stimuli, and more specifically sound effects with abrupt onsets, viewers fail to notice significant quality degradations in the image regions that are not related to the sound.

An eye tracking experiment was designed and conducted in order to confirm that a sound-emitting object in a scene attracts a viewer's gaze and also to further investigate the strength of the phenomenon against task-related objects and competing highly-salient objects present within the observer's visual field. The participants' scan paths (series of saccades and fixations) revealed that their gaze was, indeed, attracted toward the phone when the ringing sound was heard. This was the case not only for the "Freeview" condition, where the sound-emitting object had to compete for attention with the highly-salient objects present in the scene, but also for the "Task" condition, where the subjects were engaged in a demanding visual and memory task. This finding further supported our hypothesis that sound directs a viewer's visual attention toward the corresponding sound-emitting object.

7.2 Thesis Contributions

Our findings regarding the perceived motion smoothness of an animation, i.e. the perceived delivery rate, under the influence of auditory stimuli, may prove to be very sig-

nificant for applications where the delivered frame rate is one of the Quality of Service (QoS) parameters. These applications include, but are not limited to, desktop VR systems, 3D games and simulation applications, especially the multiuser networked ones, as the savings in the computational time and the transmission resources by exploiting this observed auditory bias of visual attention can be dramatic. We would not suggest the application of these findings to full-scale virtual environments without further detailed investigation, as variations in the frame rate can cause motion sickness. In addition, video compression algorithms, which have started to take into account perceptual issues for low bit-rate applications, could also benefit from our findings. By encoding frame rate control, the sudden frame skipping which results from existing techniques and degrades motion smoothness significantly, could be reduced.

In addition, our results from the experiments on the pattern of eye fixations and the perceived rendering quality under the influence of sound effects, demonstrate that sound may in fact be exploited to significantly speed up rendering, by reducing the quality of a large portion of the rendered scene without any noticeable difference to the viewer. These findings may have major implications for the developers of applications that make extensive use of sound effects, such as VR worlds and games, as the computational savings resulting from the exploitation of this observed auditory bias of visual attention can be significant.

Despite the increasing research interest in how visual stimuli can be manipulated to reduce computational complexity and/or speed up rendering by employing perceptual criteria, computer graphics researchers and developers have not as yet considered intersensory interactions when designing perceptually adaptive techniques for 3D graphics. One key reason for that is the fact that- to our knowledge- till now no research was conducted on how intersensory phenomena can be exploited in the field of 3D computer graphics. Our results are probably the first reported in this domain.

In conclusion, the compelling findings of this thesis suggest that there is great scope for further research into how crossmodal auditory-visual interactions on the perception of vi-

sual quality of animated imagery can be exploited in order to develop novel perceptually-adaptive techniques for the interactive rendering and delivery of high-fidelity 3D graphics.

7.3 Future Work

The animations employed in our experiments were pre-computed (the camera motion in the 3D scenes was predefined), in order to ensure control across conditions. Control is essential between conditions for results to be valid and this becomes harder as more interactivity is introduced. However, allowing for interactivity of the user with a computer graphics world where the camera motion is altered on the fly, would emphasise more the 3D aspects of the space and would match applied settings, such as VR worlds and games, better. An experiment needs to be designed and conducted to investigate the effect of sound on the perceived delivery rate and rendering quality in passive versus interactive settings.

The work presented in Chapter 6 will form part of a more complex selective rendering system that takes also saliency into consideration [273]. This system should be able to deliver frames at interactive rates. More specifically, a preliminary model of bottom-up visual attention for arbitrary animations which include sound emitting objects has been designed, see Figure 7.1. This model is currently restricted to a bottom-up guidance of attention toward conspicuous visual targets. It is suitable for modelling attention when there is no task at hand (e.g. spontaneous looking), or for interrupting task-level attention with potentially important events (in our case, sound effects attracting attention to the origin of the sounds). Our model is a variation of the visual attention model described by [90], its main difference being that it also takes into account the 'saliency' of objects in the scene which are emitting sounds. Sound 'contingencies' are integrated into the visual saliency-based model, by adding a "sound-emitting objects map" (SEOM) between the saliency map and the winner-take-all (WTA) mechanism. The SEOM acts as a point-wise

multiplicative filter, making the model more likely to attend to regions of the visual field that contain sources of sound. The product between the saliency and SEOM maps yields the "attention guidance map" that drives the WTA, see Figure 7.1.

The framework incorporates an object-based inhibition-of-return (IOR) mechanism with our saliency map; every object in the scene is provided with an uncertainty level, which is a measure of the completeness of a viewer's mental representation of the object. A high uncertainty level indicates that the object has not been attended to before or is a sound-emitting object, while a low uncertainty level signifies that the observer has a relatively complete representation of the object and will probably distribute his attention to new areas in the scene. The locations of objects with low uncertainty levels are inhibited in the saliency map in order to represent the reduced importance of parts of the scene that do not contain sources of sound and the viewer has already been familiarised with. This IOR strategy does not only increase the predictive power of the attentional model, but also decreases the time needed to compute the saliency map. For instances of dynamic scenes where constantly attending to the most salient location in the incoming video may be more appropriate or cases when we want the viewer to look repeatedly at interesting or informative parts of an animation, the inhibition-of-return mechanism may be alternatively turned off.

A significant future improvement to our model will be the addition of a top-down attention component as well [31, 217]. This could allow us to consider subtle factors such as task relevance when planning gaze motions. Certain objects in the environment have a heightened perceptual conspicuity due to their perceived relevance, or importance, to the current task at hand. This is true for objects of interest, as well as for objects that might be mistaken for relevant objects. Task-based attention could be achieved by increasing the importance of certain object types.

Finally, future work could also investigate the use of spatially located 3D sound, instead of merely exploiting the ventriloquist effect as we do now.

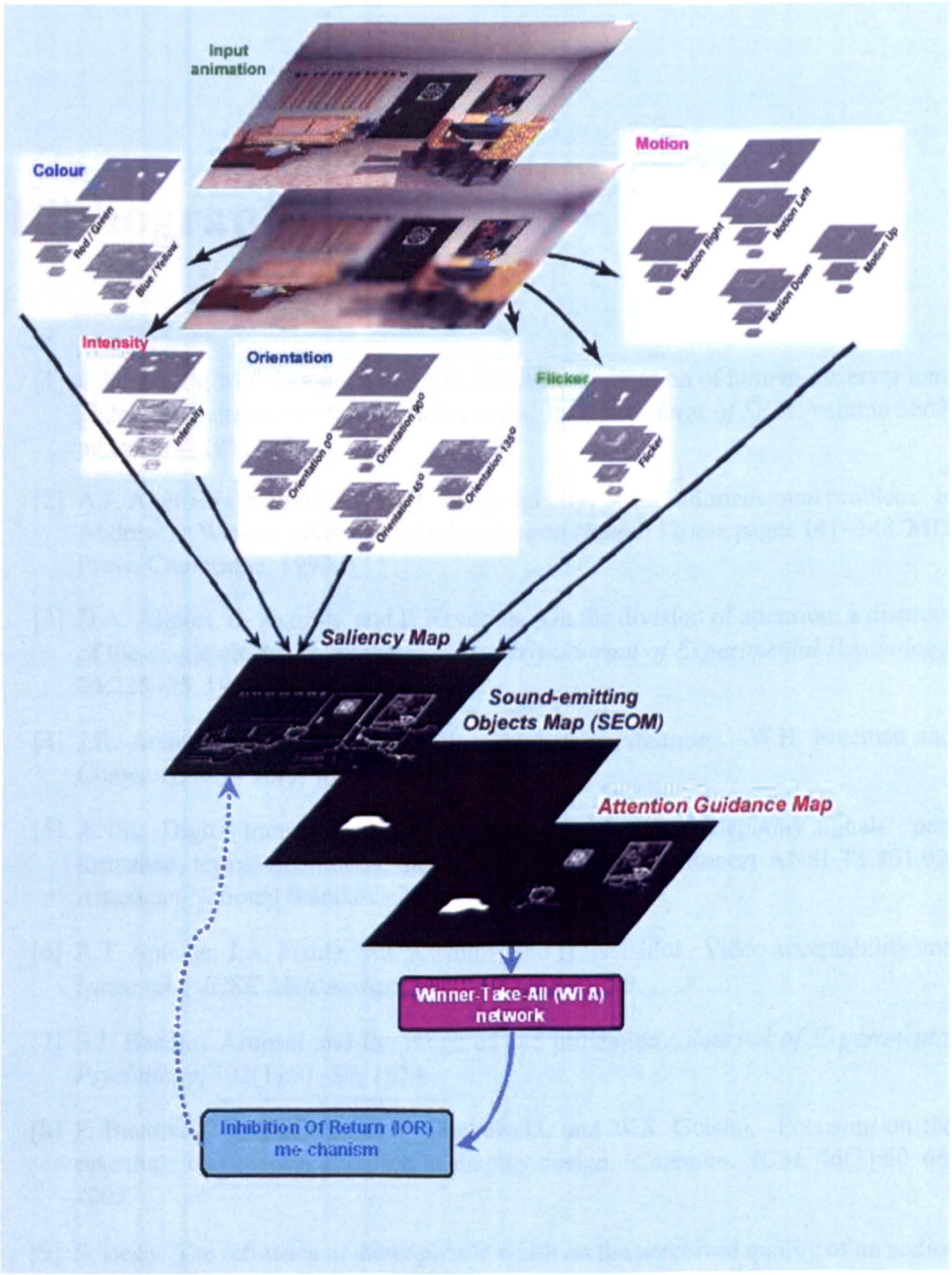


Figure 7.1: Our bottom-up visual attention model for arbitrary animations which include sound emitting objects.

Bibliography

[1] C.K. Abbey, M.P. Eckstein, and F.O. Bochud. Estimation of human-observer templates for 2 alternative forced choice tasks. In *Proceedings of SPIE*, volume 3663, pages 284–295, 1999.

[2] A.J. Ahumada Jr. and C.H. Null. Image quality: A multidimensional problem. In Andrew B. Watson, editor, *Digital Images and Human Vision*, pages 141–148. MIT Press, Cambridge, 1993.

[3] D.A. Allport, B. Antonis, and P. Reynolds. On the division of attention: a disproof of the single channel hypothesis. *Quarterly Journal of Experimental Psychology*, 24:225–35, 1972.

[4] J.R. Anderson. *Cognitive psychology and its implications*. W.H. Freeman and Company, New York, 1995.

[5] ANSI. Digital transport of video teleconferencing/video telephony signals - performance, terms, definitions, and examples. Technical Report ANSI T1.801.02, American National Standards Institute, 1995.

[6] R.T. Apteker, J.A. Fisher, V.S. Kisimov, and H. Neishlos. Video acceptability and frame rate. *IEEE Multimedia*, 2(3):32–40, Fall 1995.

[7] S.J. Bacon. Arousal and the range of cue utilization. *Journal of Experimental Psychology*, 102(1):81–87, 1974.

[8] P. Baudisch, D. DeCarlo, A.T. Duchowski, and W.S. Geisler. Focusing on the essential: considering attention in display design. *Commun. ACM*, 46(3):60–66, 2003.

[9] S. Bech. The influence of stereophonic width on the perceived quality of an audio-visual presentation using a multichannel sound system. Preprint No 4432. Presented at the 102nd Audio Engineering Society Convention, New York, March 22–25 1997.

[10] S. Bech, V. Hansen, and W. Woszczyk. Interaction between audio-visual factors in a home theater system: Experimental results. Preprint No 4096. Presented at the 99th Audio Engineering Society Convention, New York, October 6–9 1995.

[11] F. L. Bedford. Keeping perception accurate. *Trends in Cognitive Sciences*, 2:4–11, 1999.

[12] D.R. Begault. Perceptual effects of synthetic reverberation on three-dimensional audio systems. *Journal of Audio Engineering Society*, 40:895–904, 1992.

[13] D.R. Begault. Auditory and non-auditory factors that potentially in influence virtual acoustic imagery. In *Proc. AES 16th Int. Conf. on Spatial Sound Reproduction. Rovaniemi, Finland.*, pages 13–26, 1999.

[14] P. Bertelson. Ventriloquism: A case of cross-modal perceptual grouping. In G. Aschersleben, T. Bachmann, and J. Musseler, editors, *Cognitive contributions to the perception of spatial and temporal events*, pages 347–362. Amsterdam: Elsevier, 1999.

[15] P. Bertelson and G. Arhenleben. Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Reviews*, 5:482–489, 1998.

[16] P. Bertelson and M. Radeau. Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception & Psychophysics*, 29:578–587, 1981.

[17] P. Bertelson, J. Vroomen, B. De Gelder, and J. Driver. The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & Psychophysics*, 62:321–332, 2000.

[18] L.J. Biggs. Headphone-delivered three dimensional sound in npsnet. Master's thesis, Naval Postgraduate School, Monterey, California, September 1996. Available from http://citeseer.ist.psu.edu/biggs96headphonedelivered.html.

[19] M.R. Bolin and G.W. Meyer. A frequency based ray tracer. pages 409–418. ACM Press.

[20] M.R. Bolin and G.W. Meyer. A perceptually based adaptive sampling algorithm. pages 299–310. ACM Press.

[21] A. M. Bonnel and E.R. Hafter. Divided attention between simultaneous auditory and visual signals. *Perception & Psychophysics*, 60:179–190, 1998.

[22] P. A. Bourke, J. Duncan, and I. Nimmo-Smith. A general factor involved in dual-task performance decrement. *Quarterly Journal of Experimental Psychology*, 49A:525–545, 1996.

[23] G.E.P. Box, W.G. Hunter, and J.S. Hunter. *Statistics for Experimenters*. New York: Wiley, 1978.

[24] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organisation of Sound.* Cambridge, MA: The MIT Press, 1990.

[25] K. A. Briand. Feature integration and spatial attention: more evidence of a dissociation between endogenous and exogenous orienting. *Journal of Experimental Psychology : Human Perception and Performance*, 24:1243–56, 1998.

[26] D. Broadbent. *Perception and Communication.* London: Pergamon Press, 1958.

[27] P. Burt. Attention mechanisms for vision in a dynamic world. In *9th International Conference on Pattern Recognition,Beijing, China*, pages 977–987, 1988.

[28] K.O. Bushara, T. Hanakawa, I. Immisch, K. Toma, K. Kansaku, and M. Hallet. Neural correlates of cross-modal binding. *Nature Reviews Neuroscience*, 6:190–195, 2003.

[29] W. Buxton. Introduction to this special issue on nonspeech audio. *Human Computer Interaction*, 4(1):1–9, 1989.

[30] N.R. Carlson. *Psychology, The Science of Behavior.* Needham Heights, MA: Allyn and Bacon, 1993.

[31] K. Cater, À.G. Chalmers, and G. Ward. Detail to attention: Exploiting visual tasks for selective rendering. In *Proceedings Eurographics Symposium on Rendering, Leuven*, pages 270–280. ACM, June 2003.

[32] E. Catmull. A hidden-surface algorithm with anti-aliasing. pages 6–11. ACM Press.

[33] A.G. Chalmers, A. McNamara, S. Daly, K. Myszkowski, and T. Troscianko. *Image Quality Metrics.* ACM SIGGRAPH, July 2000.

[34] E.C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 23:915–919, 1953.

[35] M. Chion. *Audio-Vision: Sound on Screen.* Columbia University Press, New York, 1994.

[36] C.S. Choe, R.B. Welch, R.M. Gilford, and J.F. Juola. The "ventriloquist effect:" visual dominance or response bias? *Perception & Psychophysics*, 18:55–60, 1975.

[37] F. B. Colavita. Human sensory dominance. *Perception & Psychophysics*, 16:409–412, 1974.

[38] S. Daly. The visible differences predictor: An algorithm for the assessment of image fidelity. In Andrew B. Watson, editor, *Digital Images and Human Vision*, pages 179–206. MIT Press, Cambridge, 1993.

[39] H. de Ridder, F.J.J. Blommaert, and E.A. Fedorovskaya. Naturalness and image quality: Chroma and hue variation in color images of natural scenes. In *Proceedings of SPIE*, volume 2411, pages 51–61, San Jose, CA, 1995.

[40] K. Debattista. *Selective Rendering for High-Fidelity Graphics*. PhD thesis, University of Bristol, submitted 2006.

[41] K. Debattista, V. Sundstedt, L.P. dos Santos, and A.G. Chalmers. Selective component-based rendering. In *GRAPHITE, 3rd International Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia*, pages 13–22. ACM Press, November 2005.

[42] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18:193–222, 1995.

[43] H. Deubel and W. X. Schneider. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, 36:1827–1837, 1996.

[44] J.A. Deutch and D. Deutsch. Attention: some theoretical considerations. *Psychological Review*, 70:80–90, 1963.

[45] D. Deutsch. The psychology of hearing. *Sound and Video Contractor*, pages 34–44, September 1998.

[46] J. Driver. Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature*, 381:66–68, 1996.

[47] J. Driver. A selective review of selective attention research from the past century. *British Journal of Psychology*, 92:53–78, 2001.

[48] J. Driver and C. Spence. Crossmodal attention. *Current Opinion in Neurobiology*, 8:245–253, 1998.

[49] J. Driver and C. Spence. Crossmodal links in spatial attention. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 353:1319–1331, 1998.

[50] A. Duchowski and R. Vertegaal. Eye-based interaction in graphical systems: Theory and practice. In *SIGGRAPH 2000 Course Notes No. 5, New Orleans*, 2000.

[51] A. Dufour. Importance of attentional mechanisms in audiovisual links. *Experimental Brain Research*, 126:215–222, 1999.

[52] R. Dumont, F. Pellacini, and J.A. Ferwerda. Perceptually-driven decision theory for interactive realistic rendering. *ACM Transactions on Graphics*, 22(2):152–181, 2003.

[53] J.A. Easterbrook. The effect of emotion on cue utilization and the organization of behaviour. *Psychological Review*, 66:183–201, 1959.

[54] H.E. Egeth and L.C. Sager. On the locus of visual dominance. *Perception & Psychophysics*, 22(1):77–86, 1977.

[55] M. Eimer and E. Schröger. Erp effects of intermodal attention and cross-modal links in spatial attention. *Psychophysiology*, 35:313–327, 1998.

[56] H.C. Ellis and R.R. Hunt. *Fundamentals of Human Memory and Cognition.* Dubuque. Iowa: Wm. C. Brown, 1989.

[57] Enjoy the music.com Classic. Music definition: Rap music, 1996-2005. http://www.enjoythemusic.com/musicdefinition2.htm, Last accessed: 24-3-2006.

[58] C. W. Eridksen and J. E. Hoffman. The extent of processing of noise elements during selective encoding from visual displays. *Perception & Psychophysics*, 14:155–160, 1973.

[59] C.W. Eriksen and J.D. St. James. Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics*, 40(4):225–240, 1986.

[60] M.W. Eysenck. *Attention and arousal.* New York: Springer, 1982.

[61] R. Fendrich and P.M. Corballis. The temporal cross-capture of audition and vision. *Perception & Psychophysics*, 63:719–725, 2001.

[62] J.M. Findlay and Z. Kapoula. Scrutinization, spatial attention, and the spatial programming of saccadic eye movements. *Quarterly Journal of Experimental Psychology*, 45(4):633–47, November 1992.

[63] C.L. Folk, R. Remington, and J.C. Johnston. Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance*, 18:1030–1044, 1992.

[64] P. Fraisse. Rhythm and tempo. In Diana Deutch, editor, *The Psychology of Music*, page 174. New York: Springer-Verlag, 1982.

[65] S.P. Frysinger. Applied research in auditory data representation. In D. Farrell, editor, *Extracting meaning from complex data: processing, display, interaction (Proceedings of the SPIE/SPSE symposium on electronic imaging)*, pages 130–139, 1990.

[66] T. Funkhouser, J. Jot, and N. Tsingos. Sounds good to me! computational sound for graphics, virtual reality, and interactive systems. In *SIGGRAPH 2002 Course Notes, San Antonio, TX*, July 2002.

[67] A. Gabrielsson. Emotions in strong experiences with music. In P. N. Juslin and J. A. Sloboda, editors, *Music and emotion: Theory and research*, pages 431–449. New York: Oxford University Press, 2001.

[68] J.W. Gebhard and G.H. Mowbray. On discriminating the rate of visual flicker and auditory flutter. *American Journal of Psychology*, 72:521–528, 1959.

[69] J.J. Gibson. Adaptation, after-effect, and contrast in the perception of curved lines. *Journal of Experimental Psychology*, 16:1–31, 1943.

[70] S. Gibson and R.J. Hubbold. Perceptually-driven radiosity. *Computer Graphics Forum*, 16(2):129–141, 1997.

[71] S. Gibson and R.J. Hubbold. A perceptually-driven parallel algorithm for efficient radiosity simulation. *IEEE Transactions on Visualization and Computer Graphics*, 6(3):220–235, July-September 2000.

[72] J. Gray and A. Wedderburn. Grouping strategies with simultaneous stimuli. *Quarterly Journal of Experimental Psychology*, 12:180–184, 1960.

[73] D.P. Greenberg. A framework for realistic image synthesis. *Communications of the ACM*, 42(8):43–53, August 1999.

[74] R. Groner and M.T. Groner. Attention and eye movement control: an overview. *European Archives of Psychiatry and Neurological Sciences*, 239(1):9–16, 1989.

[75] J. Haber, K. Myszkowski, H. Yamauchi, and H.-P. Seidel. Perceptually guided corrective splatting. pages 142–152.

[76] S. Handel. *Timbre perception and auditory object identification, Hearing*. New York, NY: Academic Press, 1995.

[77] T. C. Handy, A. Kingstone, and G. R. Mangun. Spatial distribution of visual attention : Perceptual sensitivity and response latency. *Perception & Psychophysics*, 58(4):613–627, 1996.

[78] J.M. Henderson. Visual attention and eye movement control during reading and scene perception. In K. Rayner, editor, *Eye movements and visual cognition*, pages 260–283. New York: Springer-Verlag, 1992.

[79] J.M. Henderson and A. Hollingworth. Eye movements during scene viewing: An overview. In G. Underwood, editor, *Eye Guidance in Reading and Scene Perception*, pages 269–283. Oxford: Elsevier, 1998.

[80] J.M. Henderson and A. Hollingworth. The role of fixation position in detecting scene changes across saccades. *Psychological Science*, 10:438–443, 1999.

[81] J.M. Henderson and A.D. Macquistan. The spatial distribution of attention following an exogenous cue. *Perception & Psychophysics*, 53(2):221–230, 1993.

[82] J.E. Hoffman and B. Subramaniam. The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, 57:787–795, 1995.

[83] M.P. Hollier and R.M. Voelcker. Objective performance assessment: Video quality as an influence on audio perception. Preprint No 4590. Presented at the 103rd Audio Engineering Society Convention, New York, September 26–29 1997.

[84] M.P. Hollier and R.M. Voelcker. owards a multi-modal perceptual model. *BT Technology Journal*, 15(4):163–172, October 1997.

[85] Y. Horita, M. Katayama, T. Murai, and M. Miyahara. Objective picture quality scale for video coding. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, volume 3, pages 319–322, Lausanne, Switzerland, September 1996.

[86] E. Horvitz and J. Lengyel. Perception, attention, and resources: A decision-theoretic approach to graphics rendering. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 238–249, Providence, RI, August 1997. Morgan Kaufmann: San Francisco.

[87] G. Husain, W.F. Thompson, and E.G. Schellenberg. Effects of musical tempo and mode on arousal, mood, and spatial abilities. *Music Perception*, 20:149–169, 2002.

[88] D.E. Irwin. Information integration across saccadic eye movements. *Cognitive Psychology*, 23(3):420–456, July 1991.

[89] L. Itti. *Models of bottom-up and top-down visual attention.* PhD thesis, California Institute of Technology, 2000.

[90] L. Itti, N. Dhavale, and F. Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. In B. Bosacchi, D. B. Fogel, and J. C. Bezdek, editors, *Proc. SPIE 48th Annual International Symposium on Optical Science and Technology*, volume 5200, pages 64–78, Bellingham, WA, Aug 2003. SPIE Press.

[91] L. Itti and C. Koch. A comparison of feature combination strategies for saliency-based visual attention systems. In *Proceedings of SPIE, Human Vision and Electronic Imaging IV*, volume 3644, pages 373–382, San Jose, CA, January 1999.

[92] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, 2000.

[93] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[94] T. Iwaki, M. Tsukamoto, and M. Iwanaga. The effects of repeated listening to music on arousal level. *Journal of Music Perception and Cognition*, 4(1):1–9, 1998.

[95] W. James. *Principles of Psychology*. New York: Holt, 1890.

[96] J. Jonides. Toward a model of the mind's eye's movement. *Canadian Journal of Psychology*, 34:103–112, 1980.

[97] J. Jonides. Voluntary vs. automatic control over the mind's eye's movement. In J.B. Long and A.D. Baddeley, editors, *Attention and Performance IX*. Hillsdale, N.J.:Lawrence Erlbaum Associates, 1981.

[98] J. Jonides and S. Yantis. Uniqueness of abrupt visual onset in capturing attention. *Perception & Psychophysics*, 43:346–354, 1988.

[99] D. Kahneman. *Attention and Effort*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1973.

[100] B. Kapralos, M. Jenkin, and E. Milios. Auditory perception and spatial auditory systems. Technical Report Technical Report CS-2003-07, Dept. of Computer Science, York University, Dec. 2003. http://www.cs.yorku.ca/techreports/2003/CS-2003-07.html.

[101] R. Klein. Does oculomotor readiness mediate cognitive control of the visual attention. In R. Nickerson), editor, *Attention and performance III*. Hillsdale: Erlbaum, 1973.

[102] C. Koch and S. Ullman. Shifts in selective visual attention: toward the underlying neural circuitry. *Human Neurobiology*, 4(2):19–227, 1985.

[103] S.J. Korchin. Anxiety and cognition. In C. Scheerer (Ed.), editor, *Cognition: Theory, research, promise*. New York: Harper & Row, 1964.

[104] C.L. Krumhansl. An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology*, 51:336–352, 1997.

[105] D. Lamy and Y. Tsal. A salient distractor does not disrupt conjunction search. *Psychonomic Bulletin & Review*, 6:93–98, 1999.

[106] N. Lavie. Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology : Human Perception and Performance*, 21:451–468, 1995.

[107] N. Lavie and S. Cox. On the efficiency of attentional selection: Efficient visual search results in inefficient rejection of distraction. *Psychological Science*, 8:395–398, 1997.

[108] N. Lavie and J.W. de Fockert. The role of working memory in attentional capture. *Psychonomic Bulletin & Review*, 12(4):669–674, 2005.

[109] N. Lavie, A. Hirst, J.W. de Fockert, and E. Viding. Load theory of selective attention and cognitive control. *Journal of Experimental Psychology : General*, 133:339–354, 2004.

[110] N. Lavie and Y. Tsal. Perceptual load as a major determinant of the locus of selection in visual attention. *Perception & Psychophysics*, 56:183–197, 1994.

[111] P.J. Lindh and C.J. van den Lambrecht. Efficient spatio-temporal decomposition for perceptual processing of video sequences. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, volume 3, pages 331–334, Lausanne, Switzerland, September 1996.

[112] D. Lischinski, B. Smits, and D.P. Greenberg. Bounds and error estimates for radiosity. pages 67–74, July 1994.

[113] P. Longhurst. *Rapid Saliency Identification for Selectively Rendering High Fidelity Graphics*. PhD thesis, University of Bristol, December 2005.

[114] J.M. Loomis, R.L. Klatzky, and R.G. Golledge. Auditory distance perception in real, virtual, and mixed environments. In Y. Ohta and H. Tamura, editors, *Mixed reality: Merging real and virtual worlds*, pages 201–214. Tokyo: Ohmsha, 1999.

[115] Lord Rayleigh [Strutt, J.W.]. On our perception of sound direction. *Philosophical Magazine*, 13:214–232, 1907.

[116] L. Loschky and G.W. McConkie. User performance with gaze contingent displays. pages 97–104.

[117] L.C. Loschky and G.W. McConkie. Investigating spatial vision and dynamic attentional selection using a gaze-contingent multiresolutional display. *Journal of Experimental Psychology : Applied*, 8(2):99–117, 2002.

[118] L.C. Loschky, G.W. McConkie, J. Yang, and M.E. Miller. Perceptual effects of a gaze-contingent multi-resolution display based on a model of visual sensitivity. In *ARL Federated Laboratory 5th Annual Symposium - ADID Consortium Proceedings*, pages 53–58, 2001.

[119] L.C. Loschky, G.W. McConkie, J. Yang, and M.E. Miller. The limits of visual resolution in natural scene viewing. *Visual Cognition*, 12(6):1057–1092, 2005.

[120] D. Luebke and B. Hallen. Perceptually driven simplification for interactive rendering. In *Proceedings of 12th Eurographics Workshop on Rendering*, pages 221–223, 2001.

[121] F.X.J. Lukas and Z.L. Budrikis. Picture quality prediction based on a visual model. *IEEE Transactions on Communications*, 30(7):1679–1692, 1982.

[122] A. Mack and I. Rock. *Inattentional Blindness*. Cambridge, MA: MIT Press, 1998a.

[123] D.M. Mackie and L.T. (1989) Worth. Processing deficits and the mediation of positive affect in persuasion. *Journal of Personality and Social Psychology*, 57:27–40, 1989.

[124] A. Mania. *Fidelity Metrics for Virtual Environment Simulations based on Human Judgements of Spatial Memory Awareness States*. PhD thesis, Department of Computer Science, University of Bristol, August 2001.

[125] G. Marmitt and A.T. Duchowski. Modeling visual attention in vr: Measuring the accuracy of predicted scanpaths. In *EuroGraphics 2002 Proceedings, Saarbruecken, Germany*, pages 217–226. ACM, September 2 - 6 2002.

[126] D.W. Massaro. Speechreading: illusion or window into pattern recognition. *Trends in Cognitive Science*, 3:310–317, 1999.

[127] D.W. Massaro and D.S. Warner. Dividing attention between auditory and visual perception. *Perception & Psychophysics*, 21:569–574, 1977.

[128] G. Mastoropoulou, K. Debattista, A.G. Chalmers, and T. Troscianco. The influence of sound effects on the perceived smoothness of rendered animations. In *APGV 2005: Proceedings of the 2nd Symposium on Applied Perception in Graphics and Visualization*, pages 9–15. ACM SIGGRAPH, August 2005.

[129] G. Mastoropoulou, K. Debattista, A.G. Chalmers, and T. Troscianko. Auditory bias of visual attention for perceptually-guided selective rendering of animations. In *GRAPHITE 2005, sponsored by ACM SIGGRAPH, Dunedin, New Zealand*. ACM Press, December 2005.

[130] Georgia Mastoropoulou and A.G. Chalmers. The effect of music on the perception of display rate and duration of animated sequences: an experimental study. In *Theory and Practice of Computer Graphics 2004 (TPCG'04)*, pages 128–134, June 2004.

[131] T. Matsui and S. Hirahara. A new human vision system model for spatiotemporal image signals. In *Proceedings of SPIE, San Jose, CA*, volume 1453, pages 282–289, 1991.

[132] S. McAdams and E. Bigand. *Thinking in sound: The cognitive psychology of human audition*. Oxford: Claredon Press, 1993.

[133] P.A. McCormick. Orienting without awareness. *Journal of Experimental Psychology : Human Perception and Performance*, 23:168–180, 1997.

[134] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, December 1976.

[135] A. McNamara. *Comparing Real and Synthetic Scenes Using Human Judgements of Lightness*. PhD thesis, Department of Computer Science, University of Bristol, Bristol, UK, October 2000.

[136] A. McNamara, A.G. Chalmers, T. Troscianko, and I. Gilchrist. Comparing real and synthetic scenes using human judgements of lightness. In B. Peroche and H. Rushmeier, editors, *Proceedings of the Eurographics Workshop in Brno*, pages 207–219. Eurographics, June 2000.

[137] G.F. Meyer and M. Wuerger. Cross-modal integration of auditory and visual motion signals. *Neuroreport*, 12:2557–2560, 2001.

[138] D.P. Mitchell. Generating antialiased images at low sampling densities. *Computer Graphics (SIGGRAPH'87 Proceedings)*, 21(4):23–34, july 1987.

[139] S. Morein-Zamir, S. Soto-Faraco, and A. Kingstone. Auditory capture of vision: Examining temporal ventriloquism. *Cognitive Brain Research*, 17:154–163, 2003.

[140] A.K. Myers, B. Cotton, and H.A. Hilp. Matching the rate of tone bursts and light flashes as a function of flash surround luminance. *Perception & Psychophysics*, 30:33–38, 1981.

[141] K. Myszkowski. The visible differences predictor: Applications to global illumination problems. In G. Drettakis and N. Max, editors, *Rendering Techniques '98 (Proceedings of Eurographics Rendering Workshop '98)*, pages 233–236, New York, NY, 1998. Springer Wien.

[142] K. Myszkowski. Perception-based global illumination, rendering, and animation techniques. In Alan Chalmers, editor, *Proceedings of the 18th Spring Conference on Computer Graphics (SCCG 2002)*, pages 13–24, Budmerice, Slovakia, 2002. ACM Siggraph.

[143] K. Myszkowski, P. Rokita, and T. Tawara. Perceptually-Informed Accelerated Rendering of High Quality Walkthrough Sequences. In *Proc. of the 10th Eurographics Workshop on Rendering*, pages 5–18, 1999.

[144] K. Myszkowski, P. Rokita, and T. Tawara. Perception-based fast rendering and antialiasing of walkthrough sequences. *IEEE Transactions on Visualization and Computer Graphics*, 6(4):360–379, 2000.

[145] K. Myszkowski, T. Tawara, H. Akamine, and H.-P. Seidel. Perception-guided global illumination solution for animation rendering. In E. Fiume (Ed.), editor, *Proceedings of SIGGRAPH 2001*, Computer Graphics Proceedings, Annual Conference Series, pages 221–230. ACM Press / ACM SIGGRAPH, New York, 2001.

[146] K. Nakayama and M. Mackeben. Sustained and transient components of focal visual attention. *Vision Research*, 29:1631–1647, 1989.

[147] D. Navon and D. Gopher. On the economy of the human processing system. *Psychological Review*, 86(3):214–255, 1979.

[148] W.W. Nelson and G.R. Loftus. The functional visual field during picture viewing. *Journal of Experimental Psychology: Human Learning and Memory*, 6:391–399, 1980.

[149] W.R. Neuman. *Beyond HDTV: Exploring Subjective Responses to very High Definition Television.* MIT Media Library, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1990.

[150] W.R. Neuman, A. Crigler, and V.M. Bove. Television sound and viewer perceptions. In *Proceedings of the Audio Engineering Society 9th International Conference*, volume 1/2, pages 101–104, February 1991.

[151] E. Niebur and C. Koch. Computational architectures for attention. In *The Attentive Brain*, pages 164–186. MIT Press, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1998.

[152] E.L.-C. Niu. Gaze-based video compression using wavelets. Master's thesis, University of Illinois, Urbana-Champaign, Urbana, IL, 1995.

[153] D.A. Norman. *Memory and Attention, 2nd Ed.* Chichester: John Wiley, 1976.

[154] D.A. Norman and D.G. Bobrow. Visual dominance: an information-processing account of its origins and significance. *Cognitive Psychology*, 7:44–64, 1975.

[155] A.C. North and D.J. Hargreaves. Liking, arousal potential, and the emotions expressed by music. *Scandinavian Journal of Psychology*, 38:45–53, 1997.

[156] D. Noton and L. Stark. Eye movements and visual perception. *Scientific American*, 224(6):35–43, 1971.

[157] D. Noton and L. Stark. Scanpaths in the eye movements during pattern perception. *Science*, 171(968):308–311, 1971.

[158] R.E. Ornstein. *On the experience of time.* Penguin Books, Baltimore, 1969.

[159] C. O'Sullivan, J. Dingliana, G. Bradshaw, and A. McNamara. Eye-tracking for interactive computer graphics. In *Proceedings of the 11th European Conference on Eye Movements (ECEM 11) (Abstract of Talk presented at the Conference), Turku, Finland*, page S45, 2001.

[160] Carol O'Sullivan, Sarah Howlett, Yann Morvan, Rachel McDonnell, and Keith O'Conor. Star report on perceptually adaptive graphics. pages 141–164.

[161] P. Padmos and M.V. Milders. Quality criteria for simulator images: A literature review. *Human Factors*, 34(6):727–748, 1992.

[162] S.N. Pattanaik, J.E. Tumblin, H. Yee, and D.P. Greenberg. Time-dependent visual adaptation for realistic real-time image display. In E. Fiume, editor, *Proceedings of SIGGRAPH 2000*, Computer Graphics Proceedings, Annual Conference Series, pages 47–54. ACM Press / ACM SIGGRAPH, New York, 2000.

[163] S. Pefferkorn and J.-L. Blin. Perceptual quality metric of color quantization errors on still images. In *Proceedings of SPIE, San Jose, CA*, volume 3299, pages 210–220, 1998.

[164] D.R. Perrott, K. Saberi, B. Brown, and T.Z. Strybel. Auditory psychomotor coordination and visual search performance. *Perception & Psychophysics*, 48:214–226, 1990.

[165] L.R. Peterson and M.J. Peterson. Short-term memory retention of individual items. *Journal of Experimental Psychology*, 58:193–198, 1959.

[166] R. Pettersson. Attention an information design perspective!, 1999. (International Institute for Information Design (IIID), Vienna, Austria).

[167] M.I. Posner. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32:3–25, 1980.

[168] M.I. Posner, M.J. Nissen, and R.M. Klein. Visual dominance: an information-processing account of its origins and significance. *Psychological Review*, 83:151–171, 1976.

[169] M.I. Posner, C.R. Snyder, and B.J. Davidson. Attention and the detection of signals. *Journal of Experimental Psychology: General*, 109:160–174, 1980.

[170] J. Prikryl and W. Purgathofer. Perceptually-driven termination for stochastic radiosity. In *WSCG '99 (Seventh International Conference in Central Europe on Computer Graphics, Visualization and Interactive Digital Media)*, pages 418–425, Plzen-Borey, Czech Republic, 1999. University of West Bohemia.

[171] W. Prinzmetal, A. Zvinyatskovskiy, and L. Dilem. Involuntary attention have different consequences: The effect of perceptual difficult. Manuscript submitted for publication, 2005.

[172] Z.W. Pylyshyn. *Seeing and visualizing: Its not what you think.* Cambridge, MA: MIT Press, 2003.

[173] M. Radeau. Auditory-visual spatial interaction and modularity. *Current Psychology of Cognition*, 13:3–51, 1994.

[174] M. Radeau and P. Bertelson. Adaptation to auditory-visual discordance and ventriloquism in semi-realistic situations. *Perception & Psychophysics*, 22:137–146, 1977.

[175] M. Ramasubramanian, S.N. Pattanaik, and D.P. Greenberg. A perceptually based physical error metric for realistic image synthesis. In A. Rockwood, editor, *Proceedings of SIGGRAPH 1999*, Computer Graphics Proceedings, Annual Conference Series, pages 73–82. ACM Press / ACM SIGGRAPH, New York, 1999.

[176] G.H. Recanzone. Auditory influences on visual temporal rate perception. *Journal of Neurophysiology*, 89:1078–1093, 2003.

[177] R.W. Remington. Attention and saccadic eye movements. *Journal of Experimental Psychology : Human Perception and Performance*, 6(4):726–44, November 1980.

[178] R. Rensink. Seeing, sensing, and scrutinizing. *Vision Research*, 40:1469–1487, 2000b.

[179] B.H. Repp and A. Penel. Auditory dominance in temporal processing: New evidence from synchronization with simultaneous visual and auditory sequences. *Journal of Experimental Psychology : Human Perception and Performance*, 28:1085–1099, 2002.

[180] S. Rihs. The influence of audio on perceived picture quality and subjective audio-video delay tolerance. In *MOSAIC Handbook*, pages 183–187. O'Reilly and Associates, Inc., 1996.

[181] A.N. Rimmel, M.P. Hollier, and R.M. Voelcker. The influence of cross-modal interaction on audio-visual speech quality perception, 1998.

[182] I. Rock and C.S. Harris. Vision and touch. *Scientific American*, 216:96–104, 1967.

[183] H. Rushmeier, G. Ward, C. Piatko, P. Sanders, and B. Rust. Comparing real and synthetic images: Some ideas about metrics. In *Proc. of the 6th Eurographics Workshop on Rendering*, pages 82–91, 1995.

[184] M. Ruz and J. Lupianez. A review of attentional capture: On its automaticity and sensitivity to endogenous control. *Psicologica*, 23:283–309, 2002.

[185] T.A. Ryan. Interactions of the sensory systems in perception. *Psychological Bulletin*, 37:659–698, 1940.

[186] L.A. Schmidt and L.J. Trainor. Frontal brain electrical activity (eeg) distinguishes valence and intensity of musical emotions. *Cognition and Emotion*, 15:487–500, 2001.

[187] W. Schneider and R. M. Shiffrin. Controlled and automatic human information processing: 1. detection, search, and attention. *Psychological Review*, 84:1–66, 1977.

[188] B.J. Scholl. Objects and attention: The state of the art. *Cognition*, 80:1–46, 2001.

[189] R. Sekuler, A.B. Sekuler, and R. Lau. Sound alters visual motion perception. *Nature*, 385:308, 1997.

[190] L. Shams, Y. Kamitani, and S. Shimojo. What you see is what you hear. *Nature*, 408:788, December 2000.

[191] R. Shilling and B. Sinn-Cunningham. Virtual auditory displays. In K. Stanney, editor, *Handbook of Virtual Environments*, pages 65–92. Lawrence Erlbaum Associates, Mahwah, NJ. USA, 2002.

[192] S. Shimojo, C. Scheier, L. Nijhawan, R. amd Shams, Y. Kamitani, and K. Watanabe. Beyond perceptual modality: Auditory effects on visual perception. *Acoustical Science and Technology*, 22:61–67, 2001.

[193] S. Shimojo and L. Shams. Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion in Neurobiology*, 11(4):505–509, August 2001.

[194] T. Shipley. Auditory flutter-driving of visual flicker. *Science*, 145:1328–1330, 1964.

[195] D.J. Simons. Attentional capture and inattentional blindness. *Trends in Cognitive Sciences*, 4:147–155, 2000.

[196] D.J. Simons and C.F. Chabris. Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception*, 28(9):1059–1074, 1999.

[197] S. Singh and J. Hitchon. The intensifying effects of exciting television programs on the reception of subsequent commercials. *Psychology and Marketing*, 6:1–31, 1989.

[198] J.A. Sloboda and P.N. Juslin. Psychological perspectives on music and emotion. In P. N. Juslin and J. A. Sloboda, editors, *Music and emotion: Theory and research*, pages 71–104. New York: Oxford University Press, 2001.

[199] H. Song and C.C.J. Kuo. Rate control for low-bit-rate video via variable-encoding frame rates. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(4):512–521, 2001.

[200] S. Soto-Faraco and A. Kingstone. Multisensory integration of dynamic information. In G. Calvert, C. Spence, and B. Stein, editors, *Handbook of Multisensory Processes*, pages 49–67. MIT Press, 2004.

[201] C. Spence and J. Driver. Covert spatial orienting in audition: exogenous and endogenous mechanisms facilitate sound localization. *Journal of Experimental Psychology: Human Perception and Performance*, 20:555–574, 1994.

[202] C. Spence and J. Driver. Audovisual links in endogenous covert spatial attention. *Journal of Experimental Psychology : Human Perception and Performance*, 22:1005–1030, 1996.

[203] C. Spence and J. Driver. Audiovisual links in exogenous covert spatial attention. *Perception & Psychophysics*, 59:1–22, 1997.

[204] C. Spence and J. Driver. Attracting attention to the illusory location of a sound: reflexive crossmodal orienting and ventriloquism. *Neuroreport*, 11:2057–2061, 2000.

[205] C. Spence and J. Driver (Eds.). *Crossmodal space and crossmodal attention*. Oxford: Oxford University Press, 2004.

[206] C. Spence, M.E.R. Nicholls, N. Gillespie, and J. Driver. Cross-modal links in exogenous covert spatial orienting between touch, audition, and vision. *Perception & Psychophysics*, 60:544–557, 1998.

[207] C. Spence, J. Ranson, and J. Driver. Cross-modal selective attention: on the difficulty of ignoring sounds at the locus of visual attention. *Perception & Psychophysics*, 62:410–424, 2000.

[208] C. Spence, D.I. Shore, and R.M. Klein. Multisensory prior entry. *Journal of Experimental Psychology : General*, 130:799–832, 2001.

[209] B.E. Stein and M.A. Meredith. *The Merging of the Senses*. MIT Press, Cambridge, Massachusetts, 1993.

[210] R.L. Storms. *Auditory-Visual Cross-Modal Perception Phenomena*. PhD thesis, Naval Postgraduate School, Monterey, California, September 1998.

[211] D.L. Strayer, F.A. Drews, and W.A. Johnston. Cell phone- induced failures of visual attention during simulated driving. *Journal of Experimental Psychology : Applied*, 9:23–32, 2003.

[212] D.L. Strayer and W.A. Johnston. Driven to distraction: dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological Science*, 12:462–466, 2001.

[213] T. Strybel, C. Manligas, and D. Perrott. Minimum audible movement angle as a function of the azimuth and elevation of the source. *Human Factors*, 34(3):267–275, 1992.

[214] V. Sundstedt and A.G. Chalmers. Evaluation of perceptually-based selective rendering techniques using eye-movements analysis. In *Spring Conference on Computer Graphics*. ACM SIGGRAPH, April 2006.

[215] V. Sundstedt, A.G. Chalmers, K. Cater, and K. Debattista. Top-down visual attention for efficient rendering of task related scenes. In *VMV 2004- Vision, Modelling and Visualization*. Stanford, November 2004.

[216] V. Sundstedt, K. Debattista, and A.G. Chalmers. Selective rendering using task-importance maps (poster). In *APGV 2004- Symposium on Applied Perception in Graphics and Visualization*, pages 175–175. ACM SIGGRAPH, August 2004.

[217] V. Sundstedt, K. Debattista, P. Longhurst, A.G. Chalmers, and T. Troscianko. Visual attention for efficient high-fidelity graphics. In *Spring Conference on Computer Graphics (SCCG 2005)*, pages 162–168, May 2005.

[218] K.T. Tan, M. Ghanbari, and D. Pearson. An objective measurement tool for mpeg video quality. *Signal Processing*, 70(3):279–294, 1998.

[219] D.J. Tellinghuisen and E.J. Nowak. The inability to ignore auditory distractors as a function of visual task perceptual load. *Perception & Psychophysics*, 65(5):817–828, 2003.

[220] J.F. Thayer and R. Levenson. Effects of music on psychophysiological responses to a stressful film. *Psychomusicology*, 3:44–54, 1983.

[221] J. Theeuwes. Exogenous and endogenous control of attention: The effect of visual onsets and offsets. *Perception & Psychophysics*, 49:83–90, 1991.

[222] J. Theeuwes, P. Atchley, and A.F. Kramer. Attentional control within 3-d space. *Journal of Experimental Psychology : Human Perception and Performance*, 24:1476–1485, 1998.

[223] J. Theeuwes, A.F. Kramer, S. Hahn, and D.E. Irwin. Our eyes do not always go where we want them to go: Capture of the eyes by new objects. *Psychological Science*, 9:379–385, 1998.

[224] J. Tierney. Jung in motion, virtually and other computer fuzz. *New York Times*, page C1 and C9, September 16 1993.

[225] Tobii Technology AB. Tobii 50 series product description, 2004. http://www.tobii.com.

[226] P. Tole, F. Pellacini, B. Walter, and D.P. Greenberg. Interactive global illumination in dynamic scenes. In *Proceedings of SIGGRAPH*, pages 537–546, 2002.

[227] A. Treisman. Acontextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, 12:242–248, 1960.

[228] A. Treisman. Verbal cues, language and meaning in attention. *American Journal of Psychology*, 77:206–14, 1964.

[229] A. Treisman, A. Vieira, and A. Hayes. Automatic and preattentive processing. *American Journal of Psychology*, 105:341–362, 1992.

[230] J. Tumblin and H. Rushmeier. Tone reproduction for realistic computer generated images. *IEEE Computer Graphics and Applications*, 13(6):42–48, November 1993.

[231] M. Turatto, V. Mazza, and C. Umilta. Crossmodal object-based attention: auditory objects affect visual processing. *Cognition*, 96(2):B55–64, June 2005.

[232] C.J. van den Branden Lambrecht. Color moving pictures quality metric. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, volume 1, pages 885–888, Lausanne, Switzerland, 1996.

[233] C.J. van den Branden Lambrecht and O. Verscheure. Perceptual quality measure using a spatio-temporal model of the human visual system. In *Proceedings of SPIE*, volume 2668, pages 450–461, San Jose, CA, 1996.

[234] J. Vroomen, P. Bertelson, and B. De Gelder. A visual influence in the discrimination of auditory location. In *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP'98), Terrigal-Sydney*, pages 131–135, 1998.

[235] J. Vroomen, P. Bertelson, and B. De Gelder. The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & Psychophysics*, 63:651–659, 2001.

[236] J. Vroomen and B. De Gelder. Sound enhances visual perception: Cross-modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception and Performance*, 26:1583–1590, 2004.

[237] J. Vroomen and B. De Gelder. Temporal ventriloquism: Sound modulates the flash-lag effect. *Journal of Experimental Psychology: Human Perception and Performance*, 30(3):513–518, 2004.

[238] P.L Wachtel. Conceptions of broad and narrow attention. *Psychological Bulletin*, 68:417–429, 1967.

[239] J.J. Wakshlag, R.J. Reitz, and D. Zillmann. Selective exposure to and acquisition of information from educational television programs as a function of appeal and tempo of background music. *Journal of Educational Psychology*, 74(5):666–677, 1982.

[240] J.T. Walker and K.J. Scott. Auditory-visual conflicts in the perceived duration of lights, tones and gaps. *Journal of Experimental Psychology: Human Perception and Performance*, 7:1327–1339, 1981.

[241] B. Walter, G. Drettakis, and S. Parker. Interactive rendering using the render cache. In *Proc. of the 10th Eurographics Workshop on Rendering, Rendering Techniques '99*, pages 235–246, Granada, Spain, June 1999.

[242] G. Ward. The radiance lighting simulation and rendering system. In *ACM Transactions on Computer Graphics*, volume 28, pages 459–72. Computer Graphics Proceedings, Annual Conference Series, July 1994.

[243] L.M. Ward, J.J. McDonald, and D. Lin. On asymmetries in crossmodal spatial attention orienting. *Perception & Psychophysics*, 62:12581264, 2000.

[244] A. Watson and M.A. Sasse. Evaluating audio and video quality in low-cost multimedia conferencing systems. *Interacting with Computers*, 8:255–275, 1996.

[245] A. Watson and M.A. Sasse. Measuring perceived quality of speech and video in multimedia conferencing applications. In *Proceedings of the ACM Multimedia Conference*, pages 55–60, September 1998.

[246] A.B. Watson. Perceptual-components architecture for digital video. *Journal of the Optical Society of America A.*, 7(10):1943–1954, 1990.

[247] A.B. Watson. Toward a perceptual video quality metric. In *Proceedings of SPIE, Human Vision and Electronic Imaging III*, volume 3299, pages 139–147, San Jose, CA, 1998.

[248] B. Watson, A. Friedman, and A. McGaffey. Measuring and predicting visual fidelity. In E. Fiume, editor, *Proceedings of SIGGRAPH 2001*, Computer Graphics Proceedings, Annual Conference Series, pages 213 – 220. ACM Press / ACM SIGGRAPH, New York, 2001.

[249] A.A. Webster, C.T. Jones, M.H. Pinson, S.D. Voran, and S. Wolf. An objective video quality assessment system based on human perception. In *Proceedings of SPIE*, volume 1913, pages 15–26, San Jose, CA, 1993.

[250] R.B. Welch, L.D. Duttonhurt, and D.H. Warren. Contributions of audition and vision to temporal rate perception. *Perceptual Psychophysiology*, 39:294–300, 1986.

[251] R.B. Welch and D.H. Warren. Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, 88:638–667, 1980.

[252] R.B. Welch and D.H. Warren. Intersensory interactions. In L. Kaufman K. R. Boff and J. P. Thomas, editors, *Handbook of Perception and Human Performance*, chapter 25. New York: Wiley, 1986.

[253] M. Wertheimer. Untersuchung zur lehre von der gestalt ii. *Psychologische Forschung*, 4:301–350, 1923.

[254] S.J. Westen, R.L. Lagendijk, and J. Biemond. Spatio-temporal model of human vision for digital video compression. In *Proceedings of SPIE, Human Vision and Electronic Imaging II*, volume 3016, pages 260–268, San Jose, CA, 1997.

[255] C.D. Wickens. The structure of attentional resources. In R. Nickerson, editor, *Attention and performance VIII*. Hillsdale, NJ: Erlbaum, 1980.

[256] C.D. Wickens. Processing resources in attention. In R. Parasuraman and D. Davis, editors, *Varieties of attention*, pages 63–102. New York, NY: Academic Press, 1984.

[257] C.D. Wickens. *Engineering psychology and human performance, 2nd Ed.* New York, NY: Harper Collins, 1992.

[258] F.L. Wightman and R. Jenison. Auditory spatial layout. In W. Epstein and S. J. Rogers, editors, *Perception of space and motion, 2nd ed.*, page 365400. San Diego, CA, USA: Academic Press, 1995.

[259] S. Winkler. A perceptual distortion metric for digital color video. In *Proceedings of SPIE*, volume 3644, pages 175–184, San Jose, CA, 1999.

[260] S. Winkler and C. Faller. Audiovisual quality evaluation of low-bitrate video. In *SPIE/IS&T Human Vision and Electronic Imaging*, volume 5666, pages 139–148, San Jose, CA, JAN 16-20 2005.

[261] S. Winkler and C. Faller. Perceived Audiovisual Quality of Low-Bitrate Multimedia Content. *IEEE Trans. Multimedia*, 2005.

[262] C. Wood and N. Cowan. The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel? *Journal of Experimental Psychology: Learning Memory and Cognition*, 21:255–260, 1995.

[263] D.L. Woods. The psychological basis of selective attention: Implications of event-related potential studies. In J.W. Rohrbaugh, R. Parasuraman, and R Jr Johnson, editors, *Event-related-potentials: basic issues and applications*, pages 178–209. New York: Oxford University Press, 1990.

[264] L.T. Worth and D.M. Mackie. Cognitive mediation of positive affect in persuasion. *Social Cognition*, 5:76–94, 1987.

[265] W. Woszczyk, S. Bech, and V. Hansen. Interaction between audio-visual factors in a home theater system: Definition of subjective attributes. Preprint No 4133. Presented at the 99th Audio Engineering Society Convention, New York, October 6–9 1995.

[266] W. Wundt. *Grundzüge der Physiologischen Psychologies [Translation: Foundations of Physiological Psychology]*. Leipzig: Wilhelm Engelmann, 1893.

[267] S. Yantis. Attentional capture in vision. In *Converging operations in the study of selective visual attention.*

[268] S. Yantis and J.C. Johnston. On the locus of visual selection: evidence from focused attention tasks. *Journal of Experimental Psychology: Human Perception and Performance*, 16:135149, 1990.

[269] S. Yantis and J. Jonides. Abrupt visual onsets and selective attention: Evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 10:601–621, 1984.

[270] S. Yantis and J. Jonides. Abrupt onsets and selective attention: Voluntary versus automatic allocation. *Quarterly Journal of Experimental Psychology*, 28:429–440, 1990.

[271] A. L. Yarbus. *Eye Movements and Vision*. New York: Plenum Press, 1967.

[272] H. Yee. Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. Master's thesis, Program of Computer Graphics, Cornell University, 2000.

[273] H. Yee, S. Pattanaik, and D.P. Greenberg. Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. In *ACM Transactions on Computer Graphics*, volume 20, pages 39–65. M. Cohen, Ed., Computer Graphics Proceedings, Annual Conference Series, January 2001.

[274] S.N. Yendrikhovskij, F.J.J. Blommaert, and H. de Ridder. Perceptually optimal color reproduction. In *Proceedings of SPIE*, volume 3299, pages 274–281, San Jose, CA, 1998.

[275] E.J.S. Zagier. Perceptually-driven graphics. Technical Report TR97-017, ECCC, 1997.

Appendix A

Experiment on Temporal perception - Questionnaire

Age: … ….
Occupation: …

Which animated sequence had sound (according to the order of watching):
☐ the first ☐ the second

---

1) How focused were you on the watching of the animated sequences?
During the first sequence:
☐ Not at all ☐ A little bit ☐ Moderately ☐ Very much
During the second sequence:
☐ Not at all ☐ A little bit ☐ Moderately ☐ Very much

---

2) In which of the two sequences the camera was moving faster in the 3D scene;
☐ in the first ☐ in the second ☐ didn't notice any difference

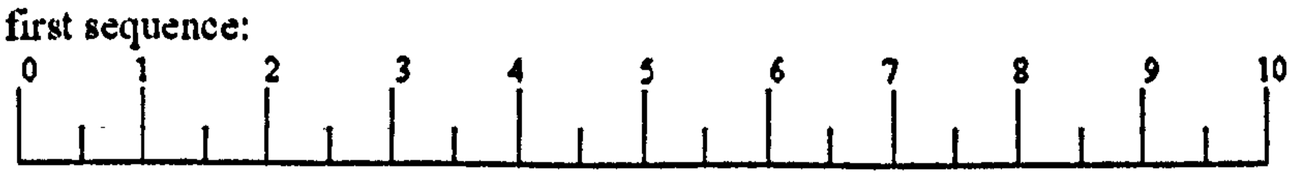
---

3) Which of the two sequences lasted longer;
☐ the first ☐ the second ☐ didn't notice any difference

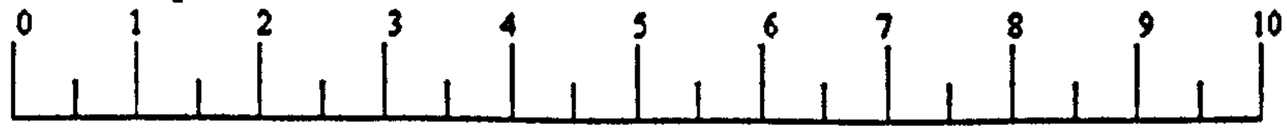
---

4) Which do you think was the duration of each sequence (in minutes)? Draw a line so that its length corresponds to the perceived duration. (The mid-interval vertical markers correspond to 30 seconds).

The first sequence:



The second sequence:



---

5) Was the musical background of the audio-visual sequence familiar to you?
☐ Not at all ☐ A little bit ☐ Moderately ☐ Very much

---

6) Did you find the music relaxing?

☐ Not at all ☐ A little bit ☐ Moderately ☐ Very much

---

7) Did you find the music exciting?

☐ Not at all ☐ A little bit ☐ Moderately ☐ Very much

---

8) Did you like the music?

☐ Not at all ☐ A little bit ☐ Moderately ☐ Very much

---

9) How relaxed comfortable did you feel during the watching of the sequences?
During the first sequence:
☐ Not at all ☐ A little bit ☐ Moderately ☐ Very much
During the second sequence:
☐ Not at all ☐ A little bit ☐ Moderately ☐ Very much

Appendix B

Eye-tracking Experiment - Resulting Scan Paths

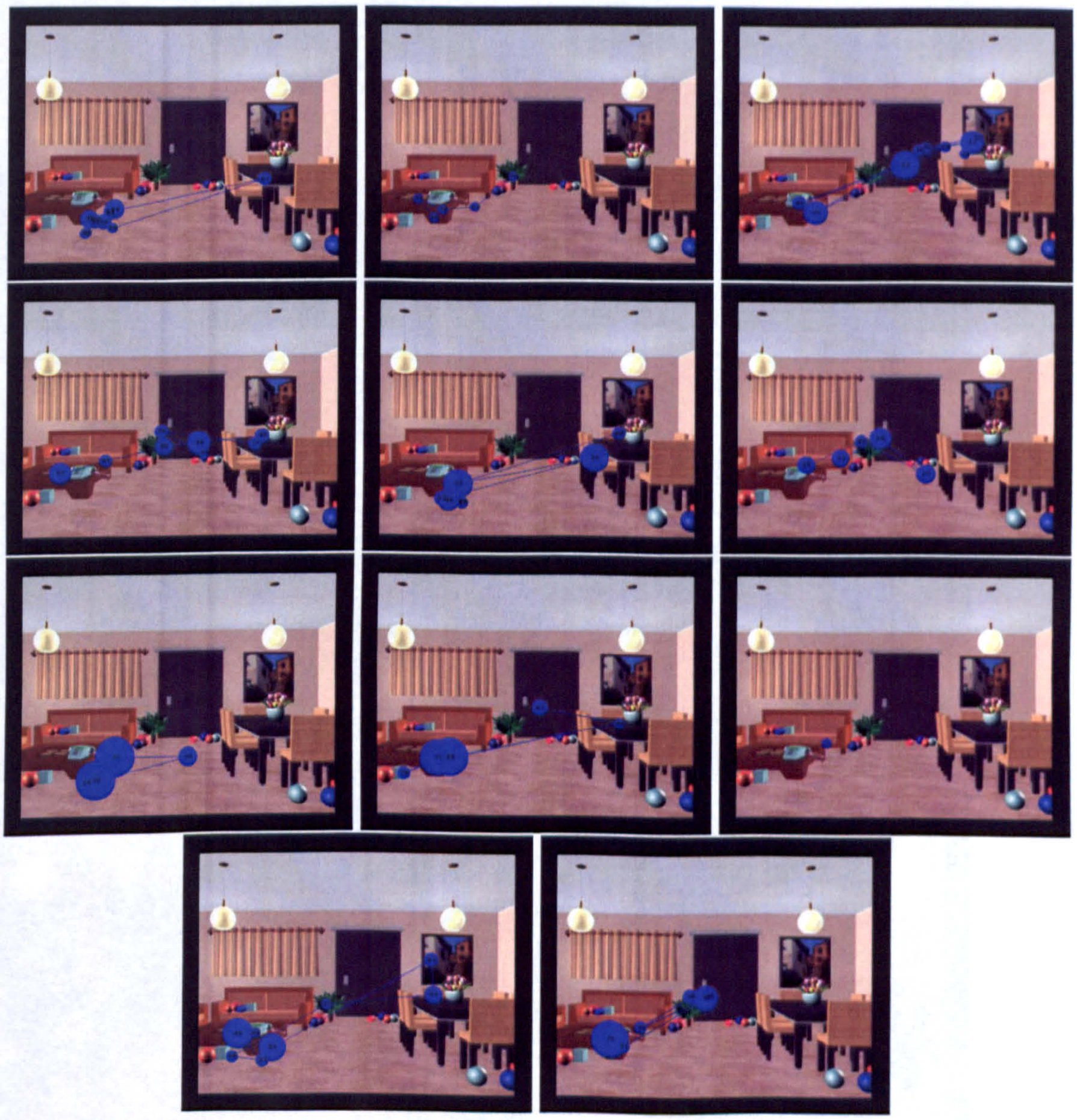


Figure B.1: Scan paths of the participants in the "Freeview-Sound Effect" group, corresponding to the 3-second period that the sound effect of the ringing phone was audible.

Figure B.2: Scan paths of the participants in the "Freeview-No Sound Effect" group (corresponding to the 3-second period that the sound effect was audible in the audiovisual animation).

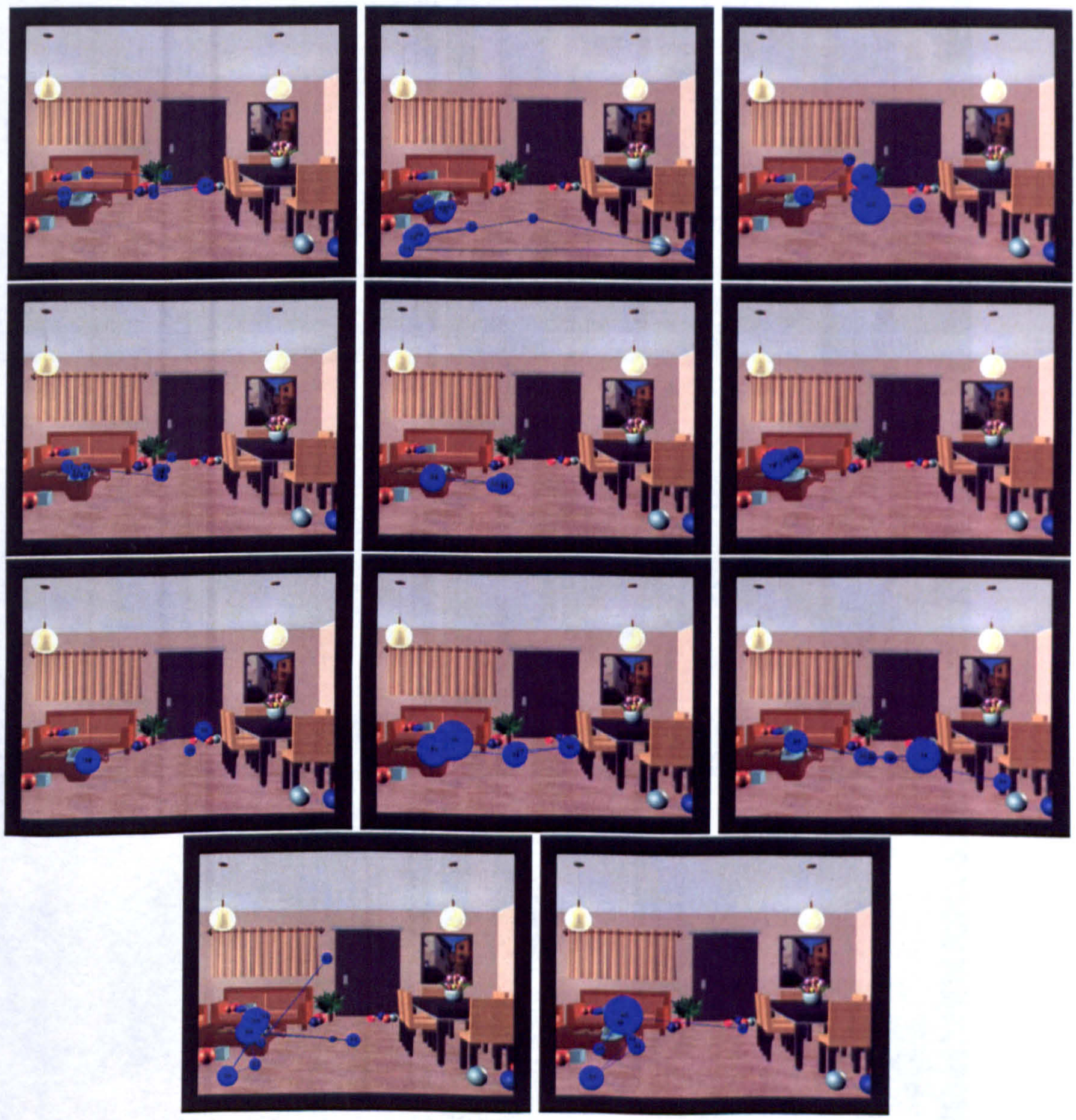


Figure B.3: Scan paths of the participants in the "Task-Sound Effect" group, corresponding to the 3-second period that the sound effect of the ringing phone was audible.

Figure B.4: Scan paths of the participants in the "Task-No Sound Effect" group (corresponding to the 3-second period that the sound effect was audible in the audiovisual animation).