TESIS DE LA UNIVERSIDAD
DE ZARAGOZA

2023 102

Sandra Malpica Mallo

Visual and multimodal perception in immersive environments

Director/es

Gutiérrez Pérez, Diego
Masiá Corcoy, Belén

ISSN 2254-7606

Prensas de la Universidad
Universidad Zaragoza

Servicio de Publicaciones

ISSN 2254-7606

Tesis Doctoral

VISUAL AND MULTIMODAL PERCEPTION IN IMMERSIVE ENVIRONMENTS

Autor

Sandra Malpica Mallo

Director/es

Gutiérrez Pérez, Diego
Masiá Corcoy, Belén

UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado

Programa de Doctorado en Ingeniería de Sistemas e Informática

2023

VISUAL AND MULTIMODAL PERCEPTION IN IMMERSIVE ENVIRONMENTS

SANDRA MALPICA

SUPERVISED BY: BELEN MASIA AND DIEGO GUTIERREZ

TESIS DOCTORAL - INGENIERÍA INFORMÁTICA

GRAPHICS AND IMAGING LAB
DEPARTAMENTO DE INFORMÁTICA E INGENIERÍA DE SISTEMAS
UNIVERSIDAD DE ZARAGOZA

DECEMBER 2022

Sandra Malpica
Visual and Multimodal Perception in Immersive Environments

The most important step a person can take. It's not the first one, is it?
It's the next one. Always the next step.

Brandon Sanderson

Abstract

Through this thesis we use virtual reality (VR) as a tool to better understand human visual perception and attentional behavior. We leverage the intrinsic properties provided by VR in order to build user studies tailored to a set of different topics: VR provides increased control over sensory information when compared to traditional media, as well as more natural interactions with the environment and an increased sense of realism. These qualities, together with the feeling of presence and immersion, increase the ecological validity of user studies made in VR. Furthermore, it allows us researchers to explore closer to real-world scenarios in a safe and reproducible way. By increasing the available knowledge about visual perception we aim to provide visual computing researchers with more tools to overcome current limitations in the field, either hardware- or software-caused. Understanding human visual perception and attentional behavior is a challenging task: measuring such high-level cognitive processes is often not feasible, more so without medical-grade devices (which are commonly invasive for the user). For this reason, we settle on measuring observable data, both qualitative and quantitative. This data is further processed to obtain information about human behavior and create high-level guidelines or models when possible.

We present the contributions of this thesis around two topics: visual perception of realistic stimuli and multimodal perception in immersive environments. The first one is devoted to visual appearance and has two separate contributions. First, we have created a learning-based appearance similarity metric by means of large-scale crowdsourced user studies and a deep learning model which correlates with human perception. Additionally, we study how low-level, asemantic visual features can be used to alter time perception in virtual reality, manifesting the interplay between visual and temporal perception at interval timing (several seconds to several minutes) intervals. Regarding the second topic, multimodal perception, we have first compiled an in-depth study of the state of the art of the use of different sensory modalities (visual, auditory, haptic, etc.) in immersive environments. Additionally, we have analyzed a crossmodal suppressive effect in virtual reality, where auditory cues can significantly degrade visual performance. Finally, we have shown how temporal synchronization is key to correctly perceive multimodal events and enhance their realism, even when visual quality is degraded.

Ultimately, this thesis aims to increase the understanding of human behavior in immersive environments. This knowledge can not only benefit cognitive science researchers, but also computer graphics researchers, especially those in the field of VR, who will be able to use our findings to create better user experiences.

Resumen

En esta tesis utilizamos la realidad virtual (VR) como herramienta para comprender mejor la percepción visual y el comportamiento atencional del ser humano. Aprovechamos las propiedades intrínsecas que proporciona la VR para construir estudios de usuarios adaptados a un conjunto de temas diversos. La realidad virtual proporciona un mayor control sobre la información sensorial en comparación con los medios tradicionales, así como interacciones más naturales con el entorno y una sensación de realismo mayor. Estas cualidades, junto con la sensación de presencia y la inmersión, aumentan la validez ecológica de los estudios de usuarios realizados en VR. Esto permite a los investigadores explorar escenarios más cercanos al mundo real de forma segura y reproducible. Al aumentar los conocimientos disponibles sobre la percepción visual, pretendemos proporcionar a los investigadores en informática gráfica más herramientas para superar las limitaciones actuales del campo, ya sean causadas por el hardware o el software. Comprender la percepción visual y el comportamiento atencional del ser humano es una tarea difícil: medir directamente estos procesos cognitivos de alto nivel no suele ser factible, más aún sin dispositivos de grado médico (que suelen ser invasivos para el usuario). Por ello, debemos medir datos relacionados observables, tanto cualitativos como cuantitativos. Estos datos se procesan posteriormente para obtener información sobre el comportamiento humano y crear pautas o modelos de alto nivel siempre que sea posible.

Presentamos las aportaciones de esta tesis en torno a dos temas: la percepción visual de estímulos realistas y la percepción multimodal en entornos inmersivos. El primer tema está dedicado a la apariencia visual y tiene dos contribuciones distintas. En primer lugar, hemos creado una métrica de similitud de apariencia basada en el aprendizaje por medio de estudios de usuarios a gran escala y un modelo de *deep learning* que correla con la percepción humana. Además, estudiamos cómo las características visuales asemánticas de bajo nivel pueden utilizarse para alterar la percepción del tiempo en realidad virtual, manifestando la interacción entre la percepción visual y temporal en intervalos temporales de hasta tres minutos. En cuanto al segundo tema, la percepción multimodal, primero hemos recopilado un estudio en profundidad del estado del arte del uso de diferentes modalidades sensoriales en entornos inmersivos. Además, presentamos un efecto de supresión intermodal en realidad virtual, en el que diferentes señales auditivas degradan significativamente la percepción visual. Por último, mostramos cómo la sincronización temporal es clave para percibir correctamente los eventos multimodales y mejorar la percepción de ciertas propiedades incluso cuando la calidad visual disminuye.

En definitiva, esta tesis profundiza en la comprensión de la percepción humana en entornos inmersivos. Este conocimiento puede beneficiar no solo a los investigadores de las ciencias cognitivas, sino también a los investigadores del campo de la informática gráfica, especialmente a los que tratan con realidad virtual, que podrán utilizar nuestros hallazgos para crear mejores experiencias de usuario.

Acknowledgements

Luckily for me, there is a lot of people to thank here. To everyone who has supported me through the last four years, to everyone who has made my day better, to those who have made my worries lighter or helped me focus in a brighter moment through the dark days: Thank you. I don't know what the future holds, but I know I will get there if it's with you.

First and foremost, thank you *Diego* for giving me this amazing opportunity. I was planning on looking for a summer job, but the alternative you offered has resulted in this thesis. You've guided me to see the big picture, to become more self-confident. I hope I will continue learning, while working with you, about all those things that books don't teach. *Belen*, thank you for always being there. For teaching me how to be more organized and direct, for helping me see the problems before they are here and plan in advance. Your perseverance, your attitude towards problems and your dedication are inspiring. Thank you both for being my supervisors these years, and thank you for going further than your role required for me. I'm certain I wouldn't have been able to finish this thesis if not for your understanding and your guidance. Thank you *Ana* for guiding me through my first steps, and for being such an amazing mentor. It's great having you here. *Dani*, the days are much more brighter with your optimism around. It's great having you through the good and bad times in the lab. Thank you *Mercedes* for making our lives easier. To the rest of the *Graphics and Imaging Lab* (past and present), thank you for everything: for the fruitful discussions, for the trips to the cafeteria and for making the lab such a nice place to work in. To the students I supervised: I hope you learned a lot and enjoyed the experience as much as I did.

To my supervisors at Adobe: *Qi* and *Zoya*. Thank you *Qi* for the amazing opportunity, working with you in the office was great. Thanks for discussing all of those ideas with me and for always striving to reach a step further. *Zoya*, thank you so much for your insights and your guidance, and for giving me a new, refreshing perspective to integrate in my work. I learned a lot from you, both in and out of research. I hope that we can keep collaborating in the future. To my supervisor at Facebook: *Alex*, thank you for your patience, for supporting me when I needed it the most and for teaching me how to pivot a project to get the most out of my abilities. To all of my collaborators and colleagues: thank you for making our works better than they would have been without you.

To my friends and family, thank you: *Garban* for being there through the good and bad of life. *Burgo* for your smoldering humor. *Cascabel* for being yourself. *Bambam* for your energy. *Lau* for your craziness. *Pendiente* for your imagination and the conversations we share. *Jueves* for being thoughtful. *Droi* for always supporting me and caring about me (this is the last one I have to do, don't worry!). *Kins* for your kindness. *Oráculo* for bearing with us and always joining us. *Primi* for creating worlds where we can forget everything bad. *Iván* for your different perspective in life. *Esther* for coming to see me with your beautiful girls after all these years. *Fani*, for keeping in touch and making me a part of your life even when we are far away. To *María*, for being the sister I've always wanted in life. To *Antonia* and *Angel*, for their support and love through these last hard years. To everyone who shares my life with me even when we don't see or talk to each other so often. To the ones who are not there anymore, because somehow you contributed to me being here today.

To my mother, *Loli*, for caring about me, raising me, and loving me unconditionally. Thank you for making the hardest of efforts so that I had a good life. Thank you for teaching me how to be a kind person and for being understanding through our differences.

To my father, *Francisco*, for being my safe place and never judging me. Thank you for giving me freedom and supporting me, for teaching me so many things only experience can give. I hope we can keep sharing our time for a lot of years to come.

And finally, thank you to my husband, *David*. I couldn't have made this without you, and I just hope I can repay all the effort you made into starting our family from now on. For the last 12 years of our lives. For all the holidays to come exploring in our van, the time spent with our friends, the relaxing days at home and the nights playing videogames and watching TV. To *Alfa* and *Delta* for being the goodest of boys and to *Baco* for being the softest of cats. To *Beta* for keeping me company almost to the end. We will miss you.

Contents

I Introduction & Overview

II Visual perception

III Multimodal Perception in immersive environments

IV Conclusion

List of Figures

List of Tables

Part I

INTRODUCTION & OVERVIEW

1
Introduction

Human senses are a key component of the interface between the world and the concepts we create in our brains. They are the only means to gather information about the reality we live in. From sensory data, the human brain has evolved to unconsciously process and discard information dynamically as needed in order to keep track of a stable, coherent version of our surroundings. This mental scheme, together with our previous experiences, is necessary for high-order cognitive processes, including those in charge of decision-making and interaction with the environment. Regarding extrapersonal space, humans rely mainly on sight [395] to retrieve information. However, additional sensory modalities are needed in order to integrate a complete notion of the world and ourselves. Furthermore, we often receive information of different modalities from a single object or event (for example, moving objects usually emit sound). We define multimodality as the neural processing of several sensory modalities that are integrated or segregated into different groups of coincident signals based on their spatiotemporal and structural coherence. In the domain of multimodal perception, we define crossmodality as the interactions that arise between several modalities. Both multimodal and crossmodal events shape human visual perception and attentional behavior, having a significant influence over our relationship with the world. Since every event and object (whether occurring naturally or designed by a human as part of a synthetic experience) will be eventually filtered through human perception, it is important to take into account how we integrate sensory information in our cognitive processes.

This thesis lies at the intersection of computer graphics and applied perception, drawing from both fields to better understand human visual perception and attentional behavior in immersive environments and virtual reality (VR). Cognitive sciences have studied visual perception for decades through traditional displays and simple, two dimensional visual cues in controlled laboratory conditions. However, our understanding of the real world is modulated not only by the extrapersonal information we receive, but also by internal cues like proprioceptive, vestibular or body motion information. These internal aspects are rarely present in studies with traditional media, especially since it is common to ask users not to move while performing visual perception experiments. This is were VR presents an advantage over traditional media. First, it provides increased control over sensory information when compared to traditional media. Instead of watching a conventional display in a laboratory room, users can see only the relevant visual information needed for the experiment. Second, VR provides closer to real-life interactions. In virtual reality we can move with our bodies, we can see the part of the environment behind us by moving our head, we can touch virtual objects and we can hear where a sound is coming from in real life. Third, the feeling of presence (the feeling of actually being in the virtual environment, including feeling connected to the virtual world and being able to interact with it) and immersion (the perception of being physically present in a non-physical world) in VR are key components that make users behave realistically in the virtual environment, responding as they would in the real world [354].

These qualities increase the ecological validity of user studies made in VR and the plausibility of their behavior. Furthermore, it allows researchers to explore closer to real-world scenarios in a safe and reproducible way. In return, our findings also directly benefit VR researchers and practitioners, who can improve the quality of VR applications through an increased understanding of visual and attentional behavior in immersive environments. At the same time, visual computing (which includes the fields that deal with images like computer graphics and computer vision) can also benefit from a greater understanding of human visual perception and attentional behavior. Visual computing has advanced exponentially through the last decades to the point where we can

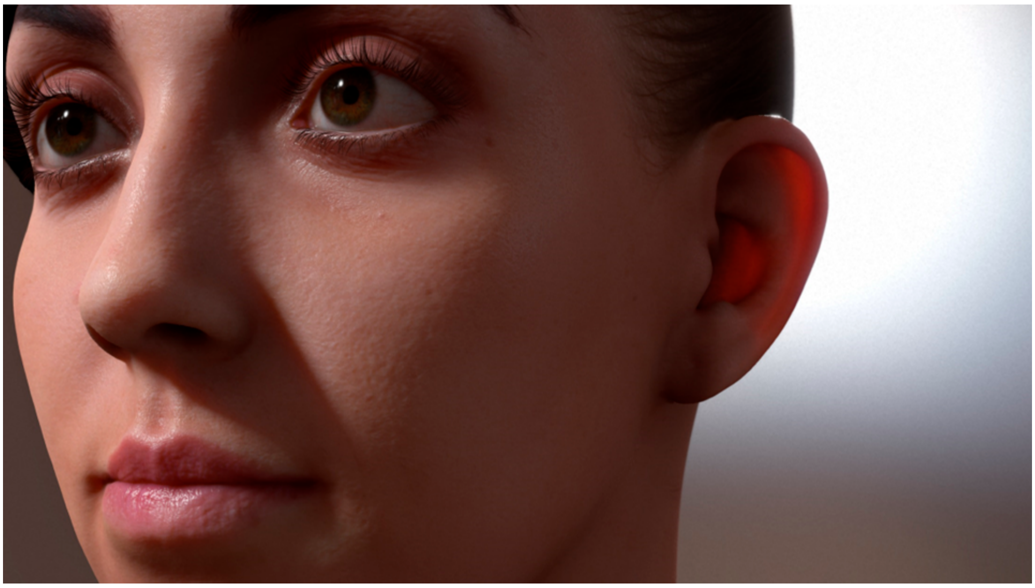


Figure 1.1: A photorealistic, real-time, physically-based rendering of Lauren, a computer generated test character by Jorge Jimenez in 2013. Note the detail and realism of the skin translucency, skin surface and appearance of the eyes.

generate highly realistic images (see Figure 1.1), but there is still room for improvement. Current challenges and problems, whether caused by a hardware or a software constraint, can leverage the particularities of human perception. For example, foveated rendering techniques achieve an increased perceived visual quality with less computational cost by tracking where the user is looking in real time and reducing image quality in the periphereral vision area. Another example, some locomotion techniques in VR like redirected walking can create a spatial distorsion in the virtual world because we are aware of the thresholds where such changes are imperceptible to humans. All in all, visual computing constantly leverages the existing knowledge about visual perception to overcome all kind of technical limitations.

Throughout this thesis we face a set of different topics, but approach them all with a common methodology. We measure user data (often behavioral, such as performance metrics, debriefing sessions, subjective preferences, eye tracking data, etc.) both qualitative and quantitative. We then extract meaningful information about human visual perception and attentional behavior and create high-level guidelines or models when possible. An increased understanding of human behavior in immersive environments will not only benefit the field of cognitive sciences, but also that of computer graphics, whose researchers and practitioners will be able to create better user experiences. In the following, we present the contributions of this thesis around two topics: visual perception of realistic stimuli and multimodal perception in immersive environments.

1.1 VISUAL PERCEPTION OF COMPUTER-GENERATED REALISTIC STIMULI

We first focus on the visual perception of realistic stimuli through Chapter 2 and Chapter 3. Although they both share a common user-centered approach, they address two different problems. Chapter 2 studies and models the notion of perceived visual similarity in the context of material appearance while Chapter 3 is focused on how visual appearance, in the context of semantic visual features like contrast and frequency, can influence specific aspects of human cognition: in particular we study the interplay of visual behavior and time perception in intervals of several seconds up to several minutes.

Humans rely on their sight to get much of the information related to extrapersonal space. In fact, humans can gather data about the illumination around a scene, its geometry and the visual appearance of objects in a glimpse [89]. How this physical, complex interaction is processed and

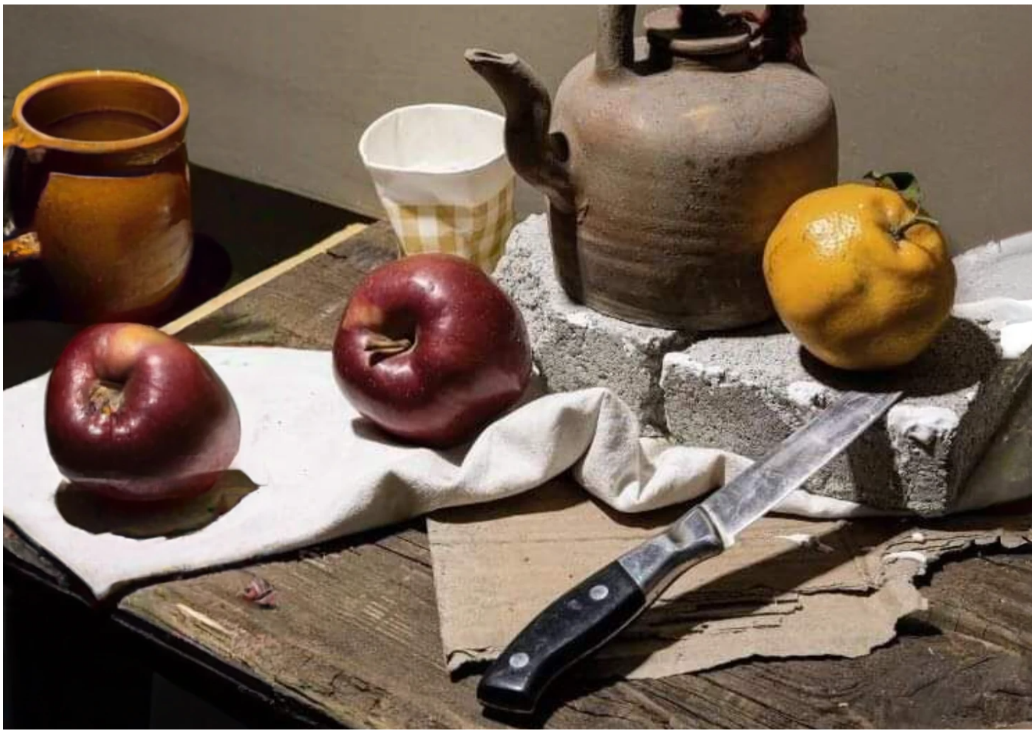


Figure 1.2: A still life picture of different objects. With a single glance we can infer different properties: how rough or hard each material will be to touch, the expected weight and relative temperature of each object, etc. Image generated with Dall-e [266] by the author.

integrated into perception is not well understood yet [279]. Just by looking at an object, a person can usually tell what the properties of the physical object are, and can tell different materials apart from each other under most circumstances (see Figure 1.2). Given the high dimensionality of material perception, many works have focused on modeling individual material attributes like glossiness, translucency or color. However, modeling how different or similar two materials are, following perceptual principles, is a well-known open problem in computer graphics for a number of reasons. Since the subjective nature of perception also plays a role in similarity, this concept is different from image similarity which can be defined as the difference between intensity patterns in two images. Additionally, materials are often represented with physically-based models that do not take into account human perception. On top of that a high number of parameters is needed to faithfully generate the appearance of an object, which means a correlation between such material models and perceived similarity is not straightforward.

Chapter 2 provides a computational model that measures appearance similarity using human perception as a learning base. Being able to compare how similar or different two objects are is an everyday ability that we perform often: to select the best piece of fruit in the supermarket, what materials will suit better in the decoration of a room or being able to distinguish a camouflaged animal in its habitat. Studying similarity (usually by pairwise comparisons, with or without a reference) allows researchers to understand how the brain represents and processes visual attributes in a non-invasive, systematic and objective way.

Traditional methods to measure similarity usually work in image space but do not take into account the influence of human perception, or the effect of different confounding factors (like geometry or illumination) into the final appearance. We propose to use a model that can learn from human perception in order to accurately reproduce the notion of human similarity. We devise a dataset of rendered images that represents a wide range of physical appearance, including different geometries, illuminations and real-world measured materials. Using online tools we run large-scale user studies to gather enough data to model how appearance similarity is perceived. With the data and the corresponding images we train a deep learning model that learns to correlate

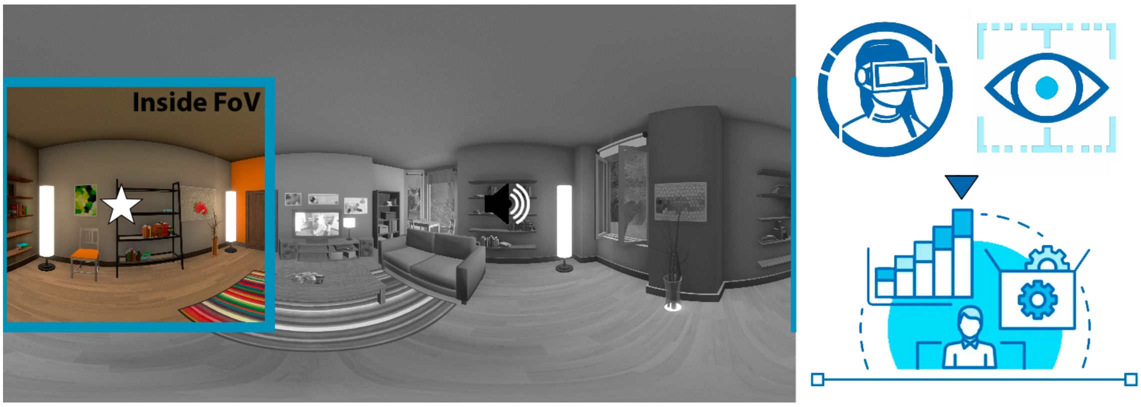


Figure 1.3: A visual target (white star) is presented inside the field of view of the user (colored area) while an auditory source is placed outside of the field of view (speaker in the gray area). We record qualitative (pre and post-test questionnaires, debriefing sessions, etc.) and quantitative metrics (eye tracking data, position in the virtual environment, task performance, etc.). The recorded data is then processed in order to obtain information about the latent cognitive processes that we want to study. The top right icons have been generated with Dall-e by the author.

with human behavior. We finally validate our model against previous work and propose several applications including material clustering and summarization, database visualization and the suggestion of materials of varying similarity.

Material perception is complex and depends on a combination of several features, including geometry, illumination, motion, physical materials, previous experience, multimodal interactions, etc. In Chapter 3 we move one step back and consider a simpler subset of asemantic visual features and their influence on perception. Visual stimuli can affect emotions which in turn modulate cognitive processes like attention, working memory and task performance [106]. In Chapter 3 we isolate visual behavior from other high-level cognitive aspects and study how visual behavior and time perception are related. Our perception of time affects our ability to process our surroundings, make predictions and interact with the environment[42]. Being able to better understand the interplay between visual behavior and time perception without the interference of other cognitive aspects will give content creators increased control over the momentum of their experiences. So far, manipulations of time perception have been leveraged in applications from medical therapy to training and performance metrics [318, 83].

We provide high-level notions of how changes in these visual features can affect time perception, both in traditional and in VR displays. We design two novel time judgment tasks that together with a carefully designed set of visual cues allow us to measure perceived changes on the passage of time in a series of user studies. This way, we disentangle the interplay between visual and time perception from other high-level cognitive processes such as emotion, memory or cognitive load. We find a consistent, significant correlation between larger visual spatiotemporal changes and a shortening of perceived time in a range of several seconds to several minutes (interval timing), while previous work had found a perceived expansion of perceived time under larger visual changes in a range of several milliseconds (millisecond timing). Shortening the perceived time could alleviate fatigue caused by prolonged sessions in VR thus potentially allowing for longer sessions in simulation and training applications and even reduce the experienced treatment time while undergoing medical care [318].

1.2 MULTIMODAL PERCEPTION IN IMMERSIVE ENVIRONMENTS

Besides sight, additional sensory modalities are necessary to increase our understanding of the world. Our brain integrates not only external information (like visual and auditory) but also internal stimuli to precisely situate itself in the environment. Due to its intrinsic characteristics, VR allows users to feel *present* in the virtual world and offers a unique way of integrating external and internal stimuli of several modalities in a natural way. Most VR systems include built-in headphones for a spatialized audio experience and support proprioception to some extent, translating the user real movement to the virtual world. Haptic feedback can be provided through a variety of gadgets, from controllers to haptic suits. Due to the high-dimensional, nonlinear neural processes in charge of all sensory perception, considering several modalities is not as simple as adding them up. In order to better understand the interplay of different modalities in immersive environments we first present a state of the art survey of multimodality in VR in Chapter 4.

We then investigate the potential crossmodal suppressive effects between auditory and visual cues in VR. Besides sight, hearing provides extrapersonal space information. This is particularly useful to alleviate the limitations of sight: in dark environments, behind occlusions or outside of the field of view (the extent of the observable world that we can see with our eyes). However, suppressive effects may appear if the spatial congruency between modalities is broken (see Figure 1.3 for an illustrative example). In Chapter 5 we present a user study that shows how visual performance can be degraded by presenting spatially incongruent sound in VR. Particularly, we present bursts of different sounds at the same time as visual targets the participants are asked to detect. The visual targets always appear inside the field of view, while the sounds are always spatially located outside of the field of view. Compared to a baseline visual-only condition participants experience a significant decrease in visual performance when sounds are present. In addition, we record eye tracking data during the experiment. We observe that even in the absence of saccades towards the sound source (an overt spatial attention redirection) the visual target is not perceived. In other words, the eyes of the participants can be located directly on the visual target and still not see it. This agrees with previous research carried out in traditional media with a more limited environment on the relationship between auditory and visual modalities. This effect could be used to induce subtle changes in the virtual environment without the user's awareness, either to guide attention or to be implemented in navigation techniques like redirected walking. Finally, we study the importance of the temporal window of integration of crossmodal events (Chapter 6), where the consistent presentation of spatial and temporal audiovisual information increases the perceived realism and physical properties of a set of materials in a virtual environment even when visual quality is degraded.

1.3 contributions and measurable results

1.3.1 *Publications*

In the following we state the publications which support the contributions of this thesis. Most of the presented work has been already published. In particular, in five journals indexed in JCR, including one paper in ACM Transactions on Graphics and presented at SIGGRAPH:

- Visual perception of realistic stimuli:
  - Perceptually-based metrics for appearance similarity (Chapter 2, Part II). The work on material appearance similarity was accepted in SIGGRAPH 2019, and published in ACM Transactions on Graphics [177]. This journal has an impact factor of 5.08, and its position in the JCR index is 8th out of 108 (Q1) in the category Computer Science, Software Engineering (data from 2019).
  - Time compression triggered by large visual changes (Chapter 3, Part II). The work on time perception was published in PLOS One [210]. This journal has an impact factor of 5.50, and its position in the JCR index is 9th out of 110 (Q2) in the category Multidisciplinary Sciences (data from 2022).
- Multimodal perception in immersive environments:
  - Multimodality in virtual reality (Chapter 4, Part III). The survey on multimodality was published in ACM Computing Surveys [218]. This journal has an impact factor of 14.32, and its position in the JCR index is 3rd out of 109 (Q1) in the category Computer Science, Theory & Methods (data from 2021).

- Auditory stimuli degrade visual performance in virtual reality (Chapter 5, Part III). The work on audiovisual illusions was published in Scientific Reports [213]. This journal has an impact factor of 4.37, and its position in the JCR index is 10th out of 142 (Q1) in the category Multidisciplinary Sciences (data from 2020).
- Crossmodal perception in virtual reality (Chapter 6, Part III). The work on crossmodal perception was published in Multimedia Tools and Applications [211]. This journal has an impact factor of 2.76, and its position in the JCR index is 224th out of 1027 (Q1) in the category Computer Science (data from 2020).

1.3.2 *Research internships*

Two research internships, totaling seven months, were carried out during this PhD:

- April 2019 – June 2019 (three months): Research Intern at *Adobe Research*, San Jose (California). Supervisors: Dr. Qi Sun and Dr. Zoya Bylinskii. As a result of this internship we published the work on time perception in virtual reality [210].
- September 2020 – December 2020 (four months): Research intern at *Facebook Reality Labs*, remote. Supervisor: Dr. Alex Locher (from the Immersive Mixed Reality Team at the Zurich office). The work carried out under this internship has been placed under a non-disclosure agreement.

1.3.3 *Supervised students*

During the development of this thesis I have supervised the following students:

- 2022. Daniel Jimenez (MSc Thesis) – Investigating auditory-triggered suppressive effects in virtual reality (9.0/10).
- 2021. Pedro Perez (BSc Thesis) – Development of a tool for data recording and visualization in virtual reality (9.0/10).
- 2021. Daniel Jimenez (MSc Internship) – Unity pipeline for perceptual experiments.
- 2020. Miguel Gomez (BSc Thesis) – Implementation and analysis of 2D scanpath prediction models (7.5/10).

1.3.4 *Research projects*

During my PhD studies I have participated in the following research project:

- CHAMELEON: Intuitive editing of visual appearance from real-world datasets. *European Research Council (ERC)*. Grant agreement No 682080. PI (in Spain): Diego Gutierrez.

1.3.5 *Reviews and conference organization*

During my PhD I have been a reviewer for a total of twelve different venues:

- 2022: Frontiers in Virtual Reality, ACM Special Interest Group on Graphics and Interactive Techniques (SIGGRAPH), IEEE International Symposium on Mixed and Augmented Reality (ISMAR), the Spanish Conference in Computer Graphics (CEIG), ACM Symposium on Applied Perception (SAP), International Journal of Human-Computer Interaction (IJHCI), the ACM Conference on Human Factors in Computing Systems (CHI) and Virtual Reality (VIRE).
- 2021: ACM Symposium on Virtual Reality Software and Technology (VRST), IEEE Conference on Virtual Reality (IEEEVR), ISMAR and VIRE.

- 2020: EuroVR International Conference (EuroVR) and VIRE.
- 2019: International Conference on Computer Graphics Theory and Applications (GRAPP), SAP and VRST.

Additionally, I also have been an International Program Committee Member for VRST 2021, a Committee Member for CEIG 2022 and the Posters Chair for ACM SAP 2022.

1.3.6 *Other activities*

During my PhD I've had an active attitude towards dissemination and outreach events, and I have participated in several talks and events to promote science, especially in those related with women in STEM. In particular, I have participated in the VR Day, the European Researcher's Night, the Girls' Day, Women Techmakers Zaragoza and the NEOCOM conference.

1.3.7 *Ethics statement and other considerations*

All experiments were carried out following the Helsinki recommendations, and ensuring data anonymization. Our experimental protocols follow the guidelines of the Consejo de Gobierno (Government Council) of Universidad de Zaragoza. At the beginning of the experiments, participants gave informed consent and were made aware of the possibility of stopping or abandoning the experiments at any point at their will.

While the author of this thesis is the leading author in many of the presented works, they have been done in collaboration with different colleagues. For this reason, the work described is contextualized at the beginning of each chapter and the contributions of the author of this thesis are discussed explicitly when needed.

Part II

VISUAL PERCEPTION

In this part we focus on the visual perception of realistic stimuli. It is important to note that, although both Chapters 2 and 3 are related with the study of visual appearance and share a user-centered approach, they address two fundamentally different problems through a common, objective methodology to evaluate human perception. The first one (Chapter 2) is focused on the perception of material appearance; it introduces a framework to model perceived similarity. The main contribution is the integration of large-scale data from human judgements within a deep learning model in order to create a similarity metric that aligns well with human perception. The second Chapter (Chapter 3) is focused on how visual appearance can influence specific aspects of human cognition. In particular, we study the existing correlation between low-level visual features and temporal perception in immersive environments. Through a series of user studies we find that larger visual changes (sequences of images or videos with higher contrast, faster stimuli, etc.) shorten perceived time in the interval (several seconds to several minutes) timing range.

2 Material Appearance Similarity

Here we describe a model to measure material similarity which strongly correlates with human similarity judgements. We create a database of varying materials, geometry and illuminations (9000 rendered images) and gather data on perceived similarity using large-scale crowdsourcing tools for the user studies (collecting over 114840 answers). We use the rendered images and gathered human judgements to train a deep learning model which learns a latent space representation that is close to human perception. Our evaluation shows that our model outperforms existing metrics. Finally, we propose several applications enabled by our metric, including appearance-based search for material suggestions, database visualization, clustering and summarization, and gamut mapping.

This work has been published in *ACM Transactions on Graphics* and presented at *SIGGRAPH 2019* [177]. I include here the full description of the work for completeness, but my main contributions can be found in Sections 2.2 (the design of the dataset), 2.3 (the design of the crowdsourced experiments), 2.5 (the evaluation of the trained network) and 2.6 (latent space visualizations, creating proof of concepts for some of the applications), as well as in the writing of the manuscript. A follow up work analyzing in more depth the role of objective and subjective measures in material similarity learning was later presented as a peer-reviewed poster in *ACM SIGGRAPH 2020 Posters* [73].

M. Lagunas, S. Malpica, A. Serrano, E. Garces, D. Gutierrez, & B. Masia
A Similarity Measure for Material Appearance
ACM Transactions on Graphics Vol. 38 (4), SIGGRAPH 2019

2.1 INTRODUCTION

Humans are able to recognize materials, compare their appearance, or even infer many of their key properties effortlessly, just by briefly looking at them. Many works propose classification techniques, although it seems clear that labels do not suffice to capture the richness of our subjective experience with real-world materials [97]. Unfortunately, the underlying perceptual process of material recognition is complex, involving many distinct variables; such process is not yet completely understood [10, 96, 208].

Given the large number of parameters involved in our perception of materials, many works have focused on individual attributes (such as the perception of gloss [274, 417], or translucency [115]), while others have focused on particular applications like material synthesis [430], editing [332], or filtering [154]. However, the fundamentally difficult problem of establishing a *similarity measure for material appearance* remains an open problem. Material appearance can be defined as "the visual impression we have of a material" [79]; as such, it depends not only on the BRDF of the material, but also on external factors like lighting or geometry, as well as human judgement [96, 1]. This is different from the common notion of image similarity (devoted to finding detectable differences between images, e.g., [408]), or from similarity in BRDF space (which has been shown to correlate poorly with human perception, e.g., [332]). Given the ubiquitous nature of photorealistic computer-generated imagery, and emerging fields like computational materials, a similarity measure of material appearance could be valuable for many applications.

Capturing a human notion of perceptual similarity in different contexts has been an active area of research recently [111, 2, 202]. In this work we develop a novel image-based material appearance similarity measure derived from a learned feature space. This is challenging, given the subjective nature of the task, and the interplay of confounding factors like geometry or illumination in the final perception of appearance. Very recent work suggests that perceptual similarity may be an emergent property, and that deep learning features can be trained to learn a representation of the world that correlates with perceptual judgements [428]. Inspired by this, we rely on a combination of large amounts of images, crowdsourced data, and deep learning. In

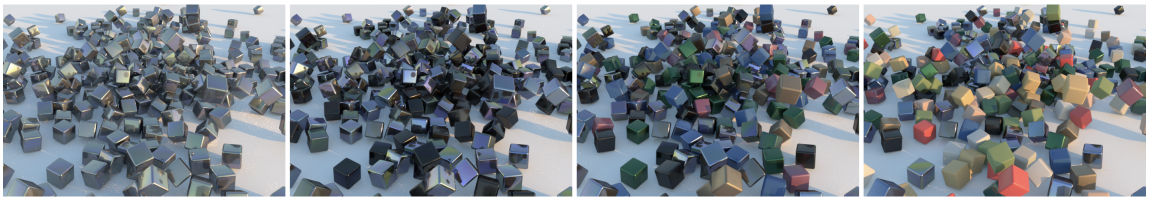


Figure 2.1: The cubes in the leftmost image have all been rendered with the same aluminum material. Our similarity measure for material appearance can be used to automatically generate alternative depictions of the same scene, where the similarity of the materials varies in a controlled manner. The next three images show results with materials randomly chosen by progressively extending the search distance from the original aluminum, from similar in appearance to farther away materials within the same dataset.

particular, we create a diverse collection of 9,000 stimuli using the measured, real-world materials in the MERL dataset [232], which covers a wide variety of isotropic appearances, and a combination of different shapes and environment maps. Using triplets of images, we gather information through Mechanical Turk, where participants are asked which of two given examples has a more similar appearance to a reference. Given our large stimuli space, we employ an adaptive sampling scheme to keep the number of triplets manageable. From this information, we learn a model of material appearance similarity using a combined loss function that enforces learning of the appearance similarity information collected from humans, and the main features that describe a material in an image; this allows us to learn the notion of material appearance similarity explained above, dependent on both the visual impression of the material, and the actual physical properties of it.

To evaluate our model, we first confirm that humans do provide reliable answers, suggesting a shared perception of material appearance similarity, and further motivating our similarity measure. We then compare the performance of our model against humans: Despite the difficulty of our goal, our model performs on par with human judgements, yielding results better aligned with human perception than current metrics. Last, we demonstrate several applications that directly benefit from our metric, such as material suggestions (see Figure 2.1), database visualization, clustering and summarization, or gamut mapping. In addition to the 9,000 rendered images, our database also includes surface normals, depth, transparency, and ambient occlusion maps for each one, while our collected data contains 114,840 answers; we provide both, along with our pre-trained deep learning framework, in order to help future studies on the perception of material appearance.

All the code, data, and models are available at: `webdiis.unizar.es/~mlagunas/publication/material-similarity/`

2.1.1 *Material perception*

There have been many works aiming to understand the perceptual properties of BRDFs [10, 99, 96, 208]; a comprehensive review can be found in the work of Thompson and colleagues [380]. Finding a direct mapping between perceptual estimates and the physical material parameters is a hard task involving many dimensions, not necessarily correlated. Many researchers focus on the perception of one particular property of a given material (such as glossiness [47, 274, 417], translucency [114, 115], or viscosity [389]), or one particular application (such as filtering [154], computational aesthetics [67], or editing [332, 251]). Leung and Malik [188] study the appearance of flat surfaces based on textural information. Other recent works analyze the influence on material perception of external factors such as illumination [137, 398, 172], motion [78], or shape [399, 127].

A large body of work has been devoted to analyzing the relationships between different materials, and deriving low-dimensional perceptual embeddings [232, 417, 332, 357]. These embeddings can be used to derive material similarity metrics, which are useful to determine if two materials convey the same appearance, and thus benefit a large number of applications (such as BRDF compression, fitting, or gamut mapping). A number of works have proposed different metrics, computed either directly over measured BRDFs [101, 255], in image space [275, 256, 371], or in reparameterizations of BRDF spaces based on perceptual traits [274, 332]. Our work is closer to the latter; however, rather than analyzing perceptual traits in isolation, we focus on the overall appearance of materials, and derive a similarity measure that correlates with the notion of material similarity as perceived by humans.

2.1.2 *Learning to recognize materials*

Image patches have been shown to contain enough information for material recognition [323], and several works have leveraged this to derive learning techniques for material recognition tasks. Bell et al. [29] introduce a CNN-based approach for local material recognition using a large annotated database, while Schwartz and Nishino explicitly introduce global contextual cues [322]. Other works add more information such as known illumination, depth, or motion. Georgoulis et al. [112] use both an object's image and its geometry to create a full reflectance map, which is later used as an input to a four-class coarse classifier (metal, paint, plastic or fabric). For a comprehensive study on early material recognition systems and latest advances, we refer to the reader to the work of Fleming [97]. These previous works focus mainly on classification tasks, however *mere labels do not capture the richness of our subjective experience of materials in the real world* [97].

Recent work has shown the extraordinary ability of deep features to match human perception in the assessment of perceptual similarity between two images [428]. Together with the success of the works mentioned above, this provides motivation for the combination of user data and deep learning that we propose in this work.

2.1.3 *Existing datasets*

Early image-based material datasets include CURet [68], KTH-TIPS [129], or FMD [343]. OpenSurfaces [28] includes over 20,000 real-world images, with surface properties annotated via crowdsourcing. This dataset has served as a baseline to others, such as the Materials in Context Database (MINC) [29], an order of magnitude larger; SynBRDF [166], with 5,000 rendered materials randomly sampled from OpenSurfaces; or MaxBRDF dataset [402], which includes synthetic anisotropic materials.

Databases with *measured* materials include MERL [232] for isotropic materials, UTIA [95] for anisotropic ones, the Objects under Natural Illumination Database [196], which includes calibrated HDR information, or the recent, on-going database by Dupuy and Jakob which measures spectral reflectance [82]. We choose as a starting point the MERL dataset, since it contains a wider variety of isotropic materials, and it is still being successfully used in many applications such as gamut mapping [371], material editing [332, 370], BRDF parameterization [357], or photometric light source estimation [200].

2.2 MATERIALS DATASET

2.2.1 *Why a new materials dataset?*

To obtain a meaningful similarity measure of material appearance we require a large dataset with the following characteristics:

- Data for a wide variety of materials, shapes, and illumination conditions.
- Samples featuring the *same* material rendered under different illuminations and with different shapes.
- Materials represented by measured BRDFs, with reflectance data captured from real materials.
- Real-world illumination, as given by captured environment maps.
- A large number of samples, amenable to learning-based frameworks.

These characteristics enable renditions of complex, realistic appearances and will be leveraged to train our model, explained in Section 2.4. To our knowledge, none of the existing material datasets features all these characteristics.

2.2.2 *Description of the dataset*

In the following, we briefly describe the characteristics of our dataset, and refer the reader to the supplementary material for further details.

Materials. Our dataset includes all 100 materials from the MERL BRDF database [232]. This database was chosen as starting point since it includes real-world, measured reflectance functions covering a wide range of reflectance properties and types of materials, including paints, metals, fabrics, or organic materials, among others.

Illuminations. We use six natural illumination environments, since they are favored by humans in material perception tasks [98]. They include a variety of scenes, ranging from indoor scenarios to urban or natural landscapes, as high-quality HDR environment maps.

The HDR illuminations are gathered from: `http://gl.ict.usc.edu/Data/HighResProbes/`

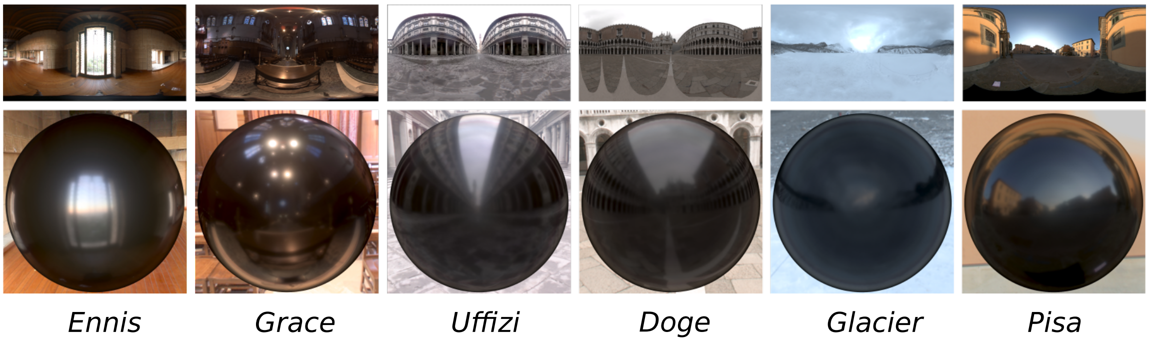


Figure 2.2: All six environment maps used in the dataset and corresponding rendered spheres with the *black-phenolic* material.

Scenes. Our database contains thirteen different 3D models, with an additional camera viewpoint for two of them, defining our fifteen scenes. It includes widely used 3D models, and objects that have been specifically designed for material perception studies [127, 399]. The viewpoints have been chosen to cover a wide range of features such as varying complexity, convexity, curvature, and coverage of incoming and outgoing light directions.

All HDR images are tone-mapped using the algorithm by Mantiuk et al. [214], with the predefined lcd office display, and color saturation and contrast enhancement set to 1.

By combining the aforementioned materials (100), illumination conditions (6), and scenes (15), we generate a total of 9,000 dataset samples using the Mitsuba physically-based renderer [413]. For each one we provide: The rendered HDR image, a corresponding LDR image, along with depth, surface normals, alpha channel, and ambient occlusion maps. While not all these maps are used in the present work, we make them available with the dataset should they be useful for future research. Figure 2.3 shows sample images for all fifteen scenes.

2.3 COLLECTING APPEARANCE SIMILARITY INFORMATION

We describe here our methodology to gather crowdsourced information about the perception of material appearance.

Stimuli. We use 100 different stimuli, covering all 100 materials in the dataset, rendered with the *Ennis* environment map. We choose the *Havran-2* scene, since its shape has been designed to maximize the information relevant for material appearance judgements by optimizing the coverage of incoming and outgoing light directions sampled [127]. Figure 2.4 shows some examples.

Participants. A total of 603 participants took part in the test through the Mechanical Turk (MTurk) platform, with an average age of 32, and 46.27% female. Users were not aware of the purpose of the experiment.

Procedure. Our study deals with the *perception* of material appearance, which may not be possible to represent in a linear scale; this advises against ranking methods [162]. We thus gather data in the form of relative comparisons, following a 2AFC scheme; the subject is presented with a triplet made up of one *reference* material, and two *candidate* materials, and their task is to answer the question *Which of these two candidates has a more similar appearance to the reference?* by choosing one among the two candidates. This approach has several additional advantages: it is easier for humans than providing numerical distances [234, 321], while it reduces fatigue and avoids the need to reconcile different scales of similarity among subjects [161].

However, given our 100 different stimuli, a naive 2AFC test would require 495,000 comparisons, which is intractable even if not all subjects see all comparisons. To ensure robust statistics, we aim to obtain five answers for each comparison, which would mean testing a total of 2,475,000 comparisons. Instead, we use an iterative *adaptive sampling* scheme [377]: For any given reference, each following triplet is chosen to maximize the information gain, given the preceding responses (please refer to the supplementary material of our work [177] for a more detailed description of the method). From an initial random sampling, we perform 25 iterations as recommended by Tamuz et al. for datasets our size; in each iteration we sample 10 new pairs for every one of our 100 reference materials, creating 1,000 new triplets. After this process, the mean information gain per iteration is less than 10^{-5}, confirming the convergence of the sampling scheme. This scheme allows us to drastically reduce the number of required comparisons, while providing a good approximation to sampling the full set of triplets.

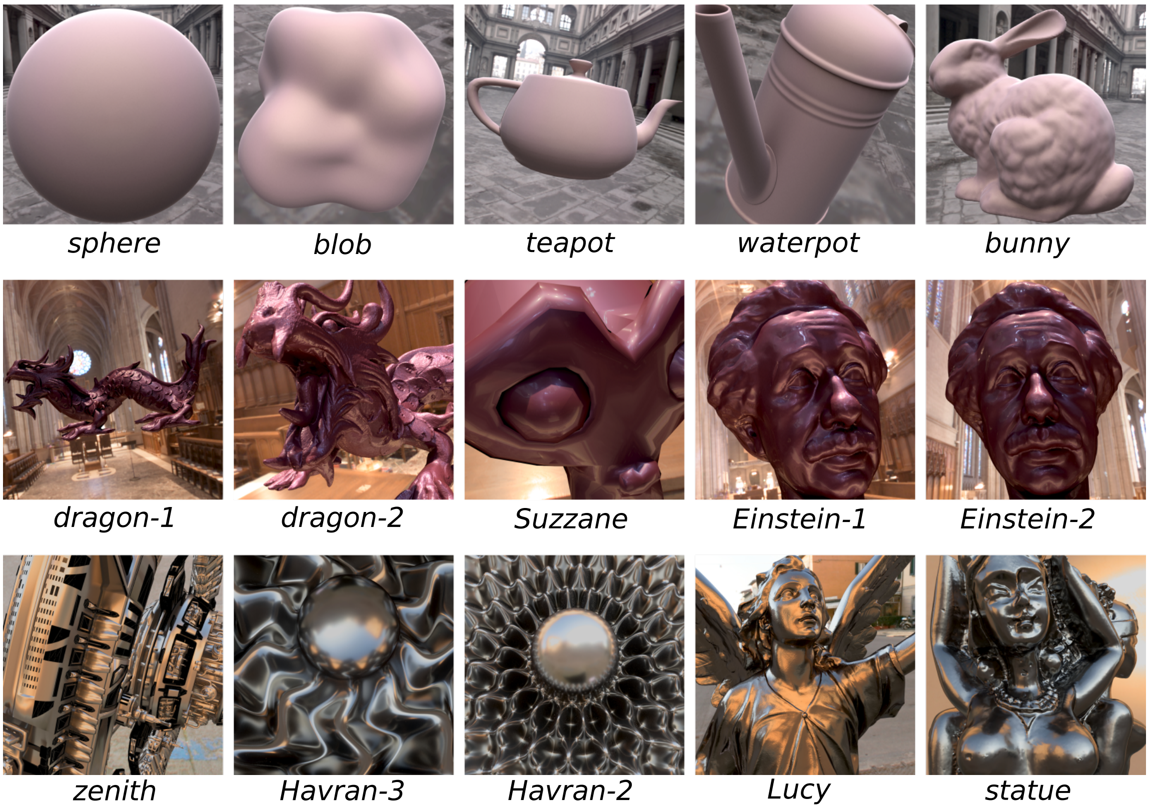


Figure 2.3: Sample images of all 15 scenes with different materials and illumination conditions. First row: *pink-felt* and *Uffizi*; second row: *violet-acrylic* and *Grace*; third row: *nickel* and *Pisa*. The 3D models *bunny*, *dragon*, *Lucy* and *statue* belong to The Stanford 3D Scanning Repository; *waterpot* (modelled by gykservy), *Suzzane* (killzone75), *Einstein* (oliverlaric), and *zenith* (KuhnIndustries) were obtained from TurboSquid.

Each test (HIT in MTurk terminology) consisted of 110 triplets. To minimize worker unreliability [413], each HIT was preceded by a short training session that included a few trial comparisons with obvious answers [302, 111]. In addition, ten control triplets were included in each HIT, testing repeated-trial consistency within participants. We adopt a conservative approach and reject participants with two or more different answers. In the end, we obtained 114,840 valid answers, yielding a participants' consistency of 84.7%.

As a separate test, to validate how well our collected answers generalize to other shapes and illuminations, we repeated the same comparisons, this time with symmetric and asymmetric triplets chosen randomly from our dataset, with the condition that they do not contain the *Havran-2* shape nor the *Ennis* illumination. For symmetric triplets, the three items in the triplet differ only in the material properties, while in asymmetric triplets geometry and lighting also vary. We launched 2,500 symmetric triplets, and found that participants' majority matched the previous responses with a 84.59% rate. When we added the same number of asymmetric triplets to the test, participants' answers held with a 80% match rate.

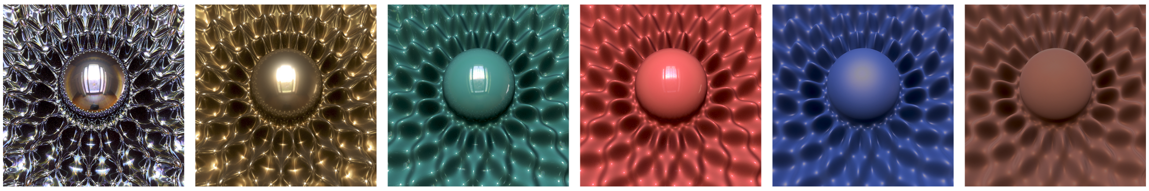


Figure 2.4: Sample stimuli for our appearance similarity collection. They correspond to the *Havran-2* scene, with materials from the MERL database, rendered with the *Ennis* environment map. In reading order: *chrome*, *gold-metallic-paint3*, *specular-green-phenolic*, *maroon-plastic*, *dark-blue-paint* and *light-brown-fabric*.

2.4 LEARNING PERCEIVED SIMILARITY

This section describes our approach to learn perceived similarity for material appearance. Given an input image ψ, our model provides a feature vector $f(\psi)$ that transforms the input image into a feature space well aligned with human perception.

We use the ResNet architecture [130], based on its generalization capabilities and its proven performance on image-related tasks. The novelty of this architecture is a residual block meant for learning a residual mapping between the layers, instead of a direct mapping, which enables training very deep networks (hundreds of layers) with outstanding performance. For training we use image data from our materials dataset (Section 2.2), together with human data on perceived similarity (Section 2.3). We first describe our combined loss function, then our training procedure.

2.4.1 *Loss function*

We train our model using a loss function consisting of two terms, equally weighted:

$$\mathcal{L} = \mathcal{L}_{TL} + \mathcal{L}_P \tag{2.1}$$

The two terms represent a perceptual triplet loss, and a similarity term, respectively. The terms aim at learning appearance similarity from the participants' answers, while extracting the main features defining the material depicted in an image. In the following, we describe these terms and their contribution.

2.4.1.1 *Triplet loss term $\mathcal{L}_{TL}$*

This term allows to introduce the collected MTurk information on appearance similarity. Let $\mathcal{A} = \{(r_i, a_i, b_i)\}$ be the set of answered relative comparisons, where r is the reference image, a is the candidate image chosen by the majority of users as being more similar to r, and b the other candidate; i indexes over all the relative comparisons. Intuitively, r and a should be closer together in the learned feature space than r and b. It is not feasible to collect user answers for all possible comparisons (n different images would lead to $n\binom{n-1}{2}$ tests); however, as we have shown in Section 2.3, the collected answers for a triplet (r, a, b) involving materials m^r, m^a and m^b generalize well to other combinations of shape and illumination from our dataset involving the same set of materials. We thus define $\mathcal{A}^M = \{(m_i^r, m_i^a, m_i^b)\}$ as the set of relative comparisons with collected answers (m^a represents the material chosen by the majority of participants). We then formulate the first term as a triplet loss [54, 319, 176]:

$$\mathcal{L}_{TL} = \frac{1}{|\mathcal{B}^A|} \sum_{(r,a,b) \in \mathcal{B}^A} \left[||f(r) - f(a)||_2^2 - ||f(r) - f(b)||_2^2 + \mu \right]_+ \tag{2.2}$$

where $f(\psi)$ is the feature vector of image ψ, and the set $\mathcal{B}^A$ is defined as:

$$\mathcal{B}^A = \left[(r, a, b) \mid (m^r, m^a, m^b) \in \mathcal{A}^M \ \wedge \ (r, a, b) \in \mathcal{B} \right] \tag{2.3}$$

with $\mathcal{B}$ the current training batch. In Eq. 2.2, μ represents the margin, which accounts for how much we aim to separate the samples in the feature space.

2.4.1.2 *Similarity term $\mathcal{L}_P$*

We introduce a second loss term that maximizes the log-likelihood of the model choosing the same material as humans. We define this probability p_{ra} (and conversely p_{rb}) as a quotient between similarity values s_{ra} and s_{rb}:

$$p_{ra} = \frac{s_{ra}}{s_{rb} + s_{ra}}, \quad p_{rb} = \frac{s_{rb}}{s_{rb} + s_{ra}} \tag{2.4}$$

These similarities are derived from the distances between r, a and b in the feature space, where a similarity value of 1 means perfect similarity and a value of 0 accounts for total dissimilarity:

$$s_{ra} = \frac{1}{1 + d_{ra}}, \quad s_{rb} = \frac{1}{1 + d_{rb}}, \quad \text{where} \tag{2.5}$$

$$d_{ra} = ||f(r) - f(a)||_2^2, \quad d_{rb} = ||f(r) - f(b)||_2^2 \tag{2.6}$$

With this, we can formulate the similarity term as:

$$\mathcal{L}_P = -\frac{1}{|\mathcal{B}^A|} \sum_{(r,a,b) \in \mathcal{B}^A} \log p_{ra} \tag{2.7}$$

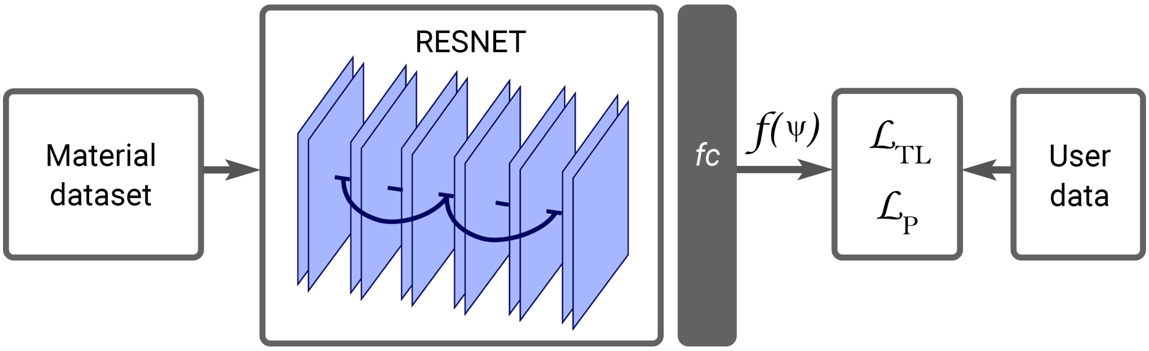


Figure 2.5: Scheme of the training process, using both image data from our material dataset, and human data of perceived similarity. We train our model so that, for an input image ψ, it yields a 128-dimensional feature vector $f(\psi)$.

2.4.2 *Training details*

For training, we remove the *Havran-2* and *Havran-3* scenes from the dataset, leading to 7,800 images (13 (scenes) × 6 (env. maps) × 100 (materials)), augmented to 39,000 using crops, flips, and rotations. These 39,000 images, together with the collected MTurk answers, constitute our training data. We use the corrected *Adam* optimization [290, 168] with a learning rate that starts at 10^{-3} to train the network. We train for 80 epochs and the learning rate is reduced by a factor of 10 every 20 epochs. For initialization, we use the weights of the pre-trained model [130] on ImageNet [75, 306]. To adapt the network to our loss function, we remove the last layer of the model and introduce a fully-connected (*fc*) layer that outputs a 128-dimensional feature vector $f(\psi)$. We use a margin $\mu = 0.3$ for the triplet loss term $\mathcal{L}_{TL}$. Figure 2.5 shows a scheme of the training procedure.

2.5 EVALUATION

We evaluate our model on the set of images of the material dataset not used during training. We employ the *accuracy* metric, which represents the percentage of triplet answers correctly predicted by our model. It can be computed as *raw*, considering each of the five answers independently as the correct one, or *majority*, considering the majority opinion as correct [417, 111]. Using our MTurk data from Section 2.3, the results are 73.10% and 77.53% respectively for human observers, indicating a significant agreement across subjects. Our model performs better than human accuracy, with 73.97% and 80.69% respectively. In other words, our model predicts the majority's perception of similarity almost 81% of the time. We include an *oracle* predictor in Table 2.1, which has access to all the human answers and returns the majority opinion; note that its raw accuracy is not 100 due to human disagreement. Figure 2.6 shows examples from our 26,000 queries where our model agrees with the majority response, while we discuss failure cases later in this section. More examples of queries and our model's answer are included in the supplementary material.

2.5.1 *Comparison with other metrics*

We compare the performance of our model to six different metrics used in the literature for material modeling and image similarity: The three common metrics analyzed by Fores and colleagues [101], the perceptually-based metrics by Sun et al. [371] and Pereira et al. [275], and SSIM [408], a well-known image similarity metric. We analyze again accuracy, and we additionally analyze *perplexity*, which is a standard measure of how well a probability model predicts a sample, taking into account the uncertainty in the model. Perplexity Q is given by:

$$Q = 2^{-\frac{1}{|\mathcal{A}|}\sum_{\Omega}\log_2 p_{ra}} \tag{2.8}$$

where $\Omega = (r, a) \in \mathcal{A}$, $|\mathcal{A}|$ is the number of collected answers, and p_{ra} is the probability of a being similar to r (Section 2.4.1). Perplexity gives higher weight where the model yields higher confidence; its value will be 1 for a model that gives perfect predictions, 2 for a model with total uncertainty (random), and higher than 2 for a model that gives wrong predictions. As Table 2.1 shows, our model captures the human perception of appearance similarity significantly better, as indicated by the higher accuracy and lower

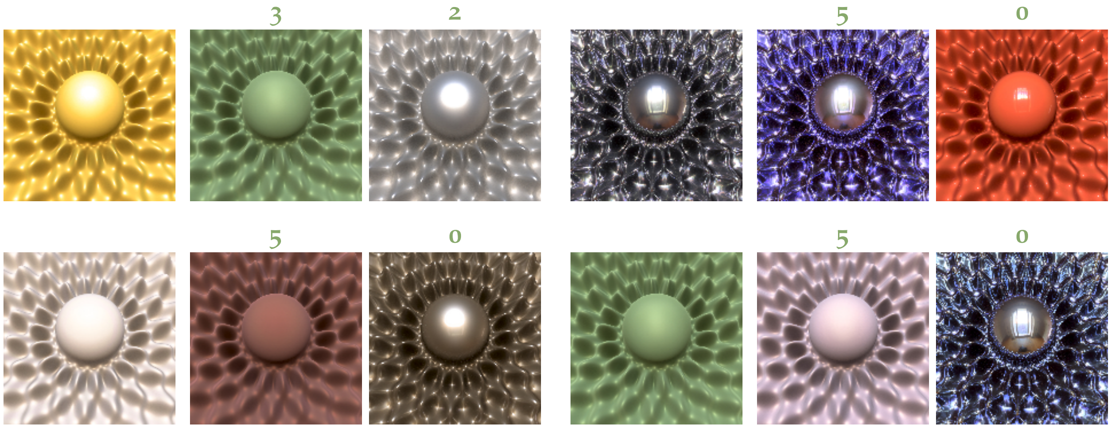


Figure 2.6: Examples from our 26,000 queries (reference, plus the two candidates) where our model agrees with the majority response (this is the case almost 81% of the time). The numbers indicate the number of votes each image received from the participants. More examples are included in the supplementary material.

perplexity values. Note that perplexity cannot be computed for humans nor the oracle, since they are not probability distributions.

Additionally, we compute the mean error between distances derived from human responses and our model's predictions, across all possible material pair combinations from the MERL dataset. To obtain the derived distances from the collected human responses, we use t-Distributed Stochastic Triplet Embedding (tSTE) [393], which builds an n-dimensional embedding that aims to correctly represent participants' answers. We use a value of $\alpha = 5$ (degrees of freedom of the Student-t kernel), which correctly models 87.36% of the participants' answers. We additionally compute the mean error for the six other metrics. As shown in Figure 2.7, our metric yields the smallest error. Error bars correspond to a 95% confidence interval.

2.5.2 *Ablation study*

We evaluate the contribution of each term in our loss function to the overall performance via a series of ablation experiments (see Table 2.2). We first evaluate performance using only one of the two terms ($\mathcal{L}_{TL}$ and $\mathcal{L}_P$) in isolation. We also analyze the result of incorporating two additional loss terms, which could in principle apply to our problem: A cross-entropy term $\mathcal{L}_{CE}$, and a batch-mining triplet loss term $\mathcal{L}_{BTL}$. The former aims at learning a soft classification task by penalizing samples which do not belong to the same class [374], while the latter has been proposed in combination with the cross-entropy term to improve the model's generalization capabilities and accuracy [108] (more details about these two terms can be found in the appendix). Last, we analyze performance using *only* these two terms ($\mathcal{L}_{CE}$ and $\mathcal{L}_{BTL}$), without incorporating participants' perceptual data. As Table 2.2 shows, none of these alternatives outperforms our proposed loss function. Although the single-term $\mathcal{L}_P$ loss function yields higher accuracy, it also outputs higher perplexity values; moreover, as Figure 2.7 shows, the mean error is much higher, meaning that it does not capture the notion of similarity as well as our model.

2.5.3 *Alternative networks*

We have tested two alternative architectures, VGG [352], which stacks convolutions with non-linearities; and DenseNet [145], which introduces concatenations between different layers. Both models have been trained using our loss function. As shown in Table 2.2, both yield inferior results compared to our model. DenseNet has a low number of learned parameters, insufficient to capture the data distribution, hampering convergence. VGG has a larger number of parameters; however, the residual mapping learned by the residual blocks in the architecture of our model yields the best overall performance.

Table 2.1: Accuracy and perplexity of our model compared to human performance, an oracle (which always returns the majority opinion), and six other metrics from the literature: RMS, RMS-cos, Cube-root [101], L2-lab [371], L4-lab [275] and SSIM [408]. For accuracy, higher values are better, while for perplexity lower are better.

Evaluation of our model

Metric	Accuracy		Perplexity	
	Raw	Majority	Raw	Majority
Humans	73.10	77.53	-	-
Oracle	83.79	100.0	-	-
RMS	61.63	64.72	3.61	3.13
RMS-cos	61.60	64.67	3.86	3.33
Cube-root	63.71	67.40	1.96	1.86
L2-lab	63.76	67.21	2.16	2.07
L4-lab	60.60	62.93	15.36	11.66
SSIM	62.35	64.74	2.02	1.94
Our model	73.97	80.69	1.74	1.55

2.5.4 *Results by category*

We additionally divide the materials into eight categories: *acrylics, fabrics, metals, organics, paints, phenolics, plastics*, and *other*, and analyze raw and majority accuracy in each. We can see in Table 2.3 how our model is reasonably able to predict human perception also within each category. For instance, although the numbers are relatively consistent across all the categories, humans perform on average slightly worse for phenolics or acrylics, and better for fabrics; our metric mimics such behavior. The only significant difference occurs within the *organics* category, where our metric performs worse than humans. This may be due to the combination of a low number of material samples and a large variety of appearances within such category, which may hamper the learning process.

2.5.5 *Failure cases*

Being on par with human accuracy means that our similarity measure disagrees with the MTurk majority 19.31% of the time. Figure 2.8 shows two examples where humans were consistent in choosing one stimuli as closer to the reference (5 votes out of 5), yet our metric predicts that the second one is more similar. In the leftmost example, the softness of shadows may have been a deciding factor for humans. In the rightmost example, humans may have been overly influenced by color, whilst our metric has factored in the presence of strong highlights. These examples are interesting since they illustrate that neither color nor reflectance are persistently the dominant factors when humans judge appearance similarity between materials.

2.6 applications

We illustrate here several applications directly enabled by our similarity measure.

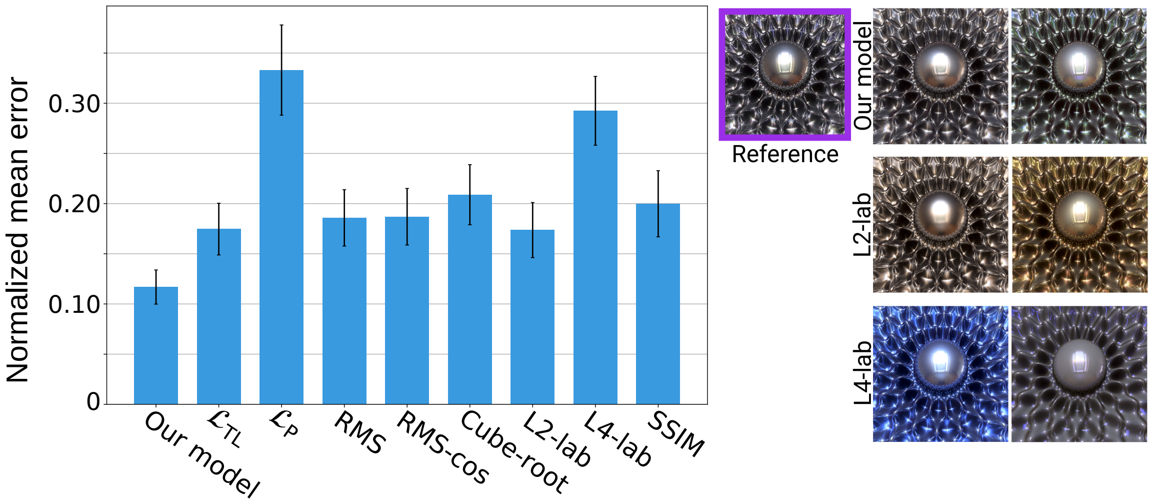


Figure 2.7: **Left**: Mean error for different metrics (each normalized by its maximum value) with respect to distances derived from human responses, across all possible pair combinations from the MERL dataset (the $\mathcal{L}_{TL}$ and $\mathcal{L}_P$ columns refer to the ablation studies in Table 2.2; please refer to the main text). Error bars correspond to a 95% confidence interval. **Right:** Representative example of the two most similar materials to a given reference, according to (from top to bottom): Our model, and the two perceptually-based metrics L2-lab [371], and L4-lab [275]. Our model yields less error, and captures the notion of appearance similarity better.

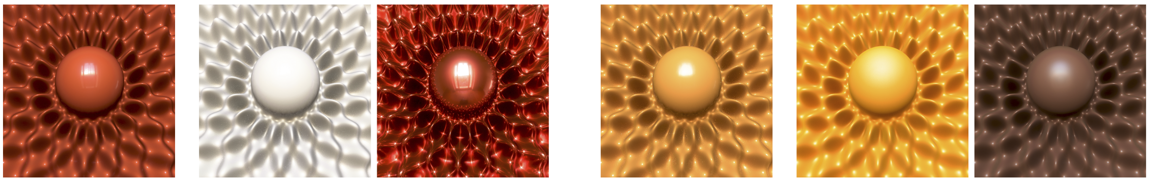


Figure 2.8: Two examples where humans' majority disagrees with our metric. For both, humans agreed that the middle stimulus is perceptually closer to the reference on the left, while our metric scores the right stimuli as more similar.

2.6.1 *Material suggestions*

Assigning materials to a complex scene is a laborious process [430, 53]. We can leverage the fact that the distances in our learned feature space correlate with human perception of similarity to provide controllable material suggestions. The artist provides the system with a reference material, and the system delivers perceptually similar (or farther away) materials in the available dataset, thus creating a controlled amount of variety without the burden of manually selecting each material. Figure 2.1 illustrates this, where the search distance is progressively extended from a chosen reference, and the materials are then assigned randomly to each cube. Suggestions need not be automatically assigned to the models in the scene, but may also serve as a palette for the artist to choose from, facilitating browsing and navigation through material databases. Figure 2.9 shows two MERL samples used as queries, along with returned suggestions from the *Extended MERL* dataset [332]. The figure shows results at close, intermediate, and far distances from the query. Additional examples can be seen in Figure 2.10, and in the supplementary material.

2.6.2 *Visualizing material datasets*

The feature space computed by our model can be used to visualize material datasets in a meaningful way, using dimensionality reduction techniques. We illustrate this using UMAP (Uniform Manifold Approximation and Projection [236]), which helps visualization by preserving the global structure of the data. Figure 2.11 shows two results for the MERL dataset, using images not

Table 2.2: Accuracy and perplexity for other loss functions, as well as for two alternative architectures (VGG and DenseNet).

ABLATION STUDY AND ALTERNATIVE NETWORKS

Model	Accuracy		Perplexity	
	Raw	Majority	Raw	Majority
$\mathcal{L}_{TL}$	69.32	74.12	1.89	1.73
$\mathcal{L}_{P}$	75.22	82.31	3.16	2.13
$\mathcal{L}_{TL}+\mathcal{L}_{P}+\mathcal{L}_{CE}$	71.82	77.53	1.76	1.66
$\mathcal{L}_{TL}+\mathcal{L}_{P}+\mathcal{L}_{CE}+\mathcal{L}_{BTL}$	71.78	77.76	1.76	1.67
$\mathcal{L}_{CE}+\mathcal{L}_{BTL}$	56.88	58.44	1.96	1.93
VGG	70.70	76.40	2.25	1.89
DenseNet	60.90	63.49	2.66	2.46
Our model	73.97	80.69	1.74	1.55

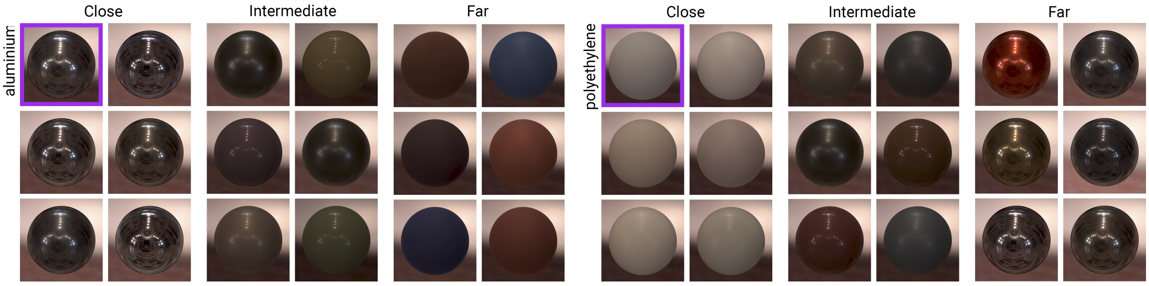


Figure 2.9: Two examples of material suggestions using our model. Queries from MERL (violet frame), and returned results for perceptually close, intermediate, and far away materials from the Extended MERL dataset.

included in the training set. On the left, we can observe a clear gradient in reflectance, increasing from left to right, with color as a secondary, softer grouping factor. The right image shows a similar visualization using only three categories: *metals*, *fabrics*, and *phenolics*.

2.6.3 *Database clustering*

For unlabeled datasets like Extended MERL, our feature space allows to obtain clusters of perceptually similar materials. To further analyze the clustering enabled by our perceptual feature space, we rely on the Hopkins statistic, which estimates randomness in a data set [22]. A value of 0.5 indicates a completely random distribution, lower values suggest regularly-spaced data, and higher values (up to a maximum of 1) reveal the presence of clusters. The Hopkins statistic computed over our 128-dimensional feature vectors for the Extended MERL dataset yields a value of 0.9585, suggesting that meaningful clusters exist in our learned feature space (Figure 2.13 shows three representative clusters using the Extended MERL database). For comparison purposes, using only *metals* in MERL the Hopkins statistic drops to 0.6935, since their visual features are less varied within that category. Figure 2.12 shows an example of material suggestions leveraging our perceptual clusters in unlabeled datasets.

The Hopkins statistic is an averaged value over 100 iterations since its computation involves random sampling of the elements in the dataset.

Table 2.3: Statistics per category. **From left to right:** Category, number of materials in each category, number of collected answers, humans' accuracy (raw and majority), accuracy of our model, and oracle raw accuracy.

Analysis per material category

Category	Materials	Answers	Humans		**Our model**		Oracle
			Raw	Majority	Raw	Majority	Raw
Acrylics	4	4719	67.27	70.69	67.57	74.18	79.89
Fabrics	14	16019	79.65	83.70	83.03	90.44	87.87
Metals	26	32337	74.20	78.90	75.63	83.10	84.54
Organics	7	8370	69.28	73.08	60.46	62.43	81.28
Paints	14	15101	74.22	78.85	75.22	81.84	84.61
Phenolics	12	13025	66.49	70.53	67.62	74.36	79.72
Plastics	11	12031	70.53	74.70	69.25	74.06	82.05
Other	12	13198	74.80	79.38	78.21	86.11	84.89
Total	100	114800	73.10	77.53	73.97	80.69	83.79

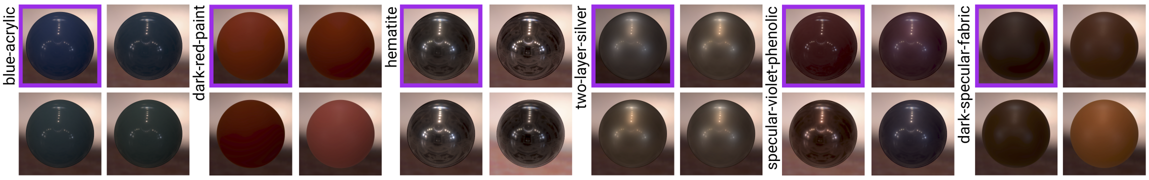


Figure 2.10: Additional material suggestion results. Queries (violet frame) and results for the closest materials in the *Extended MERL* dataset.

2.6.4 *Database summarization*

Perceptually meaningful clustering leads in turn to the possibility of database summarization. We can estimate the appropriate number of clusters using the elbow method, taking the number of clusters that explains the 95% of the variance in our feature vectors. In the 400-sample Extended MERL dataset, this results in seven clusters. Taking the closest material to the centroid for each one leads to a seven-sample database summarization that represents the variety of material appearances in the dataset (Figure 2.14).

2.6.5 *Gamut mapping*

In general, our model can be used for tasks that involve minimizing a distance. This is the case for instance of gamut mapping, where the goal is to bring an out-of-gamut material into the available gamut of a different medium, while preserving its visual appearance; this is a common problem with current printing technology, or in the emerging field of computational materials. We illustrate the effectiveness of our technique in the former. Gamut mapping can be formulated as a minimization on image space [275, 371]. We can use our feature vector $f(\psi)$ to minimize the perceptual distance between two images as

$$min_w||f(o) - f(g * w)||_2^2, \tag{2.9}$$

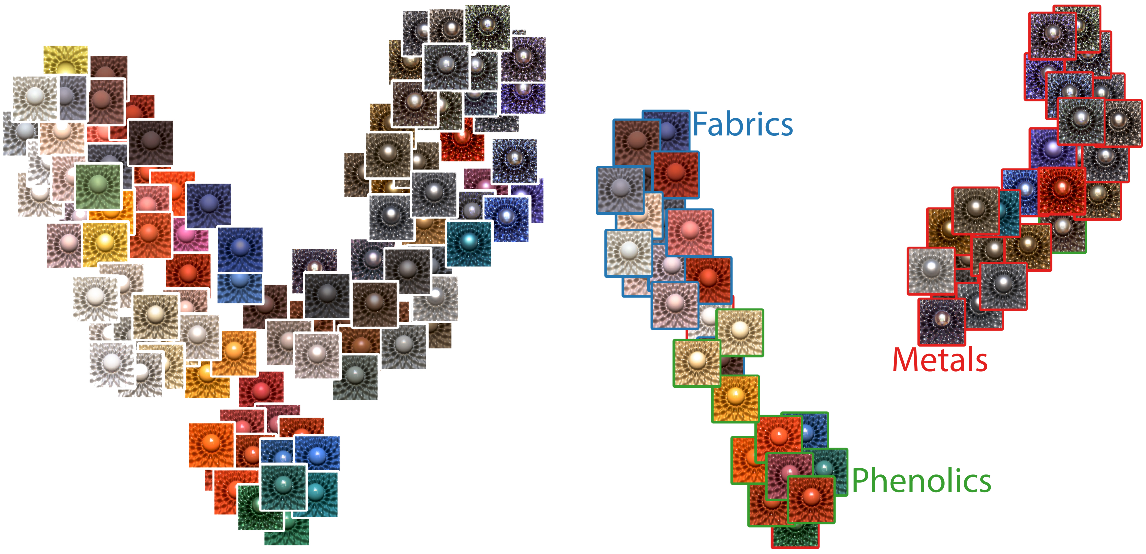


Figure 2.11: Visualization of the MERL dataset in a 2D space based on the feature vectors provided by our model, using UMAP [236]. **Left:** The entire MERL dataset. **Right:** Materials from three different categories (*metals*, *fabrics*, and *phenolics*).

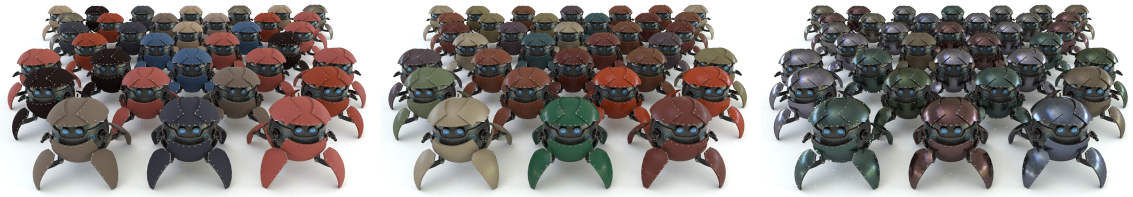


Figure 2.12: Material suggestions using our perceptual database clustering. The images show random materials assigned from three different clusters of varying appearance. The robot model (cKalten) was obtained from TurboSquid.

where o is the out-of-gamut image, and $g * w$ represents the image in the printer's gamut, defined as a linear combination of inks g [231]). Figure 2.15 shows some examples.

2.7 DISCUSSION

We have presented and validated a model of material appearance similarity that correlates with the human perception of similarity. Our results suggest that a shared perception of material appearance does exist, and we have shown a number of applications using our metric. Nevertheless, material perception poses many challenges; as such there are many exciting topics not fully investigated in this work. Several factors come into play that influence material appearance, i.e., the visual impression of a material, in a highly complex manner; fully identifying them and understanding their complex interactions is an open, fundamental problem. As a consequence of these interactions, the same material (e.g., plastic) may have very diverse visual appearances, whereas two samples of the same material may look very different under different illumination conditions [399, 98]. In aiming for material appearance similarity, we aim for a material similarity metric that can predict human judgements. There is a distinction, common in fields like psychology or vision science, between the distal stimulus—the physical properties of the material—, and the proximal stimulus—the image that is the input to perception—. The key observation here is that human perceptual judgements usually lie between these two, and our training framework and loss function are designed to take both into account. We combine the information about the physical properties

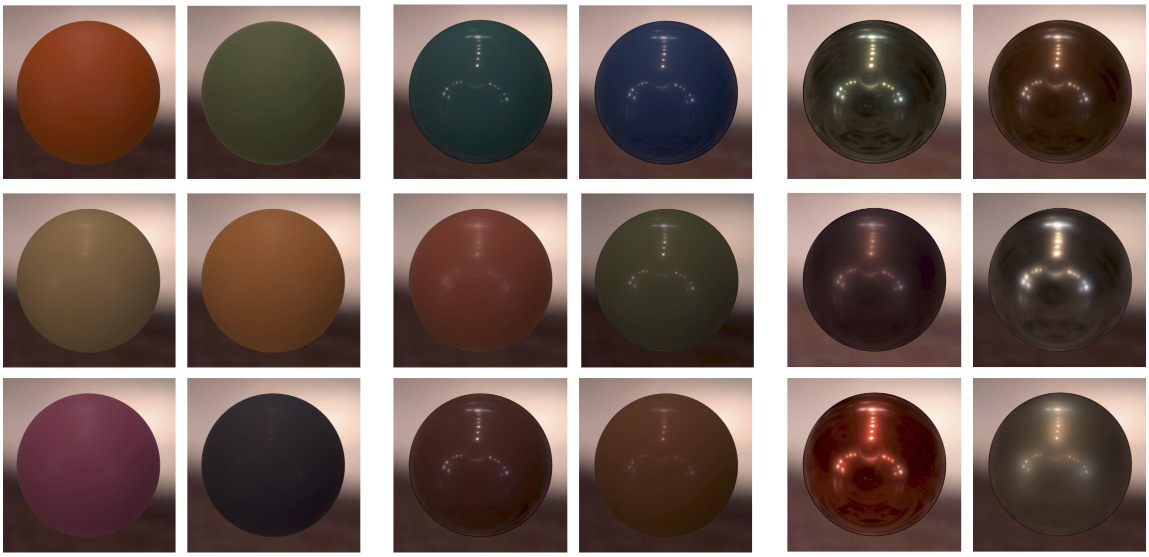


Figure 2.13: Representative samples of three clusters on the Extended MERL database. The Hopkins statistic on our feature space confirms that our similarity metric creates perceptually-meaningful clusters of materials.

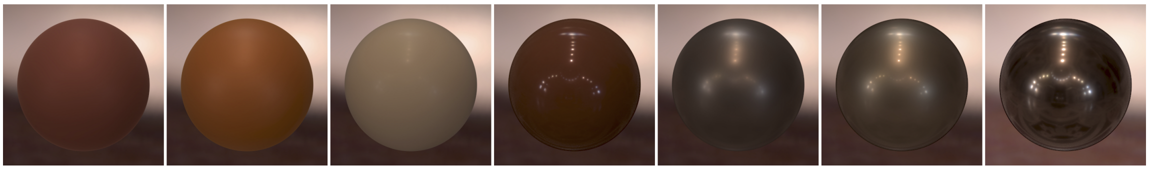


Figure 2.14: Example of database summarization for the Extended MERL dataset. These seven samples represent the variety of material appearances in the dataset.

of the material contained in the images, by having the same material under different geometries and illuminations, with the human answers on appearance similarity. In other words, a pure image similarity metric would not be able to generalize across shape, lighting or color, while a BRDF-based metric would be unable to predict human similarity judgements.

We do not attempt to identify nor classify materials (Figure 2.16). Our loss function could, however, incorporate additional terms (such as the cross-entropy and batch-mining triplet loss term discussed in the appendix) to help with classification tasks. We have carried out some tests and found anecdotical evidence of this, but a thorough analysis requires a separate study not covered in this work.

Despite having trained our model on isotropic materials, we have found that it may also yield reasonable results with higher-dimensional inputs. Figure 2.17 shows three examples from the Flickr Material Database (FMD) [343], which contains captured images of highly heterogeneous materials. We have gathered all the materials from the *fabrics*, *metals*, and *plastics* categories in the database; taking one reference from each, we show the three closest results returned by our model, using an L2 norm distance in feature space. Images were resized to match the model's input size, with no further preprocessing. Note that the search was not performed within each category but across all three, yet our model successfully finds similar materials for each reference. This is a remarkable, promising result; however, a more comprehensive analysis of in-the-wild, heterogeneous materials is out of the scope of this work.

We have also tested the performance of our model on grayscale images. In this case, we have repeated the evaluation conducted in Table 2.1 for our model, using grayscale counterparts of the images. Despite the removal of color information, we obtain results similar to those of our model on color images: A raw accuracy of 72.55 (vs 73.97 on color images), a majority accuracy of 78.64

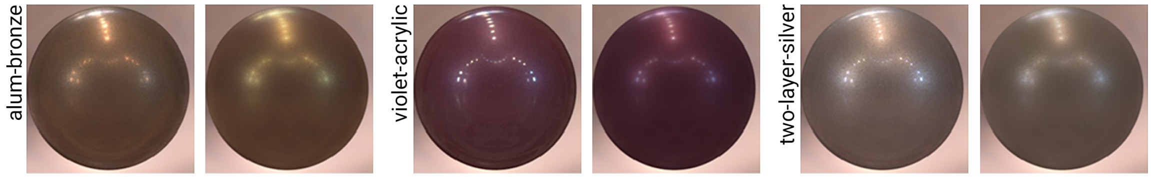


Figure 2.15: Our similarity metric can be used for gamut mapping applications, by minimizing the perceptual distance of our feature vectors. Each pair shows the ground truth (left), and our in-gamut result (right).

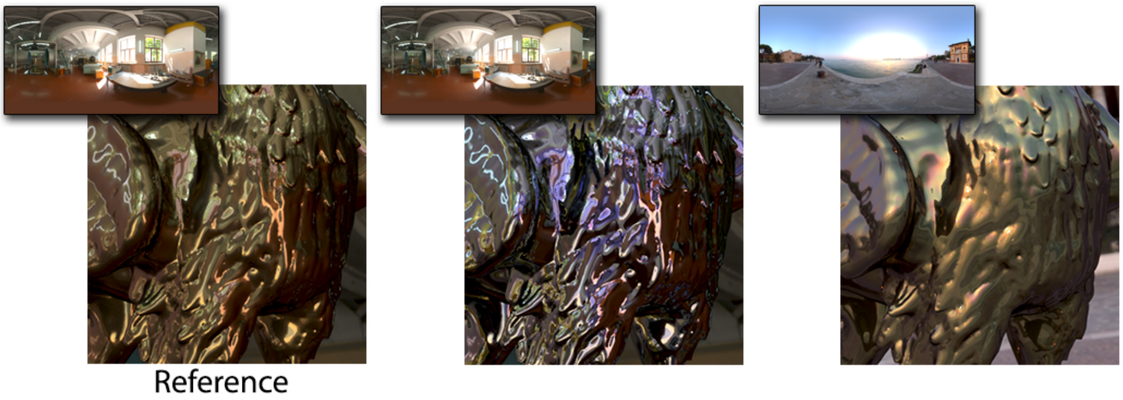


Figure 2.16: In the feature space defined by our model, the middle image (*chrome*) is closer in appearance to the reference (*brass*) than the image on the right (*brass*). The insets show the environment maps used. Our model is driven by appearance similarity, and does not attempt to classify materials.

(vs 80.69), a raw perplexity of 1.82 (vs 1.74), and a majority perplexity of 1.67 (vs 1.55). This further enforces the idea that we learn a measure of appearance similarity, and not image similarity.

To collect similarity data for material appearance, we have followed an adaptive sampling scheme [377]; following a different sampling strategy may translate into additional discriminative power and further improve our results. Our model could potentially be used as a feature extractor, or as a baseline for transfer-learning [344, 422] in other material perception tasks. A larger database could translate into an improvement of our model's predictions; upcoming databases of complex measured materials (e.g., Dupuy et al. [82]) could be used to expand our training data and lead to a richer and more accurate analysis of appearance. Our methodology for data collection and model training could be useful in these cases. Similarly, upcoming network architectures that may outperform our ResNet choice could be adopted within our framework. Finding hand-engineered features could also be an option and may increase interpretability, but it could also introduce bias in the estimation.

In addition to the applications we have shown, we hope that our work can inspire additional research and different applications. For instance, our model could be of use for designing computational fabrication techniques that take into account perceived appearance. It could also be used as a distance metric for fitting measured BRDFs to analytical models, or even to derive new parametric models that better convey the appearance of real world materials. We have made our data available for further experimentation, in order to facilitate the exploration of all these possibilities.

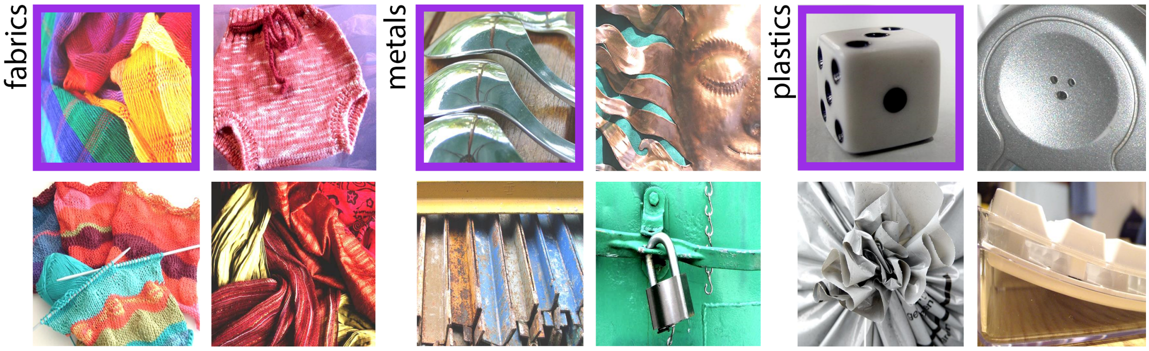


Figure 2.17: Results using highly heterogeneous materials from the FMD dataset. We show the three closest results returned by our model, from the reference materials highlighted in violet. Note that the search was performed across all three categories shown, not within each category.

3

Asemantic Visual Changes Affect Time Perception

So far we have dealt with how static visual appearance is perceived. However, our day to day lives are experienced through time, and we usually perceive a changing environment. The perception of our surroundings is created and updated constantly, with time perception playing a key role in our ability to make predictions and interact with our surroundings. Despite being a subjective experience, time perception depends not only on inner factors like the emotional state but also on the sensory information that is perceived at each moment. Time perception is known to be affected by visual appearance and sometimes it is even considered as an independent sensory modality. However, it is difficult to disentangle the effect of visual appearance from other high-level cognitive processes, since visual information can also have an effect on factors like emotional valence, cognitive load, etc. In this Chapter we focus on isolating low-level, basic features of visual appearance and studying how these affect time perception in immersive environments, since they provide a more natural and realist environment to work with than traditional media. We find that larger visual changes (elicited by sequences of images or videos with higher contrast, faster frequency, bigger field of view or higher visual complexity) shorten the perceived time.

This work has been published in PLOS One [210]. A previous work in progress was presented as a peer-reviewed poster in *Vision Science Society 2020* [209]. This work was started during my internship in Adobe under the supervision of Qi Sun and Zoya Bylinskii. While I was the leading author the rest of my coauthors collaborated with the design of the experiments and the writing. Zoya Bylinskii also run a part of the experiments.

S. Malpica, B. Masia, L. Herman, G. Wetzstein, DM. Eagleman, D. Gutierrez, Z. Bylinskii & Q. Sun
Larger visual changes compress time: The inverted effect of asemantic visual features on interval time perception
PLOS One 17(3) 2022

3.1 INTRODUCTION

Our perception of time crucially affects our ability to process our surroundings, make predictions, and act in real and simulated environments. Manipulations of time perception have been leveraged in broad applications. For instance, patients undergoing medical procedures can experience shorter durations of time, helping them cope with anxiety [318]; urban planning is also affected by time perception, since perceived waiting times in transit can be altered by the presence of basic amenities, indirectly affecting productivity [88]; and reaction times have to be considered in the context of public safety, especially for those who are driving [410]. Time perception can also be used as a proxy to measure performance in professional training, particularly when considering fitness for duty [83]. Evidence suggests that varied biological mechanisms might be involved in the perception of different temporal durations [42], including sub-second range (millisecond timing), seconds-to-hours range (interval timing), and a 24-hour range (circadian timing).

In this work, we define *asemantic visual features* as lower-level factors that are intrinsically related to visual processing (e.g., luminance contrast, temporal frequency), but not high-level cognitive aspects like emotions. In contrast, we denote factors related to high-level cognitive aspects as *semantic features*. Millisecond range time perception has been reported as being affected by purely visual, low-level patterns [421], which could be disentangled from other high-level cognitive aspects [23]. For instance, a positive correlation was found between visual magnitude and time perception at the millisecond level: higher-magnitude stimuli (e.g., stronger luminance, larger sizes, etc.) were judged to last longer in a prospective paradigm (participants knew beforehand that they would be making a temporal judgement [411]). In comparison, interval timing is the most common duration in high-order, real-world tasks, and it is known to be altered by semantic features such as emotional valence, arousal, cognitive load, or zeitgebers [318, 392, 316]. Manipulating the visual content to be sad or funny [392], varying the task difficulty (e.g., asking participants to solve either a 2D or 3D puzzle) [318], or having participants notice the changes in the illumination of a scene due to the position of the sun [316] are all examples of manipulations to semantic features that have been reported to affect time perception at interval timing durations.

In these previous works, high cognitive load or arousal generally shortened perceived time at the level of multiple seconds to more than one hour. However, since features like emotional valence and arousal are partially subjective and individualized, it is difficult to consistently and universally affect time perception by manipulating semantic features. It is much more practical to manipulate asemantic visual features, since they can be explicitly controlled (e.g., by computer rendering photo-realistic stimuli). Despite this, the effects of asemantic visual features on interval timing remain significantly under-explored. This is likely because, within interval timing, it is difficult to entirely disentangle asemantic from semantic features (e.g., manipulating color is known by content creators to elicit different emotions [159]). While past work has thus mainly considered the effects of high-level semantic features on time perception, we manipulate asemantic visual features by carefully curating the experimental stimuli.

In this work, we aim to answer the fundamental question "whether and how do asemantic visual features alter human time perception at the interval level?". Specifically, we investigate the prospective paradigm informing participants from the beginning that they will be making judgements related to time. This contrasts with the retrospective paradigm, in which participants report on the passage of time at the end of a task, without being informed in advance that a temporal judgement will be solicited. Contrary to retrospective judgements that rely on memory [400], we rely on prospective judgements to simulate ecologically valid applications and understand how time perception is affected in *real-world* scenarios in which people are actively aware of the passage of time.

Our experiments are designed to test the following hypothesis: differences in the magnitude of asemantic visual features should alter time perception in prospective judgements at interval timing durations. We initially deploy our experiments with two different display interfaces: conventional displays (CDs) and virtual reality head-mounted displays (HMDs). Since we find similar (and consistent with previous work) trends in time perception for both display conditions, we favor HMD setups for the subsequent experiments. HMDs offer full immersion, allowing for more realistic visual simulation with computer-generated stereo stimuli (thus depth cues), larger fields of view effects [425], and free viewing with dynamic head motion. They enable full control over the visual stimuli displayed to participants, regardless of whether they remain still or move. Thus, time perception can be studied in more natural conditions while maintaining control over the virtual environment [316, 412, 41, 25]. Besides, time perception is an important factor to consider in VR applications in designing a better human-computer interface. This has been emerging as a critical demand in the medical field, where, for example, the presentation of distracting content through HMDs has proven to shorten the perceived duration of chemotherapy treatments [318]. We believe that time perception manipulations could have a similar effect in training or simulation applications, allowing for longer exposure sessions.

The emotional valence, semantic or other high-level cognitive aspects of the presented content cannot always be manipulated at will without affecting the experience or the goal of a given application. Instead, we propose the manipulation of asemantic visual features to alter time perception without affecting high-level cognitive aspects. In that regard, we study four visual features pertaining to three abstraction levels: low-level spatiotemporal visual properties (*spatial luminance contrast* and *temporal frequency*), mid-level display-related properties (*field of view*), and high-level cognitive aspects (*visual complexity*, understood as the number of independent visual sources in a scene).

In our pre-studies, we replicated an experiment [421] that studied the effect of asemantic visual stimuli at the millisecond timing level as a baseline and found similar results across different viewing conditions (CD and HMD). However, directly extending the millisecond experimental design to longer intervals did not reveal consistent results. We term this phenomenon a *perceptual break*. This perceptual break may be due to the different neural mechanisms that have been found to be in charge of processing different temporal magnitudes in mammals [42, 102] where millisecond timing is processed 'automatically' [189] while interval timing is 'cognitive' (by engaging attention and working memory [190]). A perceptual break can be defined as the point where the perceptual process changes abruptly. In our case, we found that the perceived effect of visual changes on time perception was different depending on the duration of the tested period, particularly when comparing millisecond (less than one second) to interval (multiple seconds) timing.

Following, we present our main experiment (Experiment 1) and additional experiments that suggest how the effect behaves under varied conditions. With *Experiment 1* we tested how differences in the magnitude of luminance contrast, temporal frequency, and field of view affect time perception using 30s intervals. In *Experiment 2* we tested intervals of up to five minutes to check whether the effects found in Experiment 1 hold for longer intervals. Both Experiments 1 and 2 make use of half-duration judgements, in which participants are interrupted part way through a trial and asked to estimate whether more or less than half of the full timing interval has already elapsed (binary response). Finally, we design *Experiment 3* with the aim of providing a quantitative estimate of the temporal distortion, which is based on a numerical estimation of the passage of time, in seconds. See Figure 3.1 for a visual summary of the experimental procedures.

Results across all of our experimental conditions reveal a common trend: duration of time is consistently perceived as shorter when high magnitude levels of each visual feature are present, regardless of viewing conditions or experiment task. We found this trend to hold across intervals ranging from 30 seconds to 3 minutes. In particular, larger spatial luminance contrast, higher temporal frequency, larger FoV, and more complex visual content consistently shorten participants' prospective time judgments when compared with lower levels of the same visual features.

3.2 METHODS

The objectives of our experiments were to analyze the effects of asemantic visual features on prospective time judgements for interval timing (several seconds to minutes range). With that aim, we designed two different tasks (A and B) and carried out four experiments with different sets of participants for a total population size of 168 participants. In the following, we describe each of the experiments in detail. A video showing an accelerated version of the experimental procedure for Tasks A and B can be found online in Movie S1 , while a complete visualization of our results can be found in Figs 3.3-3.6 and Tables 3.2 and 3.3.

Available at `https://doi.org/10.1371/journal.pone.0265591.s001`

3.2.1 *Pre-studies*

Previous work reported that the emotional reactions (level of arousal and valence) experienced by observers while viewing short movie clips were driving factors in time perception, while the viewing conditions (display type) were not [392]. Following the work of van der Ham et al. [392], we ran an initial study to validate whether their findings held by replacing emotional factors with asemantic visual features. With that aim, we first replicated a previous experiment on millisecond timing that measured the effects of visual magnitude on time perception [421] with simple 2D stimuli of abstract patterns (Fig 3.2 A), and compared the results obtained on a conventional display, or CD (first viewing condition) with those obtained on an HMD (second viewing condition). This study was completed by five participants who observed both viewing conditions. Participants judged the duration of pairs of "small" and "large" stimuli, presented for 600-937ms using a pairwise forced-choice comparison. The stimuli were simple and abstract. Following Xuan et al.'s notation, "small"/"large" indicates the level of the visual features in the stimuli. For example, a square with low luminance is referred to as "small", while a square with high luminance is referred to as "large". Size, numerosity, and digits were similarly categorized.

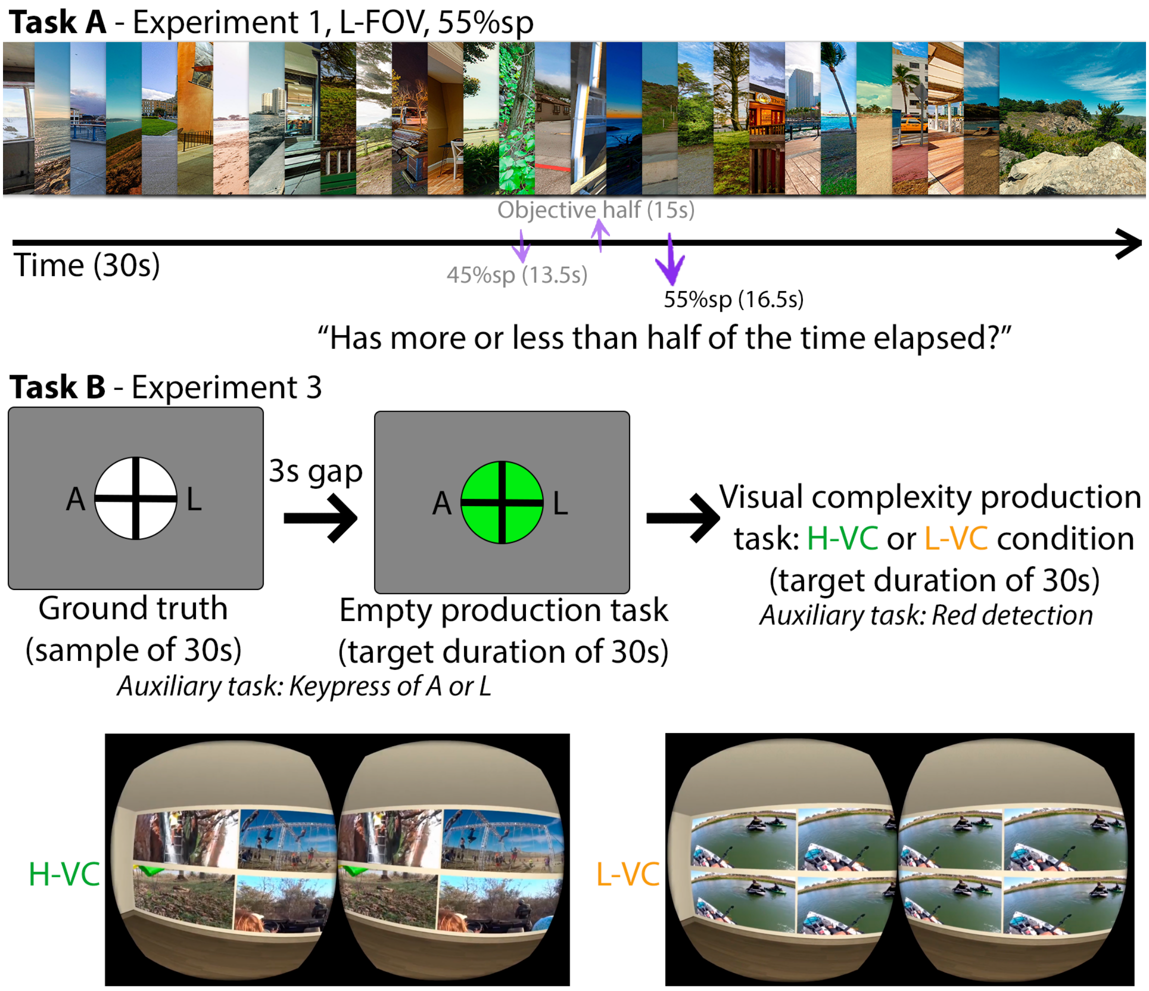

Task A - Experiment 1, L-FOV, 55%sp
Objective half (15s)
Time (30s)
45%sp (13.5s)
55%sp (16.5s)
"Has more or less than half of the time elapsed?"
Task B - Experiment 3
A
L
3s gap
A
L
Visual complexity production task: H-VC or L-VC condition (target duration of 30s)
Auxiliary task: Red detection
Ground truth (sample of 30s)
Empty production task (target duration of 30s)
Auxiliary task: Keypress of A or L
H-VC
L-VC

Figure 3.1: **Task A)** Binary task used in Experiments 1-2. Illustrative example for a single trial (L-FOV condition in Experiment 1 at the 55% sp, see Section 2.1). During a trial, participants indicated whether "more than half" or "less than half" of the duration had passed, at a temporal sampling point that was 55% of the total duration (16.5s). **Task B)** The double production task used in Experiment 3. First participants were presented with an empty scene for 30s (ground truth) while performing an auxiliary task (detecting the presence of A or L on screen with a corresponding keypress). Then they were asked to press a key after 30s had elapsed (empty production task). At this point, the empty scene was replaced with a scene in the H-VC or L-VC condition (high or low visual complexity, respectively). After the change of scene, participants had to press a key when 30 more seconds had elapsed (visual complexity production task). In the H-VC condition, the four screens of the virtual room displayed different videos at the same time. In the L-VC condition, the four screens presented the same video synchronously.

As in Xuan et al.'s work [421], we denote as *incongruent* short durations with “large” stimuli, and correspondingly, long durations with “small” stimuli, since their temporal and visual magnitudes do not match. According to the results reported in the paper, participants were inclined to judge larger stimuli as longer, regardless of their actual duration: in CDs (first viewing condition) the mean difference in error rates was 20% on average between incongruent and congruent trials, with incongruent trials having higher error rates.

We then extended Xuan et al.'s work by using an HMD (second viewing condition) and observed the same trend. We found a mean difference in error rates of 37.6% between congruent and incongruent stimuli across the five participants, with higher error rates in incongruent trials, consistent with the results trend that was found in the original work using CDs. These findings are consistent with those of van der Ham et al. [392], indicating that similar results can be obtained using HMDs and CDs. This, together with previous work, is indicative that the trend of the effect of visual stimuli magnitude on time perception should hold between HMDs and CDs.

In the following, we tried to directly extend the same experimental design to interval timing. However, our follow-up attempts to extend the work of Xuan et al. [421] to longer interval durations (5s, 10s, and 30s) did not exhibit the same effect (the mean difference in error rates between congruent and incongruent stimuli dropped to 5.2%). This trend was consistent with the *perceptual break* between millisecond and interval duration magnitudes evidenced in previous works [42]. Debriefing sessions also revealed participants' loss of engagement while observing simple stimuli for long periods, which might have contributed to the small differences observed in error rates between congruent and incongruent stimuli. We proceeded to design a more realistic scenario with 3D computer-generated models of real objects (specifically, lamps) instead of the original 2D stimuli of simple, abstract patterns. In this extended experiment, we again observed the same trends at millisecond intervals but did not find significant differences at longer durations (6.4% mean difference in error rate). However, participants suggested on post-study free written form questionnaires that they preferred the more realistic stimuli. This motivated us to design our main experiment (Experiment 1) and subsequent studies leveraging more complex and realistic stimuli (Fig 3.2 B) to maintain participant engagement throughout the experiments and attention in the timing tasks.

3.2.2 *Interval timing experiments*

In the pre-studies, we found that the effects of the absolute visual magnitude (i.e., “big” vs “small” stimuli in Xuan et al.'s experiments [421]) observed with millisecond timing did not seem to replicate when extended to interval durations, possibly due to the aforementioned perceptual break [42]. We thus design our interval timing experiments based on this knowledge, as well as inspired by Montague's theory of time perception [245]:
If [...] perceptual space and time magnitudes are essentially relative matters [...], the amount of objective time or change which appears to be present at any one moment will be measured by its ratio to the subjective change which accompanies it". In other words, it is not the absolute visual value that affects time perception, but rather the perceived changes across visual stimuli.

In our experiments, we study how changes in the *magnitude* of four different asemantic visual features affect time perception, ranging from lower to higher levels of abstraction: luminance contrast and temporal frequency (low-level), field of view (mid-level), and visual complexity (high-level). Our stimuli of choice are quickly-varying frames (in the case of both static images and short video sequences), the design choice motivated by our pre-studies suggesting that a reasonable amount of variation was required to maintain user engagement, critical to obtaining a reliable measure of time perception. Further, in potential application scenarios, it is likely that, at the interval timing durations we are considering, variation in the visual input will naturally be present.

Extending the notions of “big” and “small” from Xuan's studies, in all our experiments we similarly have two *magnitude* levels: *high* and *low* (see Fig 3.2). Regardless of the viewing conditions (different display types or visual features being manipulated) throughout our experiments, the *high* level is designed to exhibit a larger absolute value of the visual feature than the *low* level.

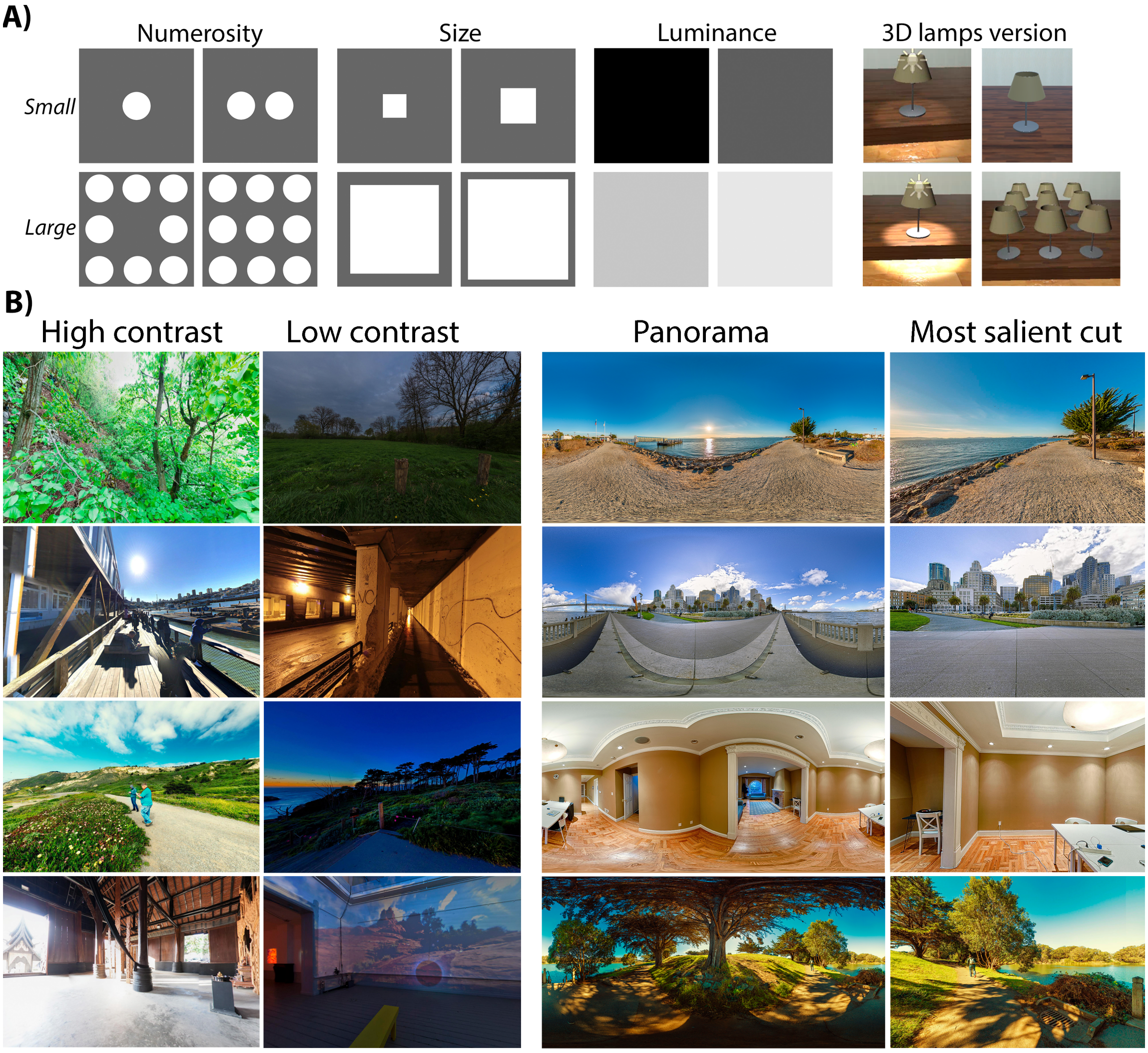


Figure 3.2: **A)** Stimuli used in the pre-studies. *Left:* The 2D, grayscale stimuli we adapted from Xuan et al [421]. *Right:* The 3D lamps used in our follow-up attempts to directly extend the procedure of Xuan et al. to interval timing. **B)** Illustrative examples of the stimuli used in our experiments depicting natural scenes. *From left to right, in each column:* high contrast (H-CON), low contrast (L-CON), 360° panoramas (H-FOV), and their associated most salient crops (L-FOV). Fig 3.1 (Task B) also shows an example of high and low visual complexity (H-VC and L-VC).

Acronym	Meaning
CD	Conventional displays
HMD	Head mounted displays
VR	Virtual reality
FRQ	Frequency
CON	Contrast
FOV	Field of view
H-CON	High magnitude level of contrast
L-CON	Low magnitude level of contrast
HFM	Histogram flatness measure [383]
H-FRQ	High magnitude level of frequency
L-FRQ	Low magnitude level of frequency
H-FOV	High magnitude level of FOV
L-FOV	Low magnitude level of FOV
sp	Sampling point
pover,45	Percentage of "more than half" responses in the 45% sp (overestimation indicator)
punder,55	Percentage of "less than half" responses in the 55% sp (underestimation indicator)
pcorr,XX	Percentage of correct answers in the XX% sp (accuracy indicator)
VC	Visual complexity
H-VC	High magnitude level of visual complexity
L-VC	Low magnitude level of visual complexity

Table 3.1: List of acronyms used through this chapter.

The changes elicited by the visual feature will also be generally perceived as larger in the high level than in the low level. We define visual changes as spatiotemporal per-pixel variations. In this sense, we assume that a sequence of stimuli with higher luminance contrast (larger per-pixel variations), temporal frequency (faster variations), field of view and visual complexity (both with more numerous variations) will all cause larger visual changes than their lower-level counterparts.

In the following, we explain each of the experiments in detail. Our results can be found in Figs 3.3-3.6 and Tables 3.2 and 3.3. Table 3.1 presents a list of the acronyms used through this chapter in order of occurrence. Our main experiment (Experiment 1) explores how different magnitudes of luminance contrast, temporal frequency and field of view affect half-duration judgements under a fixed viewing condition (static panoramas viewed in HMDs). We conducted a follow-up replication in conventional displays which show the same trend found in the work of Van der Ham et al. [392]. Experiment 2 extends the studies from 30s to up to five minutes. Finally, Experiment 3 uses a different task design (duration judgements) to give an estimate of the extent to which temporal perception gets distorted with changes in the visual complexity of the stimuli.

3.2.2.1 *Experiment 1 – Frequency, contrast, and field of view in HMDs*

In *Experiment 1* we study how variations in the magnitude of the asemantic visual features *temporal frequency* (FRQ), *luminance contrast* (CON), and *field of view* (FoV) affect time perception while watching static panoramic imagery in head-mounted displays (HMDs). All the trials tested in this experiment had a duration of 30s. We denote as a *trial* any separable part of an experiment associated with an answer from the participant. *Experiment 1* was carried out using the study design we labeled *Task A*, which evaluates how changes in the magnitude of asemantic visual features affect time estimation, for a fixed viewing condition. Participants were first exposed to a sample trial duration to set the expectations. On each subsequent trial, participants were interrupted approximately halfway through the duration and asked to make a judgement about whether more or less than half of the full trial duration had elapsed [411]. We term these binary temporal judgements *half-duration judgements*. The binary task design has two advantages: first, we wanted participants to give us quick, intuitive answers. Second, compared to asking participants to directly estimate magnitudes (i.e., how much time has passed) binary judgements are less noisy and variable, at the expense of less information. The goal of *Task A* is different from the millisecond timing experiment design presented in the pre-studies where participants compared the duration of *big vs small* stimuli in a single trial. Our goal with the pre-studies was to replicate previous work in a different viewing condition (in this case, an HMD). For *Task A* we have tried to avoid memory-related confounding factors by presenting one single manipulation of a given visual feature in each trial.

Participants and apparatus. The stimuli were presented on an HTC Vive Focus, a portable HMD (2880x1600 spatial resolution, 75Hz, 110 visual degrees FoV). A total of 89 participants took part in *Experiment 1*. To avoid excessively large numbers of trials per participant leading to fatigue or learning effects, we split the visual features tested and participants into two groups. Group 1.1 consisted of 45 participants (20 female, mean age 28.7 years) who experienced *temporal frequency x luminance contrast* (low-level visual features) changes. Group 1.2 consisted of a different set of 44 participants (18 female, mean age 26.8 years) who experienced *FoV* (mid-level property) changes. All participants had normal or corrected-to-normal vision. All were naive to the purpose and hypothesis of the study. These two affirmations hold for all the presented experiments.

Stimuli. We gathered a set of 420 panoramas from open-source web databases (Flickr, Unsplash, and Pixexid), all of them depicting natural indoor or outdoor scenes (see Fig 2). They were manually selected, excluding synthetic, cartoon, high-emotional valence (violence, parties, dramatic pictures, etc.) and written content. The stimuli were randomly arranged in sequences of 30 panoramas to compose each 30s experiment trial. Each image was shown for a total of 1.2s-1.8s, with a fade-in/fade-out effect of 0.5s between images, during which the images overlapped in order to create smooth transitions. Both the fading effect and variable image presentation durations were implemented to avoid the kind of predictable regularity that may facilitate counting. Instead of being presented for 1 second each, images in the sequence cycled through the following duration pattern: 0.4s, 0.2s, 0.4s, 0.6s, 0.8s, 0.6s. Including the additional 0.5s transition between images, a group of six images had a total duration of 6 seconds. Five of these groups were sequenced to create the 30s trials.

Luminance contrast. Images or video sequences with a high *magnitude* level of contrast (H-CON) produce larger visual changes than those with a low *magnitude* level of contrast (L-CON). A representative sample of these stimuli can be found in Fig 3.2. We calculated the contrast of our 420 panoramas taking into consideration their most salient 2D crop spanning 45 by 65 degrees of visual angle [353], and using the histogram flatness measure (HFM) [383]. The computed contrast range of our full stimuli set was [0.17, 0.72]. We selected the 120 panoramas with least contrast with an L-CON in the range [0.17, 0.60], and 120 panoramas with the highest contrast with an H-CON in the range [0.60, 0.72]. The remaining 180 panoramas were used for the temporal frequency conditions.

Temporal frequency. We created the high and low temporal frequency conditions (H-FRQ and L-FRQ, respectively) out of the same stimulus set, to manipulate temporal frequency without changing the amount of visual information presented. Specifically, starting with the stimuli in the L-FRQ condition, we inserted three black frames per second to the entire image sequence to create a flickering effect in the H-FRQ condition.

Field of view. To reduce the differences in conditions to only the main factor under investigation, we similarly used the same stimulus set for both high and low FoV conditions (H-FOV and L-FOV, respectively). In this case, we used panorama images that were either shown in full size in the H-FOV condition, or were reduced to their most salient [353] crops (2D re-projected regions of 45 x 65 degrees of visual angle, Fig 3.2) in the L-FOV condition. H-FOV was designed to show visual changes in a larger portion of the visual field compared with L-FOV. Note that the entirety of Experiment 1 was carried out in HMDs to take advantage of the expanded field of view available.

Procedure. The experimental procedure was the same for both Groups 1.1 and 1.2, using half-duration judgements for measuring the perception of time (Task A, Fig 1). Before beginning *Experiment 1*, participants were explicitly told not to count to estimate time passage. At the beginning of the experiment, they were informed of the duration of each of the trials (30s), and provided with a 30s sample trial, which was interrupted with on-screen message to denote when half of the duration (15s) had elapsed. After the sample trial, the experiment proceeded with a set of consecutive 30s trials. Within each trial, participants would be prompted to make a half-duration judgement at a given (unknown to participants) sampling point (sp) of 45% or 55% of the total trial duration (i.e., at 13.5s or 16.5s in a 30s interval): ***Has more or less than half of the time elapsed?*** Participants answered using the HMD controllers. They were instructed to answer as fast as possible, and the trial continued automatically after their response. At the end of each trial, an additional binary question appeared: ***Do you think your previous answer was correct?*** Participants again answered using the HMD controllers. Participants in Group 1.1 completed a total of 32 trials, for a total experiment duration of approximately 20 minutes. Participants were asked to make a half-duration judgement at a sampling point of 45% for half the trials, and at a sampling point of 55% for the other half. The trials were randomly ordered for each participant. Across trials for a given participant the magnitude level (high vs low) and the feature chosen (contrast vs frequency) could vary. However, within any particular trial, all the stimuli were consistent in magnitude and feature manipulation. As a reminder, the goal was to accumulate the effects of a particular manipulation over the full duration of a trial (in this case 30s) to evaluate how the manipulation affects time perception at 30s intervals. Similar to Group 1.1, participants in Group 1.2 completed a total of 8 trials, for a total experiment duration of approximately 5 minutes with the difference being a manipulation in the sampling point and magnitude of the FoV condition for this group.

Statistical Analysis. A 2 magnitude levels (high vs low)x2 visual features (contrast vs frequency)x2 sampling points (45% vs 55%) ANOVA was used to check for significant differences for the 45 participants (Group 1.1) who experienced luminance contrast and temporal frequency manipulations. A 2 (FOV levels: high vs low) x2 (sampling points) ANOVA was used to check for significant differences for the 44 participants (Group 1.2) who experienced the FoV manipulation. The answer variable was binary ("more than half" or "less than half" of the time elapsed) in both analyses. *Post hoc* analyses for this experiment can be found in the Appendix 3.A. All statistical analyses were carried out using Matlab. ANOVA post-hoc power was calculated with additional scripts [384]. Effect sizes were calculated using Harald's Toolbox for Matlab [132] (partial eta squared for ANOVA). Additionally, we analyzed Groups 1.1 and 1.2 together (89 participants) with a GLMM to check for interactions of the different visual features for completeness. The information of the GLMM analysis can be found in the Appendix 3.B.

Results. We measure and analyze: (i) the percentage of "less than half" responses in the 55% sp with respect to the total number of responses for this sampling point, which is an indicator that time felt shorter (or was underestimated) according to the half-duration judgement ($p_{under,55}$); and (ii) the percentage of "more than half" responses in the 45% sp with respect to the total number of responses for such sp ($p_{over,45}$), an indicator that time felt longer (or was overestimated). These two values are complementary to the percentage of correct responses ($p_{corr,45}$ and $p_{corr,55}$) with respect to the total number of responses, such that, for any given condition, $p_{corr,45} + p_{over,45} = 100\%$ and $p_{corr,55} + p_{under,55} = 100\%$. In Experiment 1, underestimation was more common for H-CON while overestimation was more common for L-CON. Analogously, underestimation was more common for H-FRQ and H-FOV and overestimation was more common for L-FRQ and L-FOV (see Table 3.1 and Fig 3.3). With a significance level established at p=0.05 and power of 0.895 and 0.894, both ANOVAs revealed that magnitude had a significant effect on the answers (F=45.03, p<0.001, partial $\eta 2 = 0.580$ for the three-way ANOVA, F=12.39, p<0.001, partial $\eta 2 = 0.228$ for the two-way ANOVA), while the sampling point (F=0.91, p=0.340, partial $\eta 2 = 0.012$ for the three-way ANOVA; F=0.49, p=0.482, partial $\eta 2 = 0.009$ for the two-way ANOVA) and visual features (F<0.01, p=0.964, partial $\eta 2 < 0.001$ only tested in the three-way ANOVA for frequency and contrast)

Factor	Sp	% of correct L/H ($p_{corr,45}$ or $p_{corr,55}$)	% of underestimated L/H ($p_{under,55}$)	% of overestimated L/H ($p_{over,45}$)
CON	45 (13.5s)	60.5(L)/65.8(H)	-	39.5(L)/34.2(H)
	55 (16.5s)	57.9(L)/42.8(H)	42.1(L)/57.2(H)	-
FRQ	45 (13.5s)	52.6(L)/73.7(H)	-	47.4(L)/26.3(H)
	55 (16.5s)	58.2(L)/48(H)	41.8(L)/52(H)	-
FOV	45 (13.5s)	62.8(L)/68.5(H)	-	37.2(L)/31.5(H)
	55 (16.5s)	73.7(L)/63.2(H)	26.3(L)/36.8(H)	-

Table 3.2: Results of Experiment 1. Accuracy (% of correct answers) is higher for high magnitude levels of contrast, frequency and FoV at 45% sp ("less than half" answers), while at 55% ("more than half" answers) it is always higher for low magnitude levels of contrast and frequency. At 45% sp, overestimation (incorrect responses, "more than half" answers) occurs more frequently for low magnitude levels. At 55% sp, underestimation (incorrect responses, "less than half" answers) occurs more frequently at high magnitude levels.

did not. The additional GLMM analysis yielded consistent results with the separated ANOVAs, with only the magnitude factor having a significant effect on the response variable (t=-3.7641, CI {-1.142, -0.360}, p<0.001). We observe a similar effect for high magnitude levels across the three visual features: H-CON, H-FRQ, and H-FOV all make time seem shorter compared to their low magnitude counterparts. Moreover, participants were fairly confident in their answers, as inferred from the binary responses at the end of each trial, where participants indicated if they agreed with their half-duration judgement in retrospect, after the whole interval had elapsed. We define a trial as rectified if participants believe at the end of a trial that their previous answer was wrong. Experiment 1 had a rectification percentage of less than 5%.

3.2.2.2 *Follow-up replication in Conventional Displays*

The work of Van der Ham et al. [392] suggests that similar half-duration judgements can be elicited with HMDs and conventional displays (CDs) if viewing conditions are similar. Our pre-studies follow the same trend. In Experiment 1, we addressed *interval* timing, with experiments done on HMDs only. Thus, we ran a follow-up study to replicate part of Experiment 1 on CDs. We replicated Experiment 1, but focused our attention on the two low-level visual features that should ideally generalize across display types: luminance contrast (CON) and temporal frequency (FRQ).

Participants and Apparatus. The stimuli were presented on a Samsung display (S24F350FHU, 1920x1080 spatial resolution, 60Hz) at a distance of 60cm from the viewer. Seven participants completed the follow-up study (3 female, mean age 22.4 years).

Stimuli. The stimuli used in this study consisted of a set of 420 images of natural indoor and outdoor scenes from open-source web databases with a CC license (Flickr, Unsplash, and Pixexid) depicting natural indoor and outdoor scenes. These were selected following the same exclusion criteria as in Experiment 1. The presentation of the images, duration, fade-in/fade-out effects, etc., were the same as in Experiment 1, yielding test trials that included 30 images and were 30s long. Participants completed a total of 16 trials. The duration of the experiment was 10 minutes.

Luminance contrast and temporal frequency. The different magnitude levels of CON and FRQ were achieved following the same procedure as in *Experiment 1*. The computed contrast range (HFM) of our full stimuli set was [0.55, 0.84]. We selected the 120 images with least contrast with an L-CON in the range [0.55, 0.64], and 120 images with the highest contrast with an H-CON in the range [0.71, 0.84]. The remaining 180 images were used for the temporal frequency conditions.

Procedure. This study was carried out following the same procedure described in Experiment 1, where participants indicated whether more or less than half of the time had elapsed, with the sole difference that instead of wearing an HMD and using its controllers as the input device, in this version of the study participants were entering their responses on a keyboard while looking

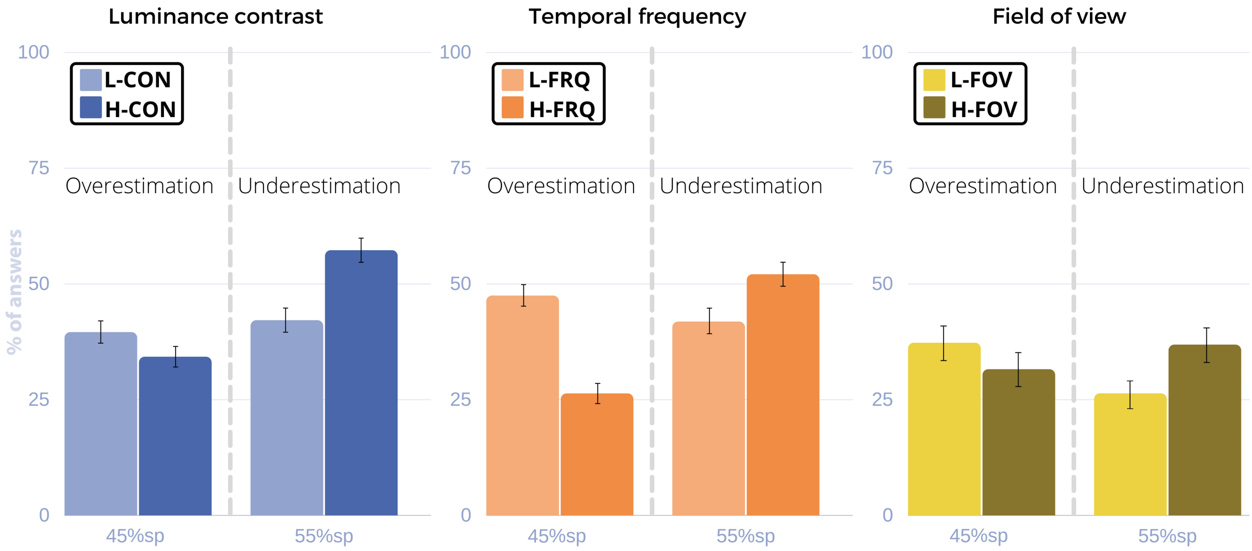


Figure 3.3: **Results for Experiment 1.** *From left to right*: aggregation of trials by luminance contrast, temporal frequency and field of view conditions. Y axis: % of answers. X axis: 45% and 55% sampling points (sp). The displayed bars of each graph correspond to $p_{over,45}$ at the 45% sp (overestimation) and to $p_{under,55}$ at the 55% sp (underestimation). The percentage of correct answers for each case is complementary to the displayed value in the graph ($p_{corr,45} + p_{over,45} = 100\%$ and $p_{corr,55} + p_{under,55} = 100\%$). Note that overestimation is always more frequent in low magnitude levels and underestimation in high magnitude levels, regardless of the visual feature.

at the CD. In this study, each participant experienced 8 conditions: 2 magnitude levels (high vs low)x2 visual features (contrast vs frequency)x2 sampling points (45% vs 55%).

Statistical Analysis. A 2x2x2 ANOVA was used to check for significant effects (with *visual feature*, *magnitude* and *sampling point* as factors). The answer variable was binary ("more than half" or "less than half" of the time elapsed).*Post hoc* analyses for this experiment can be found in the Appendix 3.A.

Results. Using the same measures described in *Experiment 1*, we found similar tendencies in time perception between HMDs and CDs. Table 3.3 and Fig 3.4 show the results of this study. The underestimation and overestimation trends follow those found in *Experiment 1*: H-CON and H-FRQ elicited higher underestimation rates while overestimation was more frequent for L-CON and L-FRQ. However, while the trend observed is the same, no significant effect was found for any of the tested factors (*magnitude* F=0.56, p=0.454; *visual feature* F=0.56, p=0.454; *sampling point* F=1.27, p=0.263).

3.2.2.3 *Experiment 2 – Extended durations up to five minutes*

To further analyze the stability of the observed time compression effect over extended periods, we tested temporal frequency (FRQ) effects at trials of longer duration. We only considered temporal frequency as a visual feature in Experiment 2 to keep the size of the experiment tractable, since we found no significant differences between the tested *visual features* in *Experiment 1*. To maintain participant engagement and for ecological validity at longer durations, *Experiment 2* employed sequences of short video clips arranged into "movies", instead of static images. The trials tested in *Experiment 2* varied in duration, including: 30s, one minute, three minutes, and five minutes.

Participants and apparatus. In the follow-up replication study we verified that similar results could be found both in HMDs and CDs when using analogous experimental set ups. Given that Experiment 2 deals with longer temporal durations, to avoid a potential confounding effect of fatigue caused by prolonged exposure [49] in HMDs, we use CDs. The stimuli were presented on a Samsung display (S24F350FHU, 1920x1080 spatial resolution, 60Hz), at a distance of 60cm from the viewer. 51 participants took part in *Experiment 2* (20 female, mean age 22.9 years).

Factor	Sp	% of correct L/H ($p_{corr,45}$ or $p_{corr,55}$)	% of underestimated L/H ($p_{under,55}$)	% of overestimated L/H ($p_{over,45}$)
CON	45 (13.5s)	58.3(L)/66.6(H)	-	41.7(L)/33.4(H)
	55 (16.5s)	62.9(L)/42.7(H)	37.1(L)/57.3(H)	-
FRQ	45 (13.5s)	52.9(L)/67.8(H)	-	47.1(L)/32.2(H)
	55 (16.5s)	61.7(L)/32.6(H)	38.3(L)/67.4(H)	-

Table 3.3: Results of the follow-up replication study with conventional displays. Accuracy (% of correct answers) is higher for high magnitude levels of contrast and frequency at 45% sp ("less than half" answers), while at 55% ("more than half" answers) it is always higher for low magnitude levels of contrast and frequency. At 45% sp, overestimation (incorrect responses, "more than half" answers) occurs more frequently for low magnitude levels. At 55% sp, underestimation (incorrect responses, "less than half" answers) occurs more frequently at high magnitude levels.

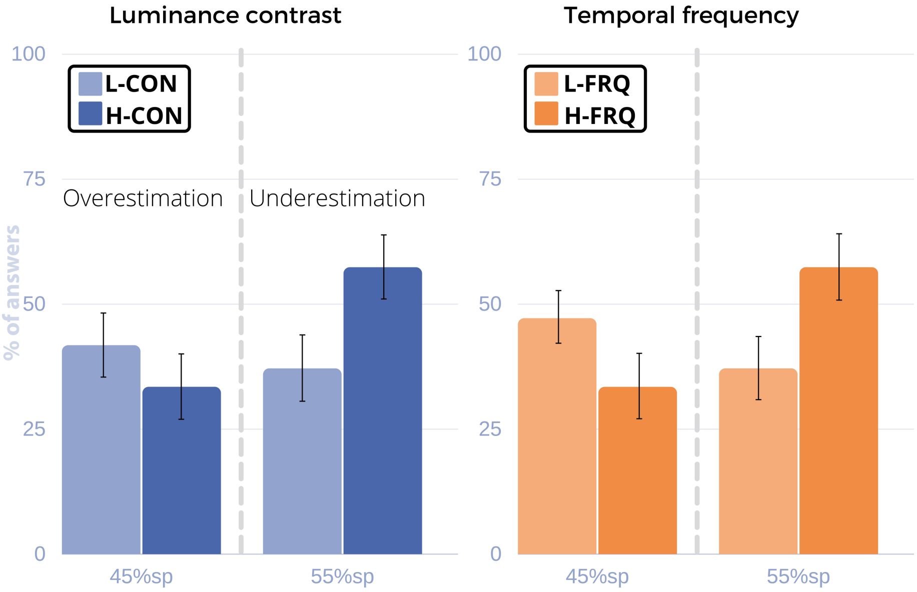


Figure 3.4: **Results for the follow-up replication study with conventional displays.** *Left*: Luminance contrast trials. *Right*: Temporal frequency trials. Y axis: % of answers. X axis: 45% and 55% sampling points (sp). The displayed bars of each graph correspond to $p_{over,45}$ at the 45% sp (overestimation) and to $p_{under,55}$ at the 55% sp (underestimation). The percentage of correct answers for each case is complementary to the displayed value in the graph ($p_{corr,45} + p_{over,45} = 100\%$ and $p_{corr,55} + p_{under,55} = 100\%$). Like in Experiment 1, overestimation is always more frequent in low magnitude levels and underestimation in high magnitude levels, regardless of the visual feature.

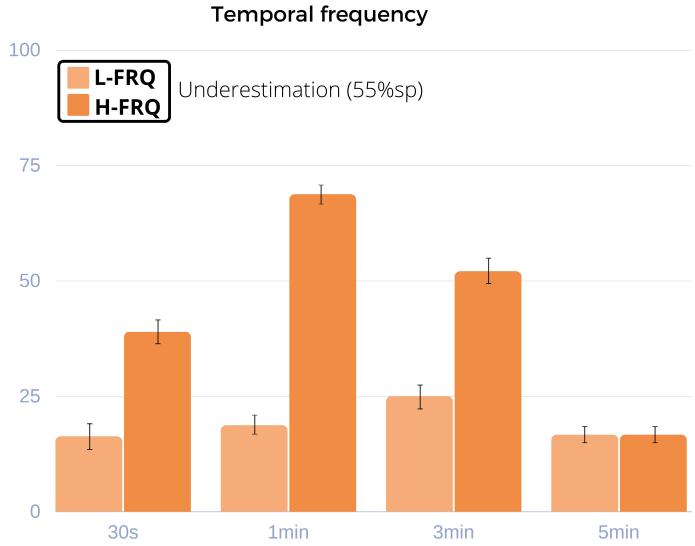


Figure 3.5: **Results for Experiment 2.** Y axis: % of answers. X axis: Temporal duration of the trials (30s, 1min, 3min, 5min). The displayed bars correspond to $p_{under,55}$ at the 55% sp (underestimation). The percentage of correct answers for each case is complementary to the displayed value in the graph ($p_{corr,55} + p_{under,55} = 100\%$). Like in Experiment 1, underestimation is more frequent in high magnitude levels, in this case in trials with a duration of up to 3minutes.

Stimuli. The videos used in *Experiment 2* consisted of 700 three-second video clips from the Moments dataset [244]. The set of videos was manually curated to exclude high arousal actions, synthetic content, text, or manipulated playback speed. Randomly ordered sequences of videos were arranged to form each trial. Since video clips had a fixed duration of 3s, the number of video clips per trial was a function of the length of the trial (10 clips for 30s trials, 20 clips for one minute trials, 60 clips for three minute trials and 100 clips for five minute trials). No transition effects were applied between different clips inside a given trial. Videos were played as continuous "movies" composed of back-to-back 3s clips.

Temporal frequency. To induce a high temporal frequency in sequences of videos, we subdivided each 3s clip into 1s cuts that were then reshuffled, effectively increasing temporal changes in visual content without changing the totality of visual information presented.

Procedure. *Experiment 2* was carried out using half-duration judgements for measuring the perception of time, following the procedure outlined in *Experiment 1*. Participants were randomly assigned to one of the four possible duration conditions, and completed all the trials with the same presentation duration to avoid bias effects due to differences in trial durations. For simplicity, we only used the 55% sampling point (prompting participants to make temporal judgements after 16.5s elapsed within 30s trials, after 33s for one minute trials, 99s for three minute trials and 165s for five minute trials), which means each duration had only two possible conditions (2 magnitude levels x 1 sampling point x 1 visual feature). Each participant completed 6 trials, for a total duration between 3 minutes (in the case of 30s trials) and 30 minutes (for five minute trials).

Statistical Analysis. Each of the sampled trial durations was analyzed separately, testing for significant differences in high vs low *magnitude* levels of FRQ with Chi-square proportions tests. The answer variable was binary, as in *Experiment 1*.

Results. With a significance level established at p=0.05 the duration of H-FRQ stimuli was significantly underestimated for trial durations up to three minutes (higher punder,55 for H-FRQ, see Fig 3.5): 16.3% of trials in L-FRQ condition vs. 38.9% in H-FRQ at 30s (χ^2=13.27, p=0.001, post hoc power=0.83, ES=0.586); 18.7% L-FRQ vs 68.7% H-FRQ at one-minute (χ^2=50.99, p=0.001, post hoc power=0.99, ES=0.635); 25% L-FRQ vs. 52% H-FRQ at three-minute trials (χ^2=15.39, p=0.001, post hoc power=0.88, ES=0.579). At five-minute trials, however, the effect was no longer present: 16.6% L-FRQ vs. 16.6% H-FRQ (χ^2=0, p=1).

3.2.2.4 *Experiment 3 – Quantification of the perceived temporal distortion*

All the previous experiments in this work used half-duration judgements, collected as binary responses (*Task A* study design) to evaluate the effects of changes to visual features on the perception of time. Experiments 1-2 confirm that the perception of time can indeed be distorted by manipulating the magnitude of a visual feature (e.g., high vs low contrast, frequency, etc.). *Experiment 3* was then designed to give an estimate of the extent to which temporal perception gets distorted, by using traditional duration judgements. In this case, rather than making a binary half-duration judgement, participants produce a numerical estimate of the elapsed duration. In *Experiment 1* we found that only differences in *magnitude* of the different visual features had a significant effect on time perception. However, we did not find any significant difference between the three tested visual features: frequency, contrast and field of view. Since all the features were equally effective, for *Experiment 3* we focused on a fourth, more abstract, feature: visual complexity. We define *visual complexity* as the number of distinct sources of visual content inside a given scene. Visual complexity was chosen as a proxy of real-world tasks that require simultaneous attention to multiple screens or sources of content. Visual complexity uses the principle common to all the previously tested features, whereby high magnitude levels of the feature trigger larger spatiotemporal per-pixel variations than their lower magnitude counterpart. In summary, in *Experiment 3* we estimate how much magnitude variations of *visual complexity* (VC) affect time perception. All the trials tested in this experiment had a target duration of 30s.

Participants and Apparatus. The stimuli were presented on an Oculus Rift CV1 HMD (2160x1200 spatial resolution, 90Hz, 110 visual degrees FoV). Eleven participants took part in Experiment 3 (5 female, mean age 25.2 years).

Stimuli. The same set of videos described in *Experiment 2* were used in this experiment. In the low visual complexity (L-VC) condition, four screens inside the virtual scene displayed the same identical video simultaneously. In the high visual complexity (H-VC) condition, each screen displayed a different video, effectively augmenting the sources of visual information inside the scene, as well as the spatial changes of the visual stimuli.

Procedure. *Experiment 3* was carried out using *Task B*: a double production task [6]. *Task B* was designed to study the magnitude of time estimation errors under larger visual changes (Fig 3.1). At the beginning of the experiment, participants were explicitly instructed not to count to get a feeling of time passing, like in *Task A*. Participants were first exposed to an *empty* virtual scene for 30s: a gray background and a fixation cross surrounded by a lighter gray circle of two degrees of visual angle situated in the middle. During these 30s, an **auxiliary task** was performed in order to prevent participants from counting. Auxiliary tasks help maintain engagement without significantly increasing cognitive load. The letters "A" and "L" would appear at a fixed position to the left or right of the fixation cross, respectively. When participants saw one of those letters in the scene, they had to press the corresponding key on a keyboard as fast as possible. A/L appeared at random intervals of 0.5-2s, 1.5° visual angle to the left or right of the fixation cross, and with a vertical size of 100 pixels. When the 30s had elapsed, and after a three-second gap, the circle around the fixation cross turned green and the **empty production** subtask started: participants had to press a key to indicate when 30 more seconds had elapsed, while continuing with the auxiliary task. After completing the empty production task, participants moved on to the next task after a keypress. Empty production tasks were used to measure individual baseline estimation in the absence of a stimulus. Participants were then shown a virtual empty room with four flat screens displaying videos on one of its walls. The **visual complexity production** subtask started: across both H-VC and L-VC conditions, participants had to indicate when 30s had elapsed by pressing a key. At the same time, they had to complete a new auxiliary task: pressing a key when something red appeared in any of the videos. Like in the first auxiliary task, the red detection task was designed with the intention of increasing engagement through the experiment and preventing users from explicitly counting. Immediately after the visual complexity production, participants completed a NASA-TLX questionnaire [126] to measure the cognitive load elicited by the visual complexity production subtask. We used a within-subjects design: each participant completed this experimental sequence twice, once with H-VC, once with L-VC, but with a different set of stimuli.

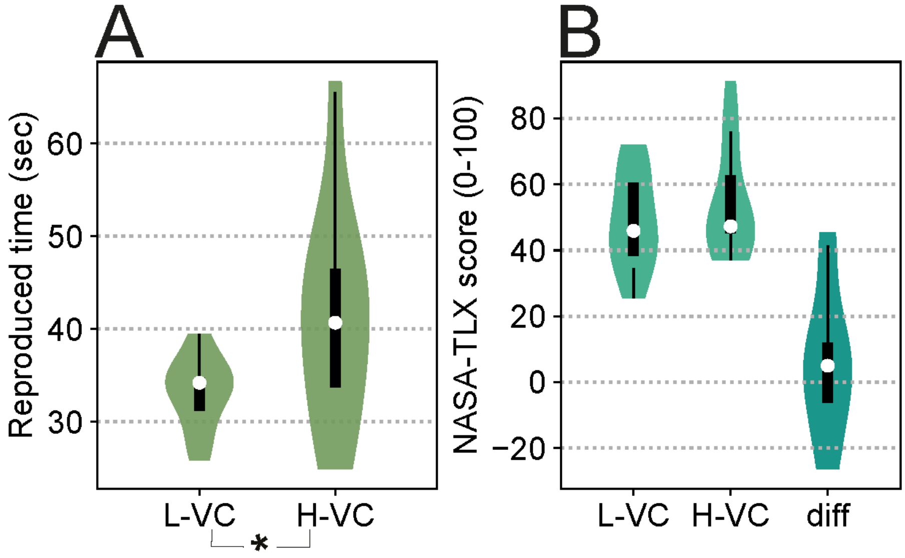


Figure 3.6: **Experiment 3 results.** **A**: Participants take longer to realize 30s have passed in H-VC, which suggests that time is compressed in the presence of larger perceived visual changes. **B**: NASA-TLX score (workload measure) for the visual complexity production tasks of each level. Right violin plot ("diff") shows individual differences in NASA-TLX score (L-VC MINUS H-VC). Note that the negative values in the diff violin plot mean that for some users L-VC was experienced as more demanding than H-VC. Significant differences are marked with an asterisk.

The order of the conditions (H-VC and L-VC) was randomized to avoid ordering effects. Finally, each participant completed an additional empty production subtask, for a total of three empty productions, from which a robust individual measure of production accuracy was obtained by averaging.

Statistical analysis. An ANOVA was carried out to compare differences in *magnitude* for H-VC and L-VC levels. NASA-TLX questionnaire differences were tested with a t-test. The answer variables were continuous (for the produced durations) or discrete (for the NASA-TLX scores). The independent variable was binary (for low and high magnitude levels of visual complexity). *Post hoc* analyses for this experiment can be found in the Appendix 3.A.

Results. The mean ratio of empty productions to ground truth duration (target duration of 30s) was 1.12, which suggests a small overestimation of time in the baseline condition. To account for individual differences in time estimation, we computed a ratio between the visual complexity production and the empty production on a per-participant basis. We fixed the empty productions as the baselines for each participant. Mean accuracy in the auxiliary task associated with the empty production (A/L detection) was 99.43%. Mean accuracy (correct keypresses divided by the total number of keypresses) in the second auxiliary task (red detection) was 99.13% (true positive rate) with only 4.15% of detection failures (false positive rate, i.e., cases in which a red object appeared on screen but there was no keypress). The mean response time was 0.416s for the auxiliary tasks. Participants took more time to indicate that 30s had passed in the H-VC condition, suggesting that time was perceived as significantly shorter under higher visual complexity (ratio of 1.38 for H-VC vs 1.10 for L-VC, power=0.407, F=4.98, p=0.0372, partial $\eta 2$=0.24, normality of distribution checked with Anderson-Darling tests). In other words, participants took, on average, 25.4% longer to perceive that 30s had passed in H-VC than in L-VC. Fig 6 illustrates this difference.

Additionally, participants completed a NASA-TLX questionnaire after each visual complexity production to provide a measure of cognitive load differences between the H-VC and L-VC condition levels. There were no significant differences in perceived workload between H-VC and L-VC productions (54.85 mean score for H-VC, 48.83 for L-VC, normal distribution checked with Anderson-Darling tests, t-test t(10)=0.8857 p=0.397, Fig 3.6).

3.3 DISCUSSION

Our series of experiments reveal consistent trends in how asemantic visual features directly alter half-duration judgements on prospective interval timing paradigms. Our results suggest that the perception of time is compressed in the presence of larger spatiotemporal visual changes (elicited in high *magnitude* levels of each visual feature) and dilated when smaller spatiotemporal visual changes are perceived. Moreover, participants felt confident about their half-duration judgements, as evidenced by the fact that they did not change their answers at the end of each trial (less than 5% rectification rates on average). Our findings hold true for different types of visual changes, and different timing intervals, up to and including 3 minutes. Debriefing sessions with participants in *Experiment 2* indicated that confounding factors like fatigue effects might be masking the main effect of visual features in the longer five-minute trials, requiring further investigation.

Experiments 1 and 2 relied on a binary task (*Task A*) to investigate the existence of a relationship between the magnitude of visual features and half-duration judgements. In *Experiment 3*, we used duration judgements to estimate *how much* time perception could be compressed or dilated using a double production task (*Task B*) and the more abstract feature of visual complexity. In *Experiment 3* the ratio of the *empty productions* (those done in the absence of visual stimuli) to the target 30s duration was 1.12. This ratio was consistent with previous literature [42], which estimated a mean 10% deviation from actual time in interval timing judgements. Following this same trend, both H-VC and L-VC made time be perceived as shorter when compared to the empty production (participants took longer to indicate 30s had passed) since both conditions contain more visual changes than an empty scene. Despite the change in task design, we observed the same trend as in *Task A*, where larger visual changes perceptually compressed time. The high accuracy achieved on the auxiliary tasks in *Experiment 3* suggest that participants were engaged throughout the experiment and concentrated on task completion when 30s trials were presented. It would be highly unlikely for participants to accurately complete these tasks while also counting, since their working memory capacity is limited [17]. Participants also self-reported, in a post-study questionnaire, that they were not counting, so we have good reason to believe that we are indeed measuring perceived estimates of time. However, *Experiment 3* had a low power (0.407, below the commonly accepted threshold of 0.8). This means that the interpretation of *Experiment 3* results, while interesting, is limited due to a small sample size. Further experiments should be carried out to confirm the preliminary effects found in this work. The fact that we found no significant differences between the perceived workload of the different conditions (H-VC vs L-VC) in Experiment 3 suggests that participants did not experience more cognitive demand in either condition. We believe our findings through Experiments 1-3 cannot be simply attributed to an increase in cognitive load due to the larger per-pixel variations (spatiotemporal visual changes) that participants have to process in the high magnitude levels of each visual feature. Instead, our results could be explained by demands on attentional resources as discussed below.

In retrospective judgments, participants are not informed in advance that a temporal judgement will be made [411], and as a result have to rely on memory in estimating the passage of time, which may result in confounds. Instead, we focused on the real-time perspective with stimuli that more closely approximate real-world applications and prospective judgements where people are aware of the passage of time. Contrary to findings for millisecond timing [42, 421], we found that larger asemantic visual changes actually shorten perceived interval time perception in a prospective setting. Our experiments suggest that all of the following visual features affect time perception: spatiotemporal visual changes (luminance contrast and temporal frequency), field of view, and the number of distinct sources of visual information (visual complexity, which in our case was the number of unique video streams). This apparent inversion of the effect when compared with millisecond timing might be explained by the fact that different timing mechanisms may guide the perception of different temporal magnitudes [42].

The perceived compression of time under large visual changes is the inverse of the well-known oddball effect [269]. In the oddball effect, no visual changes (a repetition of the same stimuli) cause time to be perceived as shorter when compared with a visual change (a different stimulus). However, this might be more related to prior expectations than to visual changes. In the oddball effect, users *get used to* watching the same stimulus repeatedly. The appearance of an unexpected, different stimulus contradicts the users' expectations, potentially *capturing their attention* [386]. In contrast, our experiments are designed to prevent participants from becoming accustomed to a particular duration or condition, for several reasons: first, the order in which trials from different magnitudes or visual features are presented to the participants is randomized. Second, we evaluate the answer of our participants at two possible sampling points of which they are unaware during the experiment, to prevent them from learning the moment at which the question presented in each trial will appear. Third, we make use of realistic, distinct stimuli for each trial in order to maintain participant engagement, which we believe prevents participants from forming specific expectations.

The stimulus information processing load is a contextual factor that can inversely affect duration judgements in prospective paradigms [427]. Although low-level visual processing might involve autonomous processes that do not need the support of cognitive resources, high order cognitive processes and early perceptual processing may not be completely independent [195]. In that sense, an increase in visual processing due to larger visual changes could be affecting timing interval mechanisms, either via an increased use of cognitive resources, which are limited, or by a shift in attention from the temporal task to the observation of the scene around the participant. A possible explanation for our findings could be found in the attentional resources theory [40]. In this work, Brown studied how non-temporal tasks affected a concurrent temporal task. The main findings showed a classic interference effect: "the concurrent nontemporal tasks caused temporal productions to become longer (longer productions represent a shortening of perceived time) and/or more variable than did timing-only conditions". Under the attentional resources theory, a limited amount of attentional resources are split between the tasks being carried out at each moment. Nontemporal tasks thus take resources away from the attention that would be otherwise allocated to the feeling of time passing. In our case, the higher volume of visual information (larger perceived visual changes) may distract attention from the passage of time, resulting in time distortions. This change in the focus of attention from temporal to non-temporal tasks is known to have a major influence on timing behavior according to the attentional-gate model [427] and related empirical studies [278]. In the attentional-gate model, allocation of attention to time acts like a gate that regulates how often or how much of the pulses produced by a temporal pacemaker are then cognitively processed. In our case, the high levels of the different visual features might elicit a shift of attention from temporal to visual processing, causing fewer temporal pulses to be processed and effectively shortening the perceived time. Conversely, low levels of our visual features might require less processing (thus attention could be shifted to temporal processing). This low attentional demand could even cause a feeling of boredom in comparison [426], which would in turn lengthen the perceived time. In fact, this feeling of boredom might be one of the reasons why we could not directly extend the experiment of Xuan et al. [421] from millisecond to interval timing in our prestudies: our participants did not have enough information to process and the different conditions were not making up for that lack of stimulation.

Limitations. Throughout our experiments, due to circumstances beyond our control, we had to adapt to a change in HMD hardware between Experiments 1 and 3, and a limited participant pool in the follow-up study to Experiment 1 (Section 2.2) and in Experiment 3. While the hardware (the HMDs) used in Experiment 1 (HTC Vive Focus) and Experiment 3 (Oculus HMD CV1) was different, we were careful to ensure that both HMDs shared the same key characteristics: both HMDs have spatial (6DoF) tracking, the same FOV (110 visual degrees) with a very similar refresh rate (75Hz for the Focus and 80 for the Oculus). Besides, the resolution of the stimuli displayed in both Experiments was the same, regardless of the native HMD resolution. We thus believe that this change did not have an effect on the answers of our participants. For this work, we derive our main conclusions mostly from the analyses of Experiment 1, which was run on 89 participants and had sufficient statistical power. For completeness, we chose to nevertheless report the results from the follow-up study to Experiment 1 and Experiment 3, despite those experiments having low power. While we do not base our conclusions on these experiments, they do provide further validation that our observed trends from Experiment 1 continue to hold under changing conditions.

Future directions. It has also been reported that aging affects time perception [59], with older adults experiencing perceptual time compression. To study how our reported effects generalize across demographic variables such as age, it would be necessary to expand the current participant pool (with an average age of 25.31 years; SD: 4.49). Moreover, we designed the experiments to ensure that higher magnitude levels of each visual feature induced larger visual changes across time. However, high magnitude levels also implied higher absolute values of each studied visual feature. Evidenced by Montague's theory [245] and the perceptual break we found in our pilot studies when trying to directly extend an experiment with simple, fixed stimuli to interval timing, we believe that the changes in time perception are caused by larger visual changes being perceived, instead of by the absolute visual value. Nevertheless, a full isolation of the magnitude change from its absolute value could help confirm this in future work. Furthermore, we only sampled two sparse levels (low and high) for the magnitude of each visual feature. A more fine-grained sampling strategy with an increased number of magnitude levels could further illuminate the observed effects in an analytical manner. Another interesting question for future investigation is the minimum required change in visual stimuli to observe differences in time perception. Our low and high levels for each visual feature were selected based on preliminary tests, but different variations between these levels might result in differently sized effects.

Conclusion. In summary, we suggest that real-world, interval-scale time perception can be compressed or dilated through asemantic and realistic visual changes. These findings imply that we can alter time perception without affecting the semantic content, making these results practical for application purposes. When designing applications where time judgements are relevant, our findings could be applied as general design guidelines: i.e., use a high contrast color palette, show changes in a larger portion of the field of view, or make faster camera cuts in a movie scene to make time seem shorter. Related to HMDs, compressing perceived time might also be helpful when controlling fatigue [49]. These findings have the potential for profound impact on practical applications, such as reducing perceived discomfort during medical treatment with virtual reality immersion [318], improving response time in highly dynamic tasks such as vehicle driving with heads-up displays, or lowering fatigue in professional training [83].

APPENDICES

3.A ANOVA POST HOC TESTS

In the following, we report the post hoc tests of our ANOVAs, including a visualization of the multiple comparison of population marginal means.

Experiment 1. Statistical Analysis. A 2x2x2 ANOVA was used to check for significant effects (with magnitude, visual feature (frequency and contrast), and sampling point as factors) for Group 1.1 (45 participants). A 2x2 ANOVA was used to check for significant effects (with magnitude differences in the FoV and sampling point as factors) for Group 1.2 (44 participants). The answer variable was binary ("more than half" or "less than half" of the time elapsed) in both analyses. With a significance level established at p=0.05 and power of 0.895 and 0.894, both ANOVAs revealed that magnitude had a significant effect on the answers (F=45.03, p<0.001, partial $\eta 2$=0.580 for the three-way ANOVA, F=12.39, p<0.001, partial $\eta 2$=0.228 for the two-way ANOVA), while the sampling point (F=0.91, p=0.340, partial $\eta 2$=0.012 for the three-way ANOVA; F=0.49, p=0.482, partial $\eta 2$=0.009 for the two-way ANOVA) and visual features (F<0.01, p=0.964, partial $\eta 2$<0.001 only tested in the three-way ANOVA for frequency and contrast) did not. No significant interactions between the fixed factors were found. Tables 3.4 and 3.5 show each ANOVA in detail, while Figures 3.7 and 3.8 show the post-hoc visualization: each pairwise comparison between the possible different conditions where only groups with different magnitude were significantly different.

Follow-up replication of Experiment 1. Statistical Analysis. A 2x2x2 ANOVA was used to check for significant differences (with visual feature, magnitude and sampling point as factors). The answer variable was binary ("more than half" or "less than half" of the time elapsed). While the trend observed is the same as in Experiment 1, no significant difference was found for any

Source	Sum Sq.	D.F.	Mean Sq.	F	P-value
Magnitude	10.936	1	10.936	45.029	<0.001
Sampling point	0.221	1	0.221	0.992	0.340
Visual Feature	<0.001	1	<0.001	0.002	0.964
userID	7.542	44	0.1714	0.706	0.927
magnitude*sp	0.003	1	0.003	0.013	0.908
magnitude*visFeat	0.059	1	0.059	0.241	0.624
sp*visFeat	0.001	1	0.001	0.005	0.943
magnitude*sp*visFeat	0.089	1	0.089	0.368	0.544
Error	338.059	1392	0.243	-	-
Total	356.996	1443	-	-	-

Table 3.4: 2x2x2 ANOVA of Experiment 1 (Group 1.1), with fixed factors: magnitude, sampling point and visual feature. The effect of each participant was considered as a random variable (userID, underlined). First and second order interactions between the fixed variables were considered. Only magnitude had a significant effect in the response variable.

Source	Sum Sq.	D.F.	Mean Sq.	F	P-value
Magnitude	2.774	1	2.774	12.390	<0.001
Sampling point	0.111	1	0.111	0.496	0.482
userID	9.275	43	0.216	0.963	0.541
magnitude*sp	0.392	1	0.392	1.752	0.187
Error	68.293	305	0.224	-	-
Total	80.845	351	-	-	-

Table 3.5: 2x2 ANOVA of Experiment 1 (Group 1.2), with fixed factors: magnitude and sampling point. The effect of each participant was considered as a random variable (userID, underlined). First order interactions between the fixed variables were considered. Only magnitude had a significant effect in the response variable.

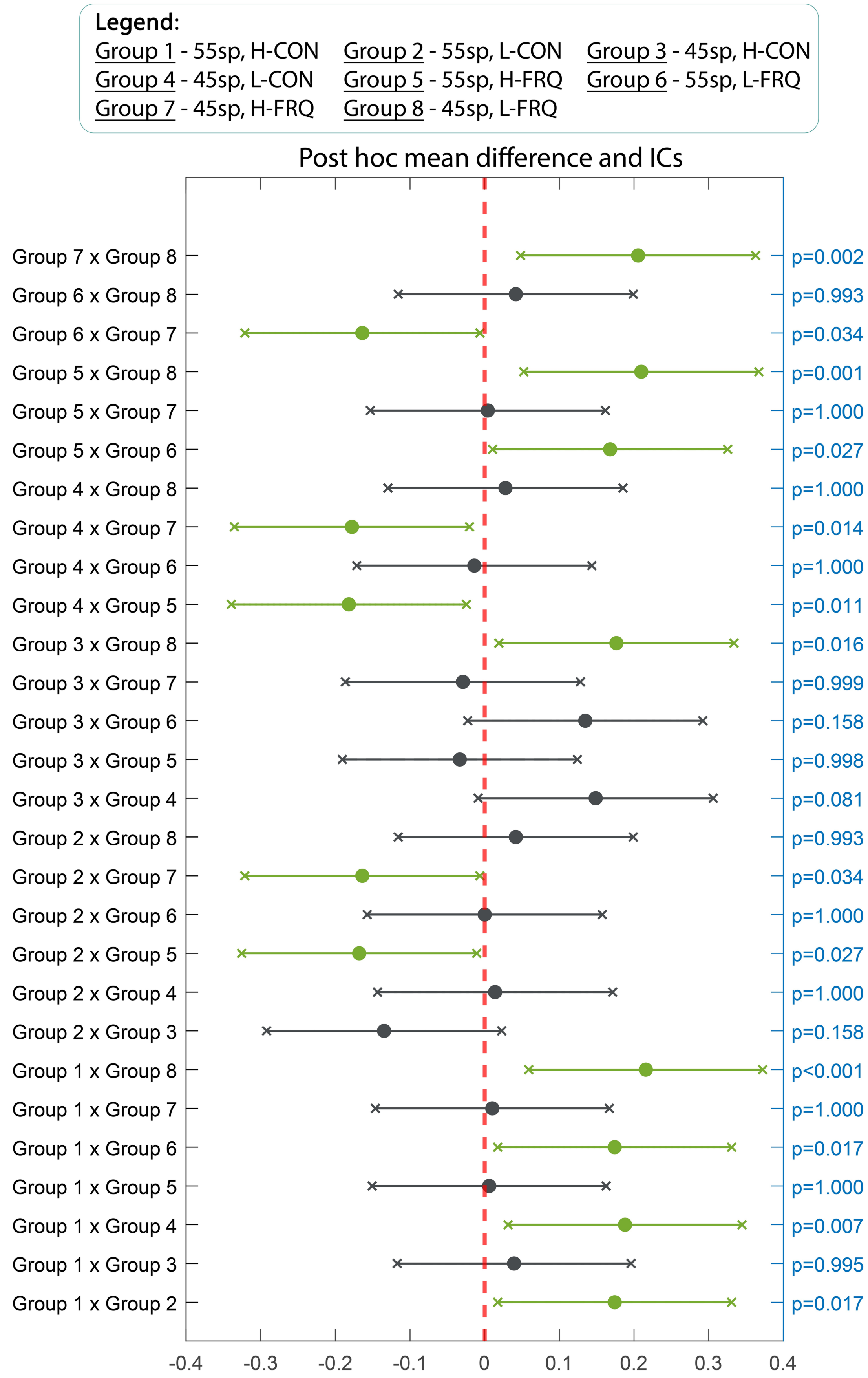


Figure 3.7: Post hoc analysis of the three-way ANOVA (Contrast and Frequency conditions). Significant differences were found only between groups with different magnitude levels. Figure A1 shows the 95% confidence interval of the mean difference between each multiple comparison, as well as the p-value for each comparison.

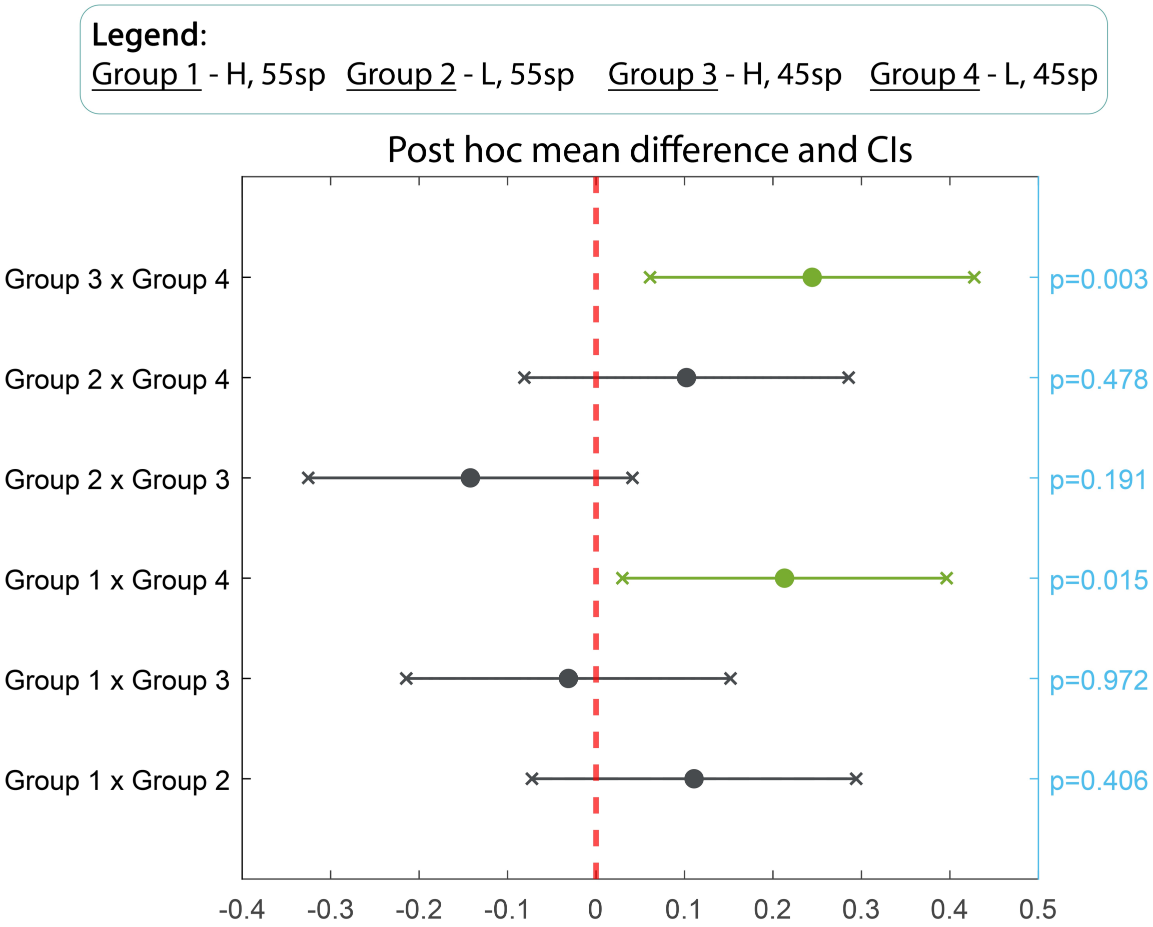


Figure 3.8: Post hoc analysis of the two-way ANOVA (FoV conditions). Significant differences were found only between groups with different magnitude levels. Figure A1 shows the 95% confidence interval of the mean difference between each multiple comparison, as well as the p-value for each comparison. Groups 1 and 3 present H-FOV conditions while Groups 2 and 4 present L-FOV conditions.

Source	Sum Sq.	D.F.	Mean Sq.	F	P-value
Magnitude	0.143	1	0.143	0.572	0.451
Sampling point	0.321	1	0.321	1.288	0.259
Visual Feature	0.143	1	0.143	0.572	0.451
userID	1.964	6	0.327	1.311	0.259
magnitude*sp	0.571	1	0.571	2.289	0.134
magnitude*visFeat	0.036	1	0.036	0.143	0.706
sp*visFeat	0	1	0	0	1
magnitude*sp*visFeat	0.321	1	0.321	1.288	0.259
Error	24.464	98	0.250	-	-
Total	27.964	111	-	-	-

Table 3.6: 2x2x2 ANOVA of the follow-up replication of Experiment 1 in CDs, with fixed factors: magnitude, sampling point and visual feature. The effect of each participant was considered as a random variable (userID, underlined). First and second order interactions between the fixed variables were considered. No significant effects were found

Source	Sum Sq.	D.F.	Mean Sq.	F	P-value
Magnitude	0.420	1	0.420	4.980	0.037
Error	1.685	20	0.084	-	-
Total	2.104	21	-	-	-

Table 3.7: ANOVA of Experiment 3 (high and low visual complexity are the two levels of the single factor -magnitude- that is tested). This factor had a significant effect on the answers of the participants.

of the tested factors (magnitude F=0.56, p=0.454; visual feature F=0.56, p=0.454; sampling point F=1.27, p=0.263) probably due to the small sample size (7 participants). Table 3.6 shows the ANOVA in detail, while Figure 3.9 shows each post-hoc pairwise comparison between the possible different conditions.

Experiment 3. Statistical analysis. An ANOVA was carried out to compare differences in magnitude for H-VC and L-VC levels. NASA-TLX questionnaire differences were tested with a t-test. The answer variables were continuous (for the produced durations) or discrete (for the NASA-TLX scores). The independent variable was binary (for low and high magnitude levels of visual complexity). Participants took more time to indicate that 30s had passed in the H-VC condition, suggesting that time was perceived as significantly shorter under higher visual complexity (ratio of 1.38 for H-VC vs 1.10 for L-VC, power=0.407, F=4.98, p=0.0372, partial $\eta 2$=0.24, normality of distribution checked with Anderson-Darling tests). The complete information about the ANOVA can be found in Table 3.7. The mean difference between groups H-VC and L-VC was -0.276 (CI [-0.534 -0.018], p=0.0372).

3.B GLMM ANALYSIS OF EXPERIMENT 1

In the following, we report an additional statistical analysis for Experiment 1. In this GLMM we analyze together participant Groups 1.1 and 1.2 (89 participants) to check for significant interactions of the visual features (luminance contrast, temporal frequency and field of view) for completeness. We find consistent results with the previous separated ANOVA analyses: only magnitude has a significant effect on time perception.

The answer variable was binary, and the significance level was established at p=0.05. The information about the GLMM can be found in Table B1. Following standard Matlab nomenclature, this is the model we tested:

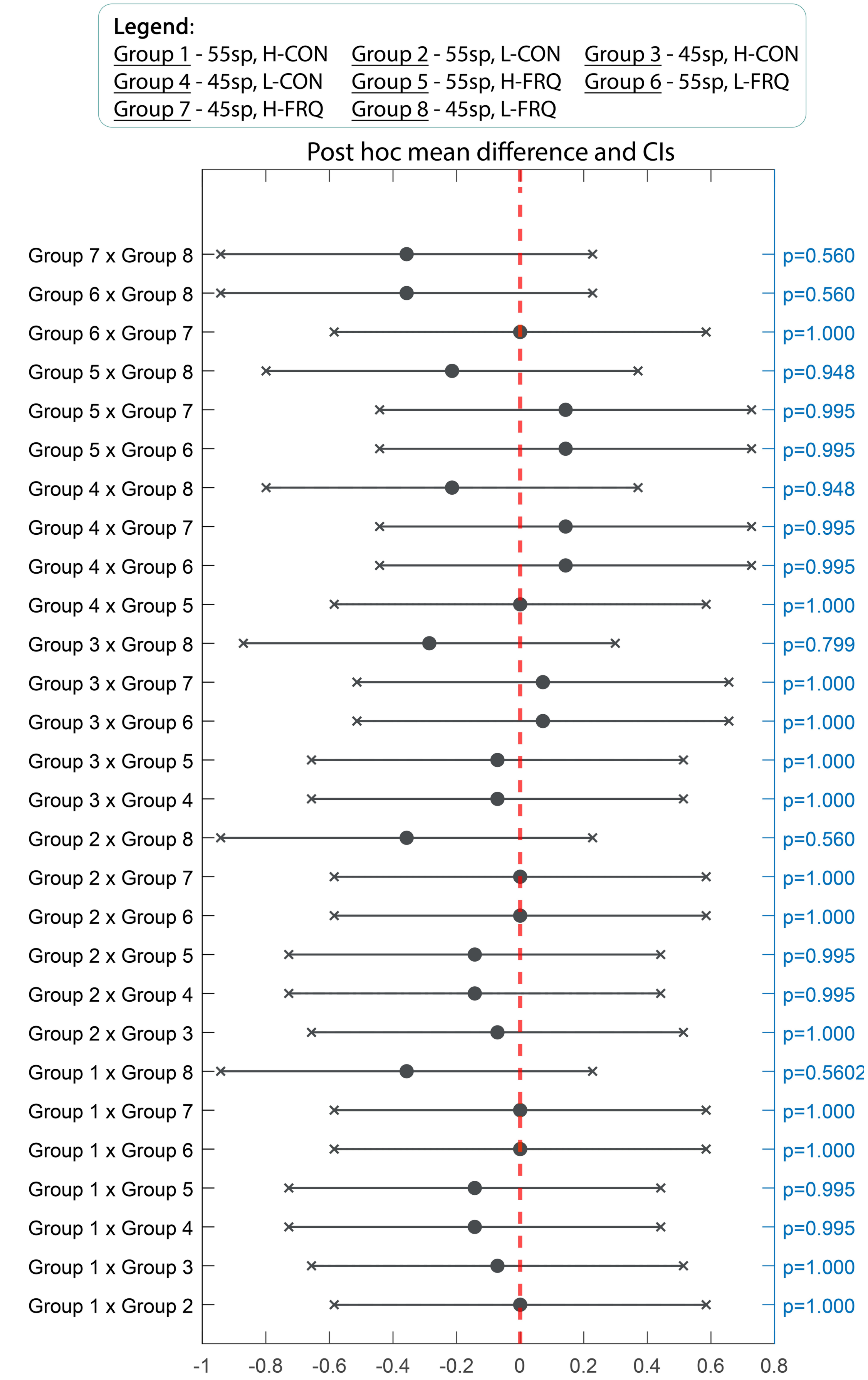


Figure 3.9: Post hoc analysis of the three-way ANOVA (Contrast and Frequency conditions). No significant differences were found. Figure A3 shows the 95% confidence interval of the mean difference between each multiple comparison, as well as the p-value for each comparison.

	Estimate	**SE**	**T-stat**	**p-value**	**CI**
(Intercept)	0.775	0.151	5.121	<0.001	{0.478, 1.071}
Sampling point	-0.175	0.201	-0.868	0.385	{-0.570, 0.220}
Magnitude	-0.751	0.199	-3.764	<0.001	{-1.142, -0.360}
Visual Feature	-0.222	0.141	-1.582	0.114	{-0.498, 0.053}
userID	-0.006	0.003	-1.946	0.059	{-0.011, 0.001}
magnitude*sp	0.144	0.282	0.512	0.609	{-0.408, 0.697}
sp*visFeat	0.152	0.182	0.837	0.403	{-0.205, 0.510}
magnitude*visFeat	0.121	0.183	0.660	0.509	{-0.238, 0.480}
magnitude*sp*visFeat	-0.344	0.261	-1.320	0.187	{-0.856, 0.167}

Table 3.8: GLMM of Experiment 1, considering Groups 1.1 and 1.2, with fixed factors: sampling point, magnitude and visual feature. The effect of each participant was considered as a random variable (userID, underlined).

$$\begin{gathered} answers = 1 + (1|userID) + samplingPoint * level + \\ samplingPoint * freq + level * freq + samplingPoint : level : freq \end{gathered} \tag{3.1}$$

$$\chi^2 - statistic \text{ vs. } constantmodel : 81.2, p - value = 2.84e - 14 \tag{3.2}$$

Part III

MULTIMODAL PERCEPTION IN IMMERSIVE ENVIRONMENTS

In this part we focus on how multimodal perception affects user experience in immersive environments. We focus on the interplay between visual and other sensory modalities. First we present an in-depth state-of-the-art review of the different modalities and the existing hardware to provide sensory feedback, their effects on several aspects of user experience and advantages of the use of several sensory modalities in virtual reality. We then focus on how a mismatch between visual and auditory modalities can result in a suppressive effect or illusion which significantly degrades visual performance. Finally we reflect on how a correctly synchronized audio source can modulate the perception of visual appearance.

4 Multimodality in Virtual Reality

Here we describe an in-depth state-of-the-art review of the potential of the use of several sensory modalities in immersive environments. We include examples of the benefits of multimodality in key aspects of user experience (including realism, presence and user performance), as well as a set of applications that have already benefited from the use of multimodality, from medicine to training and simulation.

This work has been published in ACM Computing Surveys.

D. Martin*, S. Malpica*, D. Gutierrez, B. Masia & A. Serrano
Multimodality in VR: A survey
ACM Computing Surveys 2022, 54(10s)
∗ Joint first authors

4.1 Introduction

Virtual Reality (VR) is inherently different from traditional media since it introduces additional degrees of freedom, a wider field of view, more sophisticated sound spatialization, or even gives users control of the camera. VR immersive setups (such as head-mounted displays (HMDs) or CAVE-like systems) thus have the potential to change the way in which content is consumed, increasing realism, immersion, and engagement. This has impacted many application areas such as education and training [52], rehabilitation and neuroscience [418, 313], or virtual cinematography [334]. One of the key aspects of these systems lies in their ability to reproduce sensory information from different modalities (mainly visual and auditory, but also haptic, olfactory, gustatory, or proprioceptive), giving them unprecedented potential. Although visual stimuli tend to be the predominant source of information for humans [361, 43], additional sensory information helps to increase our understanding of the world. Our brain integrates different sources of sensory feedback including both external stimuli (visual, auditory, or haptic information) and internal stimuli (vestibular or proprioceptive cues), thus creating a coherent, stable perception of objects, events, and oneself. The unified experience of the world as we perceive it therefore emerges from these multimodal cues [285, 338]. These different sources of information must be correctly synchronized to be perceived as belonging together [257, 281], and synchronization sensitivity varies depending on the context, task and individual [84]. In general, different modalities will be perceived as coming from a single event or object as long as their temporal incongruency is shorter than their corresponding window of integration [212] (for more information see Chapter 6).

When exploring virtual environments, the presence of stimuli from multiple sources and senses (i.e., multimodality) and their potential overlaps (i.e., crossmodality), may also enhance the final experience [238]. Many works have described techniques to integrate some of these stimuli to produce more engaging VR experiences, or to analyze the rich interplay of the different senses. For instance, leveraging the window of integration mentioned above may alleviate hardware limitations and lag time, producing the illusion of real-time performance; this is particularly useful when different modalities are reproduced at different refresh rates [55]. Moreover, VR is also inherently well suited to systematically study the integration process of multimodal stimuli [18], and analyze the complex interactions that occur when combining different stimuli [212] (see Figure 4.1 and Chapter 6).

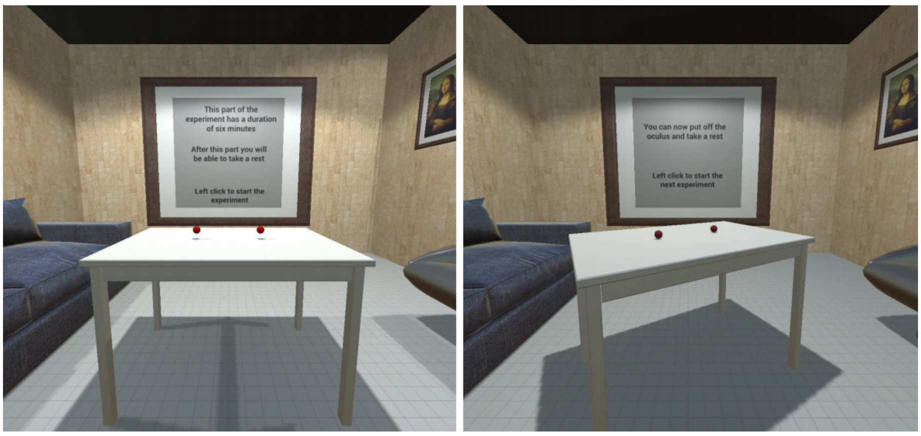

This part of the experiment has a duration of six minutes
After this part you will be able to take a rest
Left click to start the experiment
You can now put off the oculus and take a rest
Left click to start the next experiment

Figure 4.1: VR can be used to systematically analyze the interactions of multimodal information. In this example, we studied the influence of auditory signals in the perception of visual motion [212] (see Chapter 6 for further information). We found that different temporal synchronization profiles affected how the stimuli were perceived: When the visual (red balls moving) and auditory (an impact sound) stimuli were correctly synchronized, users perceived a unified event, in particular a collision between both balls.

In this survey we provide an in-depth review of multimodality in VR. Sensory modalities include information from the five senses: visual for sight, auditory for hearing, olfactory for smell, gustatory for taste, and haptic and thermal for touch. Apart from the five senses, we also consider proprioception, which can be defined as the sense of self-movement and body position, and has been defined as the *sixth sense* [60, 387]. We synthesize the existing body of knowledge with a particular focus on the *interaction* between sensory modalities focusing on visual, auditory, haptic and proprioceptive feedback; in addition, we offer an extensive overview of existing VR applications that directly take multimodality into account.

4.1.1 *The five senses*

The way we perceive the world is defined by the five senses: sight, hearing, smell, taste, and touch. Vision is the dominant sense when retrieving information of our surroundings [295]. We are capable of understanding complex scenes with varying visual patterns, we can detect moving objects in our peripheral view, and we are highly sensitive to light [381]. However, we tend to focus our visual attention in a narrow region of frontal space [361]. In that sense, we rely on hearing to retrieve information from unseen locations. Auditory stimuli can grab our attention irrespective of our orientation, and we are good at filtering out particular sounds in a noisy environment (e.g., the cocktail party phenomenon [15]). The sense of touch includes different aspects: haptic, kinesthetic (related with proprioception), temperature, pressure, and pain sensing. Touch occurs across the whole body, although our hands are our primary interface for this sense. Finally, the senses of smell and taste are closely related. They are often linked to emotions or memories, and can even trigger aversive reactions [243]. Most importantly, besides the particularities of each different sense, and as we will see through this review, our multiple senses influence each other.

4.1.2 *Proprioception*

Proprioception arises from static (position) and dynamic (motion) information [43]. It plays a key role in the concept of self, and has been more traditionally defined as "awareness of the spatial and mechanical status of the musculoskeletal framework" [388]. Proprioceptive information comes mainly from mechanosensory neurons next to muscles, tendons and joints, although other senses can induce proprioceptive sensations as well. A well-known example are visual cues inducing the phantom limb illusion [286].

Manufacturer and model	Resolution per eye	Positional tracking	Max. refresh rate (Hz)	Field of view (degrees)	Display type	Integrated audio	Price
FOVE	1280 x 1440	Yes	70	100	OLED	No - Jack 3.5mm	$600
HP Reverb - Pro	2160 x 2160	Inside-out	90	114	LCD	Built-in headphones	$649
HTC VIVE Pro	1440 x 1600	Yes	90	110	AMOLED	Built-in headphones	$799
HTC VIVE Pro 2	2448 x 2448	Yes	120	120	LCD	Built-in headphones	$800
Oculus Rift S	1280 x 1440	Inside-out	80	110	LCD	In-line speakers	$399
Oculus Quest 2	1832 x 1920	Inside-out	120	104	LCD	Stereo speakers	$399
Samsung Odyssey	1440 x 1600	Inside-out	90	110	AMOLED	Built-in headphones	$500
PlayStation VR	960 x 1080	Outside-in	120	100	OLED	No - Jack 3.5mm	$299
Valve Index	1440 x 1600	Yes	144	130	LCD	No - Jack 3.5mm	$999
Varjo VR-3	2880 x 2720	Yes	90	115	uOLED	Off-ear speakers	$3195

Table 4.1: Overview of predominant current HMD devices. For each of them, we include the resolution per eye, whether they provide positional tracking, their maximum refresh rate (in Hz), their field of view (FoV, in degrees), the type of display, and a current estimate of the final consumer price. The better specs (in terms of refresh rate and FoV) offered by Valve Index come at a higher cost, while other manufacturers opt for cheaper HMDs, potentially more affordable to consumers.

Proprioception plays an important role in VR as well. On the one hand, it helps provide the subjective sensation of *being there* [356, 307]. On the other hand, proprioception is tied to cybersickness, since simulator sickness is strongly related to the consistency between visual, vestibular, and proprioceptive information; significant conflicts between them could potentially lead to discomfort [184, 235].

4.1.3 *Reproducing sensory modalities in VR*

Important efforts have been made in VR so that all the modalities previously mentioned can be integrated. Visual and auditory feedback are the most commonly used, and almost all consumer-level devices integrate these modalities. There is currently a wide variety of manufacturers providing different HMD systems to enjoy VR at consumer level. Each of them offers devices with different capabilities and specifications, at different costs. Table 4.1 compiles an overview of the most relevant devices currently in the market. An open issue in VR is latency [85]: newer HMD models feature higher refresh rates, as well as significantly increased spatial resolution.

Usually, those displays feature a field-of-view (FoV) slightly smaller than that allowed by human peripheral vision. However, new stereoscopic rendering techniques allow to present content in 3D, and therefore perception of materials, depth, or many other cues can be achieved through visual cues [13]. *Auditory* feedback, which is often integrated in the HMD as a built-in feature (Table 4.1), is generally enabled by speakers or headphones, and spatial audio rendering techniques also support our perception of space in virtual environments [11], even enhancing perceived visual properties [212]. *Haptic* feedback is still in an exploratory phase, and can be achieved through a variety of sensors, including wearables [277, 309], physical accessories [366], ultrasounds [217], controller devices [367, 414], rotary components [155], or electric muscle stimulation [198]. Other resources like fans, hear lamps, or even spray bottles have been used to provide additional tactile stimuli in VR [415]. Recently, advances in ultrasound sensor technology have resulted in the creation of a novel haptic device that will allow for mid-air force around virtual objects and interactions [197]; this device is about to reach the consumer market [86]. *Olfactory* stimuli can be provided through smell cartridges or specific hardware [133], and electric stimulation of taste buds has been tested to generate flavors [288]. How all these different feedback modalities can be integrated depends on the particular context, with some algorithms and techniques already proposed for the most common sensory combinations, including audiovisual or audiohaptic stimuli [206].

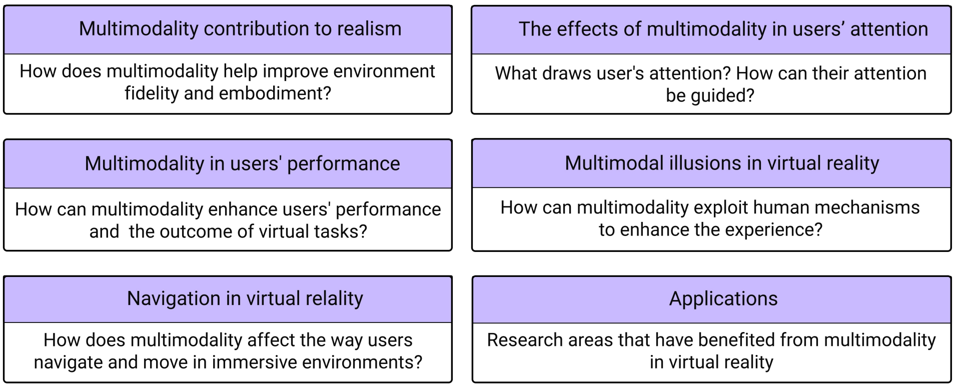

Multimodality contribution to realism
How does multimodality help improve environment fidelity and embodiment?
The effects of multimodality in users' attention
What draws user's attention? How can their attention be guided?
Multimodality in users' performance
How can multimodality enhance users' performance and the outcome of virtual tasks?
Multimodal illusions in virtual reality
How can multimodality exploit human mechanisms to enhance the experience?
Navigation in virtual relality
How does multimodality affect the way users navigate and move in immersive environments?
Applications
Research areas that have benefited from multimodality in virtual reality

Figure 4.2: Structure of this state-of-the-art report. We divide it into different areas of the VR experience in which multimodality can play a key role: user immersion, presence, and realism of the experience (Section 4.3); user attention when exploring the virtual environment (Section 4.4); user performance when completing tasks (Section 4.5); multimodal perceptual illusions that can be leveraged in VR (Section 4.6); and navigational effects of multimodality in virtual environments (Section 4.7). Finally, we review different applications where multimodality has been shown to improve the end goal (Section 4.8), and finalize with a discussion on the need for multimodality, and open avenues of research (Section 4.9).

4.1.4 *Related surveys*

Our perception of the world depends on the integration of information from multiple senses. Several works have reviewed the influence of individual senses *in isolation*, including sight [293], sound [329], or touch [186], which are the three key modalities in VR (see Table 4.5). Several surveys exist focusing on particular aspects of VR. Rubio et al. [303] systematically reviewed advances in communication, interaction, and simulation in VR, pointing out that the key factors to generate appealing virtual experiences include good interactivity, representation, gameplay, and narrative. The latter, narrative, has been explored in depth from a cinematic perspective [299] since content creators no longer have the same level of control over how viewer attention is directed: to find new ways of guiding viewer's attention, the authors reviewed current attention techniques in virtual environments, either unimodal or multimodal, emphasizing how auditory cues can be critical, and the still unexploited potential of haptic devices (see Section 4.8.3)

Other surveys target specific applications of VR, such as education or medicine [296]. For the specific case of clinical medicine, Li et al. [191] found that most of the works in the literature leverage the capabilities of haptic devices to simulate real, clinical tasks, in line with our insights (see Section 4.8.1). Freina et al. [103] reviewed works focused on using VR in education, concluding that it increases the learner's involvement and motivation, which are enhanced with multimodality (Section 4.8.2).

Some works are concerned particularly with cognitive aspects in multimodal environments. Hecht et al. [131] briefly studied how integrating multiple sources can increase presence, enhance attention and improve response time. Koelewijn et al. [169] focused on surveying works related to low-level, audiovisual interactions, concluding that both multisensory integration and attentional processes take place and can interact at multiple stages in the brain. However, multisensory overload can sometimes lead to a preference for simpler environments if not handled correctly; we delve deeper into this in Section 4.3. Other surveys have studied multimodality in *traditional media*, including cognition [362], interaction [153], human-computer interfaces [81, 122], or fusion and integration techniques [16]. Closer to the present work, Melo et al. [238] systematically studied the impact of multisensory stimuli on virtual experiences. Their study suggests that 85% of works tackling multimodality in VR report positive impacts, with only 1% of them reporting negative impacts. They also reported how multisensory experiences in VR are mainly applied in the health domain, science and engineering, teaching, or machinery; which is in line with our reports in

Section 4.8. While they report that these are the more common applications among the 105 studies they surveyed, we provide a discussion on how multimodality is impacting each of these fields.

However, and different from all these works, in this survey we focus on the integration of multimodal information and the benefits and experiences that can be achieved that way, and compile a large body of works studying not only those positive effects, but also their applicability into different disciplines.

4.1.5 *The challenges of multimodality*

One of the main challenges when considering a fully multimodal immersive experience is the gaps of empirical knowledge that exist in this field. As stated before, in this survey our main focus lies in the visual, auditory, haptic and proprioceptive modalities.

This is partly related to the fact that many modalities and their interactions remain unexplored, and there is still much to learn about them. Moreover, the available data on multimodality in VR (both referring to multimodal stimuli and to user data while experiencing multimodal VR) is scarce at best. It is also important to consider the window of integration. The necessity of synchronizing different modalities implies the need for real-time, high fidelity computation. Hardware processing limitations might also imply a constraint in what multimodal techniques should be used in different scenarios. Moreover, not all VR headsets are equally prepared to support multimodality: although most of them can give audiovisual feedback, proprioception and haptic feedback are sometimes limited, while olfactory and gustatory feedback are usually not found at all in consumer-level headsets. For example, most smartphone-based VR headsets do not include controllers, hindering the possibility of including haptic feedback. Many other basic VR systems are not able to track translations either (i.e., only have three degrees of rotatory freedom), which limits proprioceptive feedback. However, one of the most critical risks with multimodality is its definition itself. Although multimodality has the potential to improve user experience, increase immersion, or even improve performance in certain tasks, multimodal applications have to be very carefully designed, making sure that each modality has its function or purpose. This is not a trivial task, since the dimensionality of the problem grows with each added modality. Otherwise, additional modalities might distract and overload the final user, hampering the experience, and diminishing user satisfaction.

4.2 Scope and organization of this survey

In this survey we provide an in-depth review of the most significant works devoted to explore the role and effects of multimodality in virtual reality. We gather knowledge about how multiple sensory modalities interact and affect the perception, the creation, and the interaction with the virtual experience.

The structure of this survey can be seen in Figure 4.2. Since our focus is not on any specific part of the VR pipeline, but rather on the *VR experience* for the user, we have identified several areas of the VR experience in which multimodality plays a key role. First, Section 4.3 is devoted to the realism of the VR experience, which is tied to immersion and the sense of presence that the user experiences. Second, Section 4.4 looks into how multimodality can affect the attentive process of the user in the virtual environment, determining how they explore the environment and what drives their attention within it. Third, Section 4.5 delves into works that demonstrate how multimodality can help the user in completing certain tasks, essentially improving user performance in the virtual environment.

Additionally, there are a number of works devoted to analyzing multimodal perceptual illusions and their perception in VR environments. These, which we compile in Section 4.6, can be leveraged by future techniques to improve any of the aforementioned areas of the VR experience. Some other works have tackled the problem of navigation in VR, which is another integral part of the virtual experience, and Section 4.7 encompasses them. A complete perspective of all these sections is also included in Table 4.5. Finally, we devote Section 4.8 to reviewing application areas that have benefited from the use of multimodal virtual experiences, and conclude (Section 4.9) with a discussion of the potential of multimodality in VR, and interesting avenues of future research.

4.3 the effects of multimodality in perceived realism

Perceived realism elicits realistic responses in immersive virtual environments [355] , and is tied to the overall perception of the experience. There are two key factors that can lead to users responding in a realistic manner: the place illusion and the plausibility illusion [354]. The former, also called "presence" , defines the sensation of "being there", and is dependent on sensorimotor information, whilst the latter refers to the illusion that the scenario that is apparently happening is actually taking place, and is determined by the ability of the system to produce events that relate to the user, i.e., the overall credibility of the scenario being depicted in comparison with the user's expectations. Slater [354] argued that participants respond realistically to an immersive VR environment when these two factors are present. Similar observations were made in telepresence systems [364], where sensorially-rich mediated environments were proved to actually elicit more realistic responses.

Increasing the feeling of presence can therefore enhance the experience by eliciting more realistic responses from the users, and as aforementioned, increasing the perceived realism has a positive impact in the feeling of presence. This actually depends on both the virtual environment where the user is placed, and its own representation in there. As happens in the real world [116], all human modalities play a fundamental role, and must be correctly integrated, to construct a coherent notion of the both the virtual environment, and the self. In this section, we will therefore focus on how multimodal cues can affect perceived realism, by affecting both the perception of the environment, and the perception of the self.

4.3.1 *Perception of the environment*

The perceived realism of virtual environments is a key concern when designing virtual experiences, therefore many works have been devoted to investigate how multimodality and crossmodality can indeed help achieve sensorially-rich experiences. While multimodality refers to the binding of different inputs from multiple sensory modalities, crossmodality involves interactions between different sensory modalities that influence the perception of one another [179, 363]. Chalmers et al. [48] discussed how crossmodal effects in human multisensory perception can be exploited to selectively deliver high-fidelity virtual environments, for instance, rendering with higher visual quality those items related to the current auditory information of the scene, allowing to reduce computational costs in unattended regions of the virtual environment. This work also reports that humans perceive sensory information with more or less attention depending on the task they are executing (i.e., some task require more attention to particular types of stimuli), or if they have already been preconditioned to that kind of virtual environment (e.g., they are used to it).

Traditionally, sound has proven to facilitate visual perception, including enabling a better understanding of the environment, yielding a more comfortable experience, or even increasing performance of visual-related tasks [148, 340]. Seitz et al. [325] conducted a ten-day experiment where two groups of people were trained for auditory-visual motion-detection tasks, one with only visual, and the other with audiovisual stimuli.

Although all of them improved their performance over time, those trained with multimodal stimuli showed significantly better performances. Various works have been thus devoted to this audiovisual integration: Morgado et al. [246] presented a system that generates ambisonic audio for 360º panoramas, so that auditory information is represented in a spherical, smoother way (see Figure 4.3, left). Similarly, Huang et al. [146] proposed a system that automatically adds spatialized sounds to create more realistic environments (see Figure 4.3, right), validating by means of user studies the overall preference of this solution in terms of realism. Indeed, different soundscapes (a sound or combination of sounds created from an immersive environment) are able to increase the sense of presence in VR [328], and as Liao et al. [192] studied, combining visual and auditory zeitgebers (periodically occurring natural phenomena which act as cues

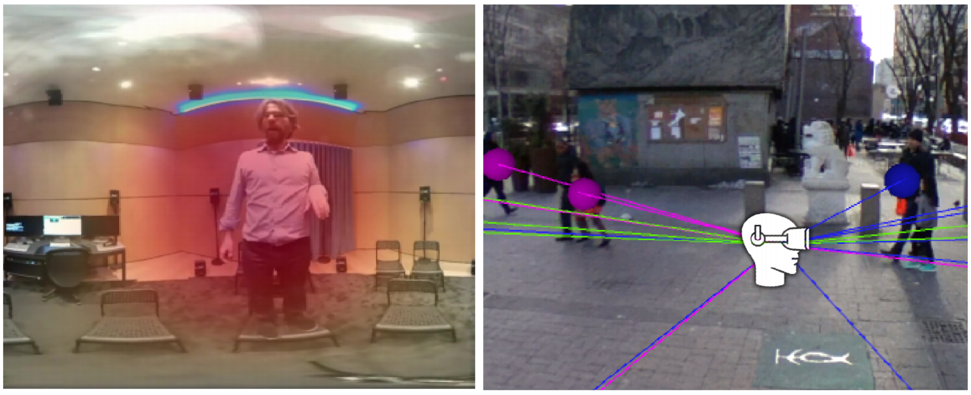


Figure 4.3: Including correct and coherent auditory information in the virtual environment has been proved to increase realism and immersion. *Left*: A system that automatically generates ambisonic information that creates a smoother acoustic experience for the scene [246]. *Right*: A framework to include auditory information into 360° panoramas depending on the elements in the scene [146]. In both cases, their validation experiments yield users' preference when auditory information is included, and an overall increase in the perceived realism and immersion.

in the regulation of biological rhythms), which act like synchronizers, could actually enhance presence, even influencing time perception. All these previous works suggest that using auditory information, either spatialized or not, enhances the realism of the experience, although some of them warn about the potential backfire of increasing the cognitive load, which can negatively impact users' confidence [157].

However, multimodal integration can also present some drawbacks: Akhtar and Falk [3] surveyed current audiovisual quality assessment and found that auditory information may cause discomfort and decrease the quality of the virtual experience [305]. To avoid negative effects during multimodal integration, different sensory cues should be not only realistic, but also coherent to the environment and between them. Proprioception also plays an important role in eliciting realism, as it contributes to the feeling of the user being there.

Although some works have demonstrated that some manipulations in virtual movement directions and distances can be unnoticeably performed (either by manipulating the environment itself through the game engine or by modifying the real-to-virtual mapping of users' movement) [333, 180], users tend to expect their virtual movements to match their real ones, to maintain a coherent experience.

In this line, Mast and Oman [226] studied the so-called visual reorientation illusions: When the environment is rotated above a given noticeable threshold in any axis, users can perceive that the expected vertical axis does not match the virtual one, and conflicts between visual and vestibular cues may arise, potentially causing motion sickness. Although the effect of this illusion is stronger for elder users [142], an incoherent spatial estimation in VR can potentially diminish the perceived realism.

Including additional modalities can also enhance environment realism. In particular, giving realistic feedback with respect to what users expect to happen actually increases plausibility. Normand et al. [265] showed that it is possible to induce a body distortion illusion by synchronous visual-tactile and visual-motor correlations (see Figure 4.4). In a similar fashion, Hoffman compared the realism of virtually touching an object with that of touching it physically at the same time [138], yielding a significant increase in perceived realism when the object was physically touched too. Similar results were obtained with taste and olfactory cues [140]: They found a preference on smelling and physically biting a chocolate bar in contrast to only virtually biting it. The level of presence achieved depends on the different combinations of sensory feedback, and multi-sensory systems have been proved to be superior to traditional audio-visual virtual systems in terms of the sense of presence and user preference [157]. Similar conclusions were obtained by Hecht et al. [131], who reported that multimodality led to a faster start of the cognitive process, which ultimately contributed to an enhanced sense of presence. However, and even if the benefits of multimodal integration are widely known and shared between researchers and practitioners, there is still much to learn about the limits and drawbacks of multisensory integration, and studying up to what point multimodal interaction can be safely applied to increase perceived realism in different scenarios remains an interesting future avenue.

4.3.2 *Perception of the self*

Virtual experiences are designed for humans, and in many occasions, users are provided with a virtual representation of themselves. This is a very effective way of establishing their presence in the virtual environment, hence contributing to the place illusion [354]. This representation does not need to be visually realistic, but it has to be coherent enough with the users' actions and expectations to maintain the consistency of the experience. In the following, we review different works that have leveraged multimodality in virtual reality to achieve consciousness of the self and embodiment, and therefore to create realistic representations of the users.

Having the feeling of being in control of oneself is possibly one of the main characteristics that VR offers [354]. The feeling of presence is possible without being in control; however, being able to control a virtual body highly increases this illusion [320]. The sense of embodiment gathers the feeling of owning, controlling, and being inside a body. As Kilteni et al. [165] reported, this depends on various subcomponents, namely the sense of self-location (a determined volume in space where one feels to be located), the sense of agency (having the subjective experience of action, control, intention, motor selection and the conscious experience of will), and the sense of body ownership (having one's self-attribution of a body, implying that the body is the source of the experienced sensations). Other factors like the proximity of virtual objects to the body also have an effect on the sense of embodiment [324]. All these concepts (such as presence or embodiment) are intrinsic characteristics that VR can achieve, and they yield the self-consciousness feeling that makes VR so different from other media.

Multimodality has been largely studied as a means to enhance those sensations. Particularly, presence is tied to the integration of multiple modalities, and many works have demonstrated how it is increased when multiple sensory information is combined [312], as opposed to unimodal (i.e., only visual) systems [157]. For instance, Gonçalves et al. [116] designed an experiment where three groups of people were exposed to virtual environments including different amount of modalities in the presented stimuli; and reported how users experiencing more modalities reported a higher involvement. Moreover, they remark the positive impact of including haptic feedback in an experience. Blanke et al. [33] discussed the relevance of a series of principles to achieve a correct sensation of bodily self-consciousness, requiring body-centered perception (hand, face, and trunk), and integrating proprioceptive, vestibular, and visual bodily inputs, along with spatio-temporal multisensory information. Sakhardande et al. [308] presented a systematic study to compare the effect of tactile, visual, visual-motor, and olfactory stimuli on body association in VR, with the latter having the strongest effect on body association. Similar insights were proposed by Pozeg et al. [283], who demonstrated the importance of first-person visual-spatial viewpoints for the integration of visual-tactile stimuli, in this case for the sense of leg ownership. The main factors to build embodiment and body-ownership in VR have been widely studied [223]. Spanlang et al. [358] presented technical guidelines to create a core virtual embodiment system, defining three key aspects: (i) a VR module to handle creation, management, and rendering of all virtual entities, (ii) a head-tracking module to map real movements to the virtual environment, and (iii) a display module to present the whole environment. However, designing experiences that are too realistic can have negative aspects and be a drawback in certain specific cases: for example, group pressure of alien virtual avatars can result in users performing potentially harmful actions towards others that they would not normally carry out [254].

The sense of moving (which depends on agency and body ownership, as previously mentioned) is also key to achieve self-consciousness. Kruijff et al. [173] presented a work showing that adding walking-related auditory, visual, and vibrotactile cues could all enhance the participants' sensation of self-motion and presence. Various works have been presented in this line, e.g., investigating the integration of tendon vibrations to give standing users the impression of physically walking [170].

Figure 4.4: *Left*: Synchronizing different modalities increases the feeling of presence and the perception of the self. Moreover, multimodality can even create a distortion of that perception: Normand et al. [265] presented a study where a body distortion illusion is achieved by synchronous visual-tactile and visual-motor correlations. *Right*: Some works have studied how different physical and behavioral factors can directly affect, and even manipulate, embodiment [254], and therefore, the perception of the self.

Sometimes locomotion is not possible, and it has to be externally generated, e.g., by means of a virtual walking system for sitting observers using only passive sensations such as optic flow and foot vibrations [228]. However, these techniques are akin to creating the well-known self-motion illusion: although users are not actually moving, their brain unconsciously assumes they are moving, and their body sometimes generates postural responses [77] to control their stability. Meyer et al. [240] studied the impact of having multimodal (visual, auditory, and haptic) anchor points in the virtual environment in users' postural sway. They report how incongruent cues diminish perceived realism. However, they also remark on the complexity of providing dynamic tactile signals in VR, which leaves an interesting research line in how to exploit tactile cues to increase presence. Some other works have also explored alternatives for cases when locomotion is not feasible, for instance proposing and evaluating a virtual walking system for sitting observers using only passive sensations such as optic flow and foot vibrations [228].

Other modalities may also play an important role in users' self-consciousness: several works have shown that multimodality can dramatically increase the sense of presence [107], although confidence levels for certain tasks are higher in traditional (i.e., audio-visual) virtual environments, due to a higher cognitive load [157]. Besides additional modalities, other factors such as immersion and emotion have been analyzed and argued to have a clear impact on the sense of presence [24]. In particular, audiovisual content eliciting emotional responses (like sadness) can increase engagement and presence, somehow bypassing the immersive effects of specific displays.

As reported in some of the aforementioned works, multimodality presents some challenges and limitations: Gallace et al. [107] focused on the ones associated with the simultaneous stimulation of multiple senses, including the senses of touch, smell, taste, and even nioceptive (i.e., painful) sense, given the cognitive limitations in the human sensory perception bandwidth when users have to divide their attention between multiple sensory modalities. Moreover, situations where some modalities violate interpersonal space may also lead to diminishing presence and comfort [416]. Ultimately, achieving user's self-consciousness depends on finding the right balance between different multimodal cues, and the users' comfort, confidence, and capacity to integrate them. Establishing guidelines towards this balance remains one of the most interesting avenues in multimodal interaction.

4.4 The effects of multimodality in users' attention

When users are exploring or interacting with a virtual environment, different elements or events can draw their attention. Visual attention influences the processing of visual information, since it induces gaze to be directed to the regions that are considered more interesting or relevant (salient

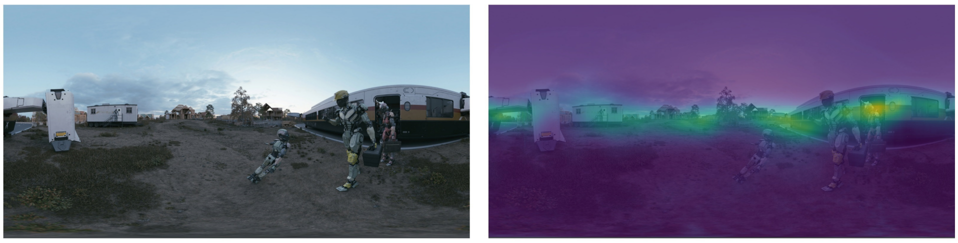


Figure 4.5: Saliency maps show the likelihood of users directing their attention to each part of the scene. Most of the current literature has been devoted to estimating saliency in unimodal, visual stimuli. This image shows the recent visual saliency estimation method proposed by Martin et al. [220] (*Left*: Input panorama. *Right*: Estimated saliency). It has been shown that each sensory modality has the potential of influencing users' attentional behavior, therefore, there is a need for further exploration of multimodal saliency in VR.

regions). The saliency of different regions results from a combination of top-down attentional modulation mechanisms (task-based) and the bottom-up multisensory information these regions provide (feature-based), creating an integrated saliency map of the environment [382]. As discussed in the previous section, VR setups may produce realistic responses and interactions, which can be different from traditional media due to the differences in perceived realism and interaction methods. Therefore, some works have been devoted to understanding saliency and users' attention in VR, offering some key insights about head-gaze coordination and users' exploratory behavior in VR. For example, Sitzmann et al. [353] detected the *equator bias* when users are freely exploring omnistereo panoramas: They observed a bias towards gazing at the central latitude of the scene (equator bias), which often corresponds to the horizon plane.

Saliency has been widely studied, both inside and outside the VR field [420]. Users are more likely to turn their attention and interact with those regions of the scene that provide more sensory information. Therefore, knowing a priori which parts of the scene are more salient may help anticipate how users will behave. So far, most of the works on saliency in VR have been carried out following a unimodal perspective [429, 220]: although several senses can be considered to create a saliency map, they all leverage users' head position and gaze orientation to create probabilistic maps indicating the chances of a user looking at each part of the virtual scene (see Figure 4.5). Based on the study of visual cues, various works have presented systems able to predict users' gaze, depending on the environment and also on user's previous behavior [144, 219].

Multimodality in saliency estimation has been only tackled in traditional media: the integration of visual and auditory information for saliency prediction in videos has been widely explored [69, 242]. All these approaches work under the assumption of audiovisual correlation: moving elements are the source of the auditory cues. In a different approach, Evangelopoulos et al. [87] proposed the addition of text information in the form of subtitles when speech was present in the auditory stream. In their work, saliency was considered as a top-down process, since the interpretation of the subtitles, a complex cognitive task, can distract viewer's attention from other parts of the scene.

Multimodality in saliency prediction for VR still remains in early phases, and only very few works have been devoted to it. Chao et al. [50] proposed the first work that studies user behavior (including saliency corresponding to sound source locations, viewing navigation congruence between observers, and the distribution of gaze behavior) in virtual environments containing both visual and auditory cues (including both monaural and ambisonic sounds). However, there are still many open avenues for future research: Visual saliency and gaze prediction in VR is still in an early phase, and the effects of auditory cues in saliency on virtual scenarios remain to be further explored. Auditory cues in VR may produce more complex effects and interactions than in traditional scenarios, since sound sources are not always in the user's field of view, and there might be several competing audiovisual cues. Additionally, investigating how other senses interact and predominate in saliency and attention can be useful for many applications, specially for content creation. Furthermore, with the proliferation of data-driven methods, it is also crucial to elaborate

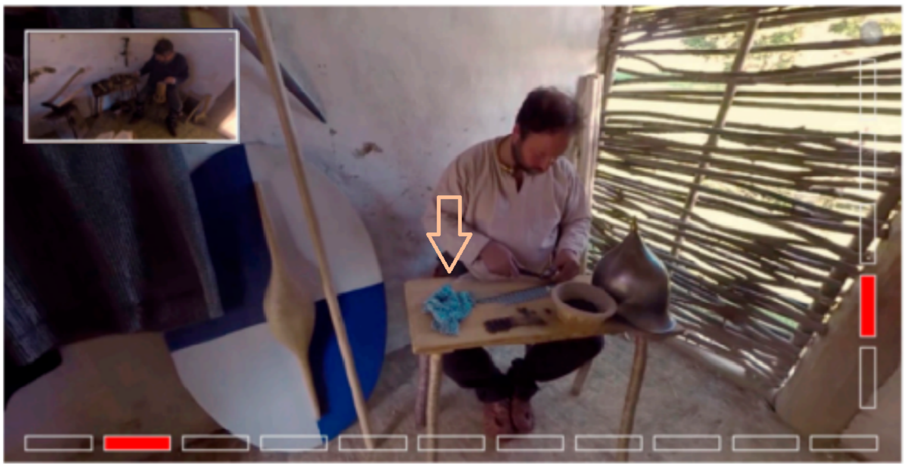


Figure 4.6: Examples of visual guidance methods in VR, adapted from Rothe et al.'s review on users' guidance for cinematic content [299]. Three visual guidance techniques are presented in this image: Arrows pointing regions of interest, picture-in-picture techniques that show information of rear regions, and the position of a point of interest marked with red bars. Most of these techniques are intrusive, hence they may break immersion and realism. With the addition of multimodal cues, guidance can be facilitated while maintaining a positive user experience.

datasets that encompass enough variety of multimodal stimuli to support the formulation of new multimodal attentional models.

Although there is still much to learn about how multimodal cues compete and alter users' behavior, it is well known that multimodality itself has consequences on how users behave in immersive environments [385, 50] . One of the main difficulties when designing and creating content for VR lies in the fact that users typically have control over the camera, and therefore each user may pay attention to different regions of the scene and create a different experience [334, 215]. Therefore, it is usually hard to make assumptions about users' behavior and attention. To support the creation of engaging experiences that convey the creators' intentions, multimodality can be exploited, so that cues from different modalities can induce specific behaviors and even guide users' attention.

For the case of attention, understanding and guiding users attention in VR has been a hot topic during the last years. Various works have explored the use of visual guiding mechanisms, such as central arrows and peripheral flicker to guide attention in panoramic videos [317]. The recent work of Wallgrun et al. [406] compares different visual guiding mechanisms to guide attention in 360° environments [406]. Lin et al. [194] proposed a picture-in-picture method that includes insets of regions of interest that are not in the current field of view, so users are aware of all the elements that surround them. Inducing the users to direct their attention to a specific part of the scene has also been explored, for example, using focus assistance techniques [193], such as indicating the direction of the relevant part, or automatically orienting the world so that users do not miss that part of the experience. Following this line, Gugenheimer et al. [121] presented a motorized swivel chair to rotate users until they were focusing on the relevant part of the scene, while Nielsen et al. [258] forced virtual body orientation to guide users attention to the most relevant region. Other techniques directly let the viewer press a button to immediately reorient the scene to the part containing the relevant information [271]. It is worth mentioning that these kind of techniques have to be taken into consideration with caution, since they can cause dizziness or discomfort due to visual-vestibular conflicts. We refer the reader to Rothe et al.'s work [299] for a complete survey about guidance in VR (see Figure 4.6).

However, guidance techniques are not necessarily constrained to visual manipulations. Multimodality can be also exploited to guide, focus and redirect attention in VR, in many cases achieving more subtle, less intrusive methods. This is important to maintain the users experience, as intrusive methods can alter the sense of presence, immersion, or suspension of disbelief (the temporary acceptance as believable of events or places that would ordinarily be seen as incredible).

As shown in previous sections, sound can help enhance the virtual experience. Besides, it can also be used to manipulate or guide users attention. Rothe et al. [301] demonstrated that the attention of the viewer could be effectively directed by sound and movements, and later [300] investigated and compared three methods for implicitly guiding attention: Lights, movements, and sounds, showing that sounds elicit users' exploratory behavior, while moving lights can also easily draw attention. Other works have explored various unobtrusive techniques combining auditory and visual information, showing that auditory cues indeed reinforce users' attention being drawn towards specific parts of the environment [39]. Similar insights were obtained by Masia et al. [224], who investigated the impact of directional sound during cinematic cuts in VR, finding that in the presence of directional sound cues, users converge much faster to the main action after a cut, even if the sound is misaligned with the region of interest. Given the importance of including sound in VR experiences, Bala et al. [20] presented a software for adding sound to panoramic videos, and studied how sound helped people direct their attention. Later, they examined the use of sound spatialization for orientation purposes [21]. In particular, they found that full spatial manipulation of sound (e.g., locating music in a visual region of interest) helped guide attention. In a similar fashion, some works have studied how to design sound to influence attention in VR [310], and how decision making processes are affected by auditory and visual cues of diegetic (i.e., sounds emanating from the virtual environment itself) and non-diegetic (i.e., sounds that do not originate from the virtual environment itself) origins [44]. However, non-diegetic cues need to be analyzed and presented carefully: the work by Peck et al. [273] showed that a distractor audio can be successful at fostering users' head rotations (and thus redirection); however, users considered this method as unnatural. It has been also suggested that too many sound sources in a VR cinematic video can produce clutter, and therefore hinder the identification of relevant sound sources in the movie [299]. How to use multimodality for guiding users' attention has many open possibilities for further investigation. In this context, establishing guidelines regarding which senses to use, how to combine them, and up to what extent each of them can surpass the others remains for now a complex, unresolved task.

4.5 multimodality in users' performance

Understanding how users perform different tasks in VR is key for developing better interfaces and experiences. Although in many cases task performance highly depends on the users' skills and experience, there are many scenarios where multimodality can play an important role in this aspect: By integrating multiple sensory information we can mimic better the real world, and this can lead to higher performance in different scenarios, comparable to real life. Additionally, multimodal VR technologies are becoming a very powerful tool for training and education, specially in scenarios that can be expensive, or even dangerous, in real life. In those cases, multimodality can help complete some tasks in a shorter period, or with a higher accuracy [131].

The effects of multimodality in task performance have been largely studied in traditional media. Lovelace et al. [199] demonstrated how the presence of a task-irrelevant light enhances the detectability of a brief, low-intensity sound. This behavior also holds in the inverse direction: Concurrent auditory stimuli could enhance the ability to detect brief visual events [262]. Therefore, integrating audiovisual cues may diminish the risk of users losing some relevant information. In a similar line, Van der Burg et al. [390] reported that a simple auditory *pip* drastically decreased detection time for a synchronized visual stimuli. These effects are not only present in audiovisual stimuli: tactile-visual interactions also affect search times for visual stimuli [391]. In most of these works, the experiments were carried out in laboratory conditions with simple stimuli, and therefore studying their applicability and limitations in more complex scenarios remains an interesting avenue. Furthermore, Maggioni et al. [207] studied the potential of smell for conveying and recalling information. They compared the effectiveness of visual and olfactory cues, and their combination in this task, and demonstrated that olfactory cues indeed improved users' confidence and performance. Therefore, the integration of multiple cues has been widely proven to be effective in terms of detectability and efficiency.

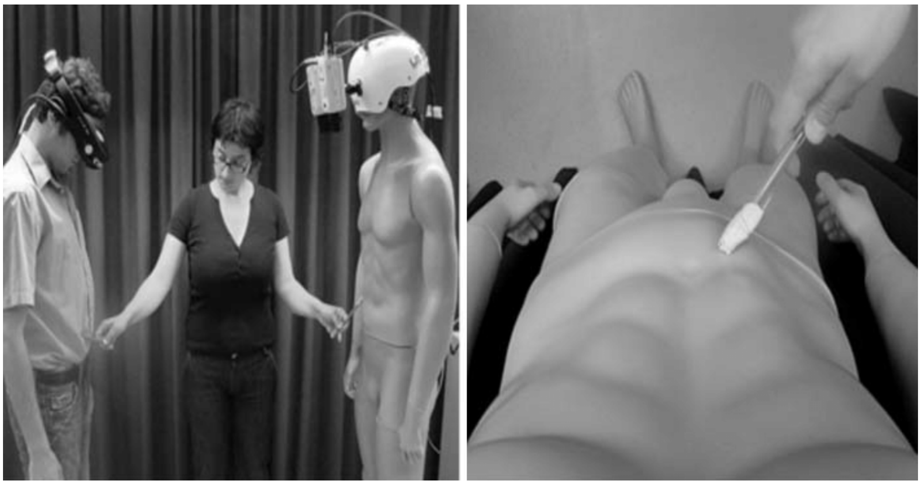


Figure 4.7: Multimodality illusions can change the way users perceive both themselves and the environment. For instance, Petkova et al. [276] studied how proprioceptive and haptic cues could lead to body ownership illusions.

As Hecht et al. [131] studied, this improvement in terms of performance also holds for multimodal VR: When there are multiple senses involved, users start their cognitive process faster, thus they can pay attention to more cues and details, resulting in a richer, more complete and coherent experience. Performance in spatial tasks can be greatly benefited from multimodality [109, 9]. Auditory cues are extremely useful in spatial tasks in VR, and therefore have been widely explored: The effect of sound beacons in navigation performance when no visual cues are available has been explored [405], with some works proving that navigation when no visual information is available is possible using only auditory cues [120]. Other works have exploited this, proposing a novel technique to visualize sounds, similar to how echolocation would work in animals, which improved the space perception in VR thanks to the integration of auditory and visual information [297], or combining the spatial information contained in echoes to benefit visual tasks requiring spatial reasoning [109]. Other senses have also been explored with the goal of enhancing spatial search tasks: Ammi and Katz [9] proposed a method coupling auditory and haptic information to enhance spatial reasoning, and thus improving performance in search tasks. Direct interaction tasks can be also enhanced by multimodality: auditory stimuli has been proven to facilitate touching a virtual object outside a user's field of view, hence creating a more natural interaction [167]. Egocentric interaction is also likely to happen, and proprioception plays an important role in those cases. Poupyrev et al. [280] presented a formal study comparing virtual hand and virtual pointer as interaction metaphors, in object selection and positioning experiments, yielding that indeed both techniques were suitable for different interaction scenarios.

As aforementioned, when developing VR experiences requiring users to complete some tasks, the integration of multiple modalities can increase their performance and spatial reasoning, leading to better, more consistent results. Furthermore, adding certain modalities (e.g., olfactory or haptic information) is not always easy, especially at the consumer level. Enabling these modalities within current consumer-level devices (Table 4.1) remains a future avenue that would not only greatly benefit multimodality in terms of performance, but it would also improve the whole experience. In spite of that, in some cases, combining several modalities can lead to the opposite effect, suppressing or diminishing some abilities [213], hence special care must be paid when designing multimodal experiences (Section 4.1.5).

4.6 multimodal illusions in vr

Multimodality can be leveraged to trick the self perception of the users, or to alter how they perceive the world around them, by means of facilitatory or inhibitory (suppressive) effects, which can have direct implications on how users behave in the virtual environment. Being able to manipulate the experience can be very useful in certain contexts and applications: for instance, sometimes it can be useful to guide the user towards a particular aspect of the virtual environment without disrupting the experience (e.g., in cinematography or videogames). A forced guidance could lead to a reduced immersion feeling, or even rupture of the suspension of disbelief. In other cases, physical space is constrained, and manipulating users' movement may allow to reduce the necessary physical space to complete a task [333]. Manipulating the experience can be also useful to reduce simulator sickness, for instance, by means of manipulating camera control depending on some characteristics such as velocity, acceleration, or scene depth [143]. Although this can be done using a single modality, the use of multimodal cues can improve the effectiveness of these techniques.

Illusion refers to an incorrect perception or interpretation of a real, external stimulus. It can lead to interpreting reality in several ways. Any healthy person can experience illusions without experiencing any pathological condition. However, not every person is affected in the same way by an illusion. Illusions can have physiological (e.g., an after-image caused by a strong light [152]) or cognitive (e.g., the Rubin vase [272]) components. They have been widely studied, as understanding illusions yields valuable information about the limitations of human senses, and helps understand the underlying neural mechanisms that help create the perception of the outside world. Moreover, illusions can allow the altering of users responses to certain tasks, even increasing performance. For instance, Chauvel et al. [51] conducted some experiments where non-golfers practiced putting golf balls, some of them with manipulated holes to enhance their visual acuity. Those who trained under these conditions showed a more effective learning outcome, and a better performance when trying in real-life scenarios. In this subsection we will focus on multimodal illusions or effects, or how illusions in other senses can affect visual perception. For visual only illusions, we refer the reader to The Oxford Compendium of Visual Illusions [342].

Multimodal illusions can be useful for boosting accuracy in certain tasks. For example, multisensory cues can improve depth perception when using handheld devices by simulating tactile responses when holding, or interacting with a virtual object with a force feedback system [38, 372]. Using a small number of worn haptic devices, Glyn et al. [185] improved spatial awareness in virtual environments without the need for creating physical prototypes. Instead of applying contact (haptic feedback) at the exact physical point of the users body that was touching a virtual object, they used a small, fixed set of haptic devices to convey the same information. Their work was based on the funneling illusion [26], in which the perceived point of contact can be manipulated by adjusting relative intensities of adjacent tactile devices. Visuo-haptic illusions allow not only to better perceive the virtual space, but also to feel certain virtual object properties, like weight, that are not easy to simulate. Even further, these properties can be unnoticeably altered when combining multiple sensory information. Carlon [46] showed that users' perception of heaviness can be unnoticeably altered when manipulating their movements in a virtual environment.

The rubber hand illusion is an illusion where users are induced to feel like a rubber hand is part of their body. In VR, proprioceptive and haptic cues can lead to a similar feeling induced either for an arm [424] or for the whole body [276] (see Figure 4.7). Similarly, proprioception can also be altered by modifying the virtual avatar (i.e., distorting the position or length of the virtual arms and hands) while retaining body ownership, allowing users to explore a bigger area of the virtual environment with their body [93]. Regarding audiovisual illusions, the well known McGurk effect has been replicated in VR. The McGurt effect happens when the audio of a syllable is paired with visual stimuli of a second syllable, raising the perception of a third, different syllable. This illusion has been used to study how audio spatialization affects speech perception, suggesting that sounds can be located at different positions and still create a correct speech experience [349, 347]. It was also found that the spatial mismatch does not affect immersion levels, suggesting that computational resources devoted to audio localization could be decreased without affecting the overall user experience. Another interesting audiovisual illusion that appears both in conventional media and VR is the ventriloquist effect, where auditory stimuli coming from a distant source seem to emerge from an actors' lips. The best located or dominant modality (usually vision) overrides the spatial information of the weak modality giving raise to the apparent translation of sound to the location of the visual stimulus [4]. In this sense, auditory stimuli are affected by visual cues [314], with visual stimuli influencing the processing of binaural directional cues of sound localization. In a complementary way, auditory perception can also act as a support for visual perception, orienting users to regions of interest outside the field of view [175]. Not every audiovisual illusion has to do with speech. In the sound-induced flash illusion [339], a single flash paired with two brief sounds was perceived as two separate flashes. The reverse illusion also happened when two flashes were concurrent with a single beep, raising the perception of a single flash.

In addition to illusions, in which new stimuli are sometimes created, there is also the phenomenon of perceptual suppression, in which one stimulus is no longer (completely or partially) perceived due to an external circumstance. For example, visual suppression is often present in the human visual system. The human brain has evolved to discard visual information when needed to maintain a coherent and stable image of the surrounding environment. Two good examples of visual suppression are blinks and saccades [35, 369], which avoid the processing of blurry information without causing perceptual breaks. Perceptual suppression has been demonstrated and used both in conventional media and in VR [369], usually allowing for environmental changes without the users awareness, which is useful in many applications, such as navigation in VR. It has also been studied how stimuli of a given modality can alter or suppress information of a different modality, usually visual. In particular, for traditional media, both auditory [135] or haptic [150] stimuli can suppress visual stimuli. Functional imaging studies [183] suggest that crossmodal suppression occurs at neural levels, involving sensory cortices of different modalities. Crossmodal suppression has not been widely studied in VR. However, a recent study [213] shows that auditory stimuli can degrade visual performance in VR using a specific spatiotemporal layout (see Figure **??** and Chapter 5 for further information). Nonetheless, there is still much to investigate about traditional illusions and whether they still hold in virtual environments. The interaction between senses, and in particular the predominance of some of them against the rest, may also diminish or enhance these phenomena, and therefore remains an interesting avenue for future work. We thus believe that a deeper study of crossmodal interactions, both facilitatory and inhibitory, could greatly benefit VR applications, as well as increase our knowledge on sensory perceptual processing in humans.

4.7 multimodality in navigation

As discussed in previous sections, agency has an effect on the feeling of realism in a virtual experience, and it is achieved when users feel that their avatar responses are coherent with their real actions, which has a direct implication on body ownership (which also depends on other factors [407]). One characteristic that makes VR intrinsically different from any other traditional media, and that contributes to the users' feeling of control of themselves, is its ability to reproduce each movement of the user into the virtual world. Virtual environments naturally elicit exploration, which usually requires the user to move across the virtual environment. In many cases, movement is heavily constrained by the physical space available [333], and therefore a complete 1:1 reproduction of the movement is not feasible. Enabling full locomotion in a VR application (i.e., allowing the user to freely move in the virtual space) would increase the possibilities of the virtual experience.

However, designers and practitioners are aware of the limited size of physical spaces in which users can consume VR. Redirected walking techniques (RDW) emerged in the pursuit of alleviating this limitation: these techniques propose different ways to subtly or overtly manipulate either the user or the environment during locomotion, in order to allow the exploration of virtual worlds larger than the available physical space. Nilsson et al. [260] presented an overview of research works in this field since redirected walking was first practically demonstrated. Nevertheless, most of these works rely on visual manipulations: some of them exploit only visual cues or mechanisms, such as saccades [369] or blinks [180] to perform inadvertent discrete manipulations, whereas others exploit continuous manipulations that remain unnoticed by users [333, 229]. However, these previous works do not exploit cues from other sensory modalities. As we have presented along this work, integrating multiple senses can take these kind of techniques a step further.

Rhythmic auditory stimuli affect how we move [205], and auditory stimuli can be therefore used to actively manipulate our motion perception. Serafin et al. [330] described two psychophysical experiments showing that humans can unknowingly be virtually turned about 20% more or 12% less than their physical rotation by using auditory stimuli: with no visual information available, and with an alarm sound as the only informative cue, users could not reliably discriminate whether their physical rotations had been smaller or larger than the virtual ones. Nogalski and Fohl [263] presented a similar experiment, aiming for detection thresholds for acoustic redirected walking, in this case by means of wave field synthesis: by designing a scenario surrounded by speakers, and with no visual information available, they demonstrated that some curvature gains can be applied when users walk towards, or turn away from some sound source.

Their work yielded similar rotation detection thresholds of ±20%, which is additionally in line with other works, proving the ability of acoustic signals to manipulate users' movements [90] and the potential benefits of using auditory stimuli in complex navigational tasks. Later, Rewkowski et al. [291] confirmed that RDW with auditory distractors can be safely used in complex navigational tasks such as crossing streets and avoiding obstacles. Nilsson et al. [261] revealed similar detection thresholds for conditions involving moving or static correlated audio-visual stimuli. Additionally, Nogalski and Fohl [264] summarized how users behavior significantly varies between audio-visual and auditory only stimuli, with the latter yielding more pronounced and less constant curvatures than with audio-visual information.

Many other sensory modalities can be used both to manipulate user's virtual movement, improving agency and therefore leading to a more natural navigation. Hayashi et al. [128] presented a technique that allows to manipulate the mapping of the user's physical jumping distance and direction. Jumping is an action strongly correlated to proprioception, but it is usually unfeasible due to the available physical space. Manipulating the virtual distance when jumping can allow users to physically jump even when space is constrained, hence proprioceptive cues and realism can be maintained in the experience. Campos et al. [45] also introduced an integration of visual and proprioceptive cues for travelled distance perception, demonstrating that body-based cues contributed to walked distance estimation, attributable to vestibular inputs. Matsumoto et al. [230] presented a combination of redirected walking techniques with visuo-haptic interaction and a path planning algorithm. Haptic feedback directly applied to feet can also influence audiovisual self-motion illusions [259]. Exogenous cues (i.e., any external information coming from the environment) can also play a role in these kind of manipulations. Feng et al. [91] examined the effects, influence and interactions of multi-sensory cues during non-fatiguing walking, including movement of directional wind, footstep vibrations, and footstep sounds, yielding results that evidenced the improvement on user experience and realism when these cues were available.

In some cases, motion is not possible at all, hence it is necessary to generate an external, visual motion. This self-motion illusion is commonly known as vection, and sometimes leads to some postural responses (pursuing a correct vestibular and proprioceptive integration of information). It has been demonstrated that auditory cues increase vection strength in comparison with purely visual cues [164], and that moving sounds enhance circular vection [294]. Moreover, vection may also depend on the environment itself: Meyer et al. [240] explored which factors actually modulate those postural responses, and showed that real and virtual foreground objects serve as static visual, auditory and haptic reference points. Some of the experiments in these works were carried out under rigidly controlled setups and in laboratory conditions, and therefore may not apply to free viewing or other complex conditions. Exploring the effectiveness (or degrading effects) of these insights in more complex scenarios can be an interesting future avenue for research. Finally, the effects of other senses besides auditory and haptics in navigation remain unexplored.

4.8 applications

We have reviewed different aspects of multimodality in VR, as well as crossmodal interactions between the different sensory modalities, together with different achievable effects. A summary of all those benefits that multimodality can lead to in VR can be seen in Table 4.5.

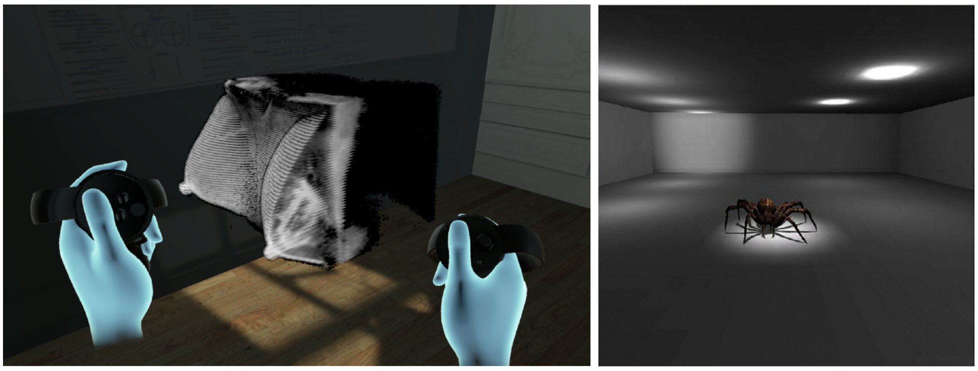


Figure 4.8: Two representative examples of different applications of multimodality in medicine. *Left:* Data visualization and manipulation frameworks [284] are important in medical and surgical education and training, and multimodality may enhance the realism and immersion, thus achieving better learning transfer. *Right:* Multimodal VR is a key tool for phobia treatments, since it is able to create realistic environments that face users against their fears, without actually exposing them [345].

Different disciplines have leveraged these benefits to enhance different VR applications, showing that multimodality can indeed deliver more realistic and immersive VR experiences. While the application scenarios range across many disciplines, here we focus on applications of multimodality to three areas where VR has made a critical impact: medicine, training and education, and entertainment.

4.8.1 *Medicine*

The potential uses of VR for medical applications have been studied for decades, and research in this field has evolved together with the virtual technologies. Satava et al. [373] presented a review about how VR has become an integral technology for medicine both for professionals and patients alike: from medical image visualization and preoperative planning to teaching and simulation, including teleinterventions and rehabilitation.

Other works focused more deeply on the use of VR in the areas of surgical planning, interoperative navigation, and surgical simulations [296, 315]. The possibility of virtualizing a real human body previously scanned and watching it from a far more realistic, immersive perspective than through a conventional display is of great use for health professionals. This has been possible, to a large extent, due to the increasingly photorealistic representation of the anatomy (both in terms of physical tissue properties and of physiological parameters) that virtual environments are achieving.

One of the most pervasive applications of VR in medicine is training, since it can provide a realistic environment for training without the risks of its real counterpart. The enhanced realism and immersion that multimodality provides can lead to improved training and education. Lu et al. [201] presented an audio-visual platform for medical education purposes. One step further, multimodal setups including haptics have been proposed for medical surgery training, where the realism of the feedback significantly improved the learning effect, for both virtual [147] and augmented [125] reality interfaces. In the area of medical visualization, Prange et al. [284] also exploited virtual environments and presented a multimodal medical 3D image system where users could walk freely inside a room and interact with the system by means of speech, and manipulate patients' information with gestures (see Figure 4.8).

Multimodal VR applications are however not constrained to medical training and visualization areas: psychological research relying on VR has also experienced an unprecedented growth, as Wilson and Soranzo reviewed [418], emphasizing both the advantages (e.g., greater control over stimulus presentation, safe exposure to adverse conditions, etc.) and challenges (e.g., VR-induced side effects) of VR in this area. Similarly, Bohil et al. studied the latest advances in VR technology and their applications to neuroscience research [34], highlighting its high compatibility with

Application	Example work	Additional involved senses (other than vision)				Brief description
		Audition	Proprioception	Haptics	Other	
Rehabilitation	Fordell et al. [100]	✗	✓	✓	✗	Chronic neglect treatment, with a force feedback interface.
	Sano et al. [313]	✓	✓	✓	✗	Multimodal sensory feedback to reduce phantom limb.
Phobia treatment	Viaud et al. [401]	✓	✗	✗	✗	Effects of auditory feedback in agoraphobic patients.
	Mülberger et al. [250, 249]	✓	✓	✗	✗	Multimodality short- and long-term effects on fear of flight.
	Hoffman et al. [139]	✓	✗	✓	✗	Illusions of touching to reduce fear of spiders.
OCD therapy	Cipresso et al. [58]	✗	✓	✗	✗	Different instructions to analyze behavioral syndromes.
Medical data visualization	Prange et al. [284]	✓	✓	✓	✗	Visualize and manipulate patients' medical data in 3D.
Surgery training	Hutchins et al. [147]	✓	✗	✓	✗	Medical training simulator with haptic feedback.
	Harders et al. [125]	✗	✗	✓	✗	Medical training simulator with AR features.
Medical education	Lu et al. [201]	✓	✗	✗	✗	Virtual platform to educate on medicine.

Table 4.2: Example works of different medical applications where multimodality plays an important role.

medical imaging technologies (such as functional magnetic resonance imaging - fMRI), which allow for a high degree of ecological validity and control over the therapeutic experience.

Other areas that have leveraged the benefits of multimodality are rehabilitative medicine and psychiatry, where significant progress has been made. Psychiatric therapies can benefit from multimodality, since different aspects of behavioral syndromes can be extensively analyzed in virtual environments: Given the suitability of VR to manipulate the virtual world and control certain tasks, it has proven to be a fitting paradigm to treat diseases like OCD [58] or Parkinson's disease [57].

Phobia treatment is one of the main areas leveraging the benefits of multimodal environments: The realism that multimodality offers over visual-only VR experiences enhances these experiences, and increases the effectiveness of the treatment. In addition, VR allows exposing patients to their fears in a safe and highly controlled way, minimizing any potential risks of exposure therapy. Shiban et al. [345] studied the effect of multiple context exposure on renewal in spider phobia (see Figure 4.8), suggesting that exposure in multiple contexts improves the generalizability of exposure to a new context, therefore helping patients to reduce the chances of future relapses. The work of Hoffman et al. [139] went a step further: they explored not only whether VR exposure therapy reduces fear of spiders, but also concluded that giving patients the illusion of physically touching the virtual spider increases treatment effectiveness. Muhlberger et al. [250] studied the effect of VR in the treatment of fear of flying, exploiting not only visual and acoustic cues, but also proprioceptive information, since motion simulation may increase realism and help induce fear. Later, they studied the long-term effect of the exposure treatment [249], proving its efficacy in treating the fear of flying. The effect of auditory feedback has been studied in other domains, such as the particular case of agoraphobic patients [401], where multimodality increases patients' immersion feeling, hence facilitating emotional responses. However, those techniques should be applied with caution, since large exposures to VR scenarios may hinder patients' ability to distinguish between the real and the virtual world [149], leading to the disorder known as Chronic Alternate-World Disorder (CAWD).

Rehabilitation has also leveraged advances in VR, yielding impressive results. Sano et al. [313] demonstrated that phantom limb pain (which is the sensation of an amputated limb still attached) was reliably reduced when multimodal sensory feedback was included in the VR therapy of patients with brachial plexus avulsion or arm amputation. Fordell et al. [100] presented a treatment method for chronic neglect, where a VR forced feedback interface provided sensorimotor activation in the contra-lesional arm, which combined with visual scanning training, yielded improvements in activities of daily life requiring spatial attention, and an improvement in transfer to real life. Moreover, spatialized sound was also beneficial to improve rehabilitation of postural control dysfunction [409].

Table 4.2 compiles examples leveraging multimodal VR for medical applications. As for the future, VR has the potential to serve medicine even in extreme situations. Virtual care has become an option to foster personalized connections between doctors and patients when in-person appointments are not possible, continuously adapting to the realities of the COVID-19 pandemic [216].

4.8.2 *Education and training*

Training and education are areas in which VR holds great promise, and in which it has already begun to show its capabilities: Jensen and Konradsen presented a review on the use of VR headsets for training and education, and showed that in many cases, better learning transfer can be achieved in this medium compared to traditional media [156].

In education, VR has been widely studied as a new paradigm for teaching: Designing ad-hoc environments helps create adequate scenarios for each learning purpose, hence facilitating the transfer of knowledge to real life scenarios. Stojvsic et al. [365] reviewed the literature on VR applications in education, and conducted a small study showing that teachers perceived benefits in introducing immersive technologies, since students were more motivated and immersed in the topic of interest. At the same time, childhood education processes have been shown to be improved by leveraging multimodality in virtual environments, by means of human-computer interaction methods including feedback and interaction from multiple modalities [56], or somatic interaction (hand gestures and body movements) [92, 8].

Many frameworks regarding VR in teaching and education have been studied and evaluated, demonstrating that using virtual manipulatives (i.e., virtual interaction paradigms) which provide multimodal interactions actually yields richer perceptual experiences than classical methodologies in the cases of e.g., mathematics learning [268] or chemistry education [5]. In the case of the latter, a virtual multimodal laboratory was designed, where the user could perform chemistry experiments like in the real world, through a 3D interaction interface with also audio-visual feedback, which indeed improved the learning capabilities of students.

Similarly, Tang et al. [378] introduced an immersive multimodal virtual environment supporting interactions with 3D deformable models through haptic devices, where not only gestures were replicated but also touching forces were correctly simulated, hence generating realistic scenarios. One step further, Richard et al. [292] surveyed existing works including haptic or olfactory feedback in the field of education, and described a simulation VR platform that provides haptic, olfactory, and auditory feedback, which they tested in various teaching scenarios, demonstrating they affected student engagement and learning positively, and obtaining similar insights as other reviews in educational scenarios, such as in STEAM (science, technology, engineering, arts and mathematics) classrooms [376].

It is worth mentioning that multimodality can also help alleviate sensory impairments, since environments can be designed to maximize the use of the non-affected senses. Following this idea, Yu and Brewster [423] studied the strengths of a multimodal interface (i.e., with speech interactions) against traditional tactile diagrams in conveying information to visually impaired and blind people, showing benefits of this approach in terms of the accuracy obtained by users.

One widespread technique to enhance learning leverages the so-called *serious games*, which enable learning by means of interactive, yet enriching video-games. Checa and Bustillo [52] reviewed the use of immersive VR serious games in this context, and their positive effect on learning processes and transfer. Multimodal VR can actually benefit the learning process of these learning-based serious games [76], since multisensory feedback can enhance many of the cognitive processes involved. Covaci et al. [63] presented a multisensory educational game to investigate how olfactory stimuli could contribute to users' learning experience: It made the experience more enjoyable, but also led to an improvement in users' performance and overall learning.

As aforementioned, multimodal VR has potential in the transfer of knowledge. Given this, it is well suited for simulating and training complex and usually expensive real-life skills requiring high cognitive loads. Gopher [118] highlighted how virtual multimodal training conditions give better results when compared with traditional training conditions in many domains, including

Application	Example work	Additional involved senses (other than vision)				Brief description
		Audition	Proprioception	Haptics	Other	
Education	Christopoulos and Gaitatzes [56]	✓	✓	✗	✗	Children education on history
	Alves et al. [8]	✗	✓	✗	✗	Serious games for children education on history
	Ali et al. [5]	✓	✗	✗	✗	Children education on chemistry
	Tang et al. [378]	✗	✗	✓	✗	Education on deformable materials
	Lu et al. [201]	✓	✗	✗	✗	Education on medicine
	Richard et al. [292]	✓	✗	✓	Olfactory	Education on physics
Accessibility in education	Yu and Brewster [423]	✓	✗	✓	✗	Accessibility for blind people
Serious games	Deng et al. [76]	✗	✗	✓	✗	Review on multimodality for serious games
Skill training	Gopher [118]	✓	✗	✓	✗	Review on multimodality for skill training
	Boud et al. [37]	✗	✗	✓	✗	Skill training for industrial processes
	Crison et al. [65]	✓	✗	✓	✗	Skill training for industrial processes
	Ha et al. [124]	✗	✗	✓	✗	Skill training for virtual prototyping
	MacDonald et al. [204]	✓	✗	✗	✗	Skill training for air traffic control

Table 4.3: Example works of different education and training applications where multimodality plays an important role.

sports, rehabilitation, industry, or surgery; with the latter being the core of Van der Meijden et al's work [394], which reviewed the use of haptic feedback for surgery training, concluding how the addition of this information yields positive assessments in the majority of the cases and even reduce surgical errors. Transferring learning from training simulators to real life situations is one of the most relevant parts of the learning process, and multimodality has been proved to enhance it [182].

In the manufacturing industry, many processes require learning specific skills, and multimodal virtual environments can offer new ways of training. Some works have studied the usability of VR for a manufacturing application such as the assembly of components into a final product, where proprioception and haptic manipulation was required [37]. Other works have proposed a virtual system dedicated to train workers in the use and programming of milling machines, offering visual, audio and haptic (force) feedback [65], also replacing the use of conventional mechanical milling machines. Since fine motor skills can be transferred to the performance of manual tasks, other studies have analyzed the effectiveness of virtual training in the specific case of industry in contrast to real-life training [282]. At the end, the aforementioned works on virtual skill training agree that virtual training could replace real training, since learning is correctly transferred, and the virtual counterparts are usually less expensive and time-consuming. VR is also extremely helpful for assembly and maintenance processes (e.g., virtual prototyping [71]), since it provides a cheap method to directly inspect, interact with, and modify 3D prototypes without the need of a physical industrial manufacturing process [335]. In this context, haptic feedback might be crucial to provide feedback in assembly simulations [124].

Other complex tasks can also benefit from multimodal virtual training. MacDonald et al. [204] focused on the air traffic control problem, and evaluated the relevant aspects of the auditory modality to improve the detection of sonic warnings, including the best design patterns to maximize performance, signal positioning, and optimal distances on the interaural axis depending on the sound amplitudes. Real-time acoustic spatialized simulation can be also used in architecture, when designing acoustic isolation, or studying how sound will be propagated through an indoor environment [404].

All the works mentioned in this section concluded that multimodality offered higher user engagement than unimodal or traditional environments, leading to a better experience and learning transfer. Training in virtual environments has proven to be useful, especially in contexts

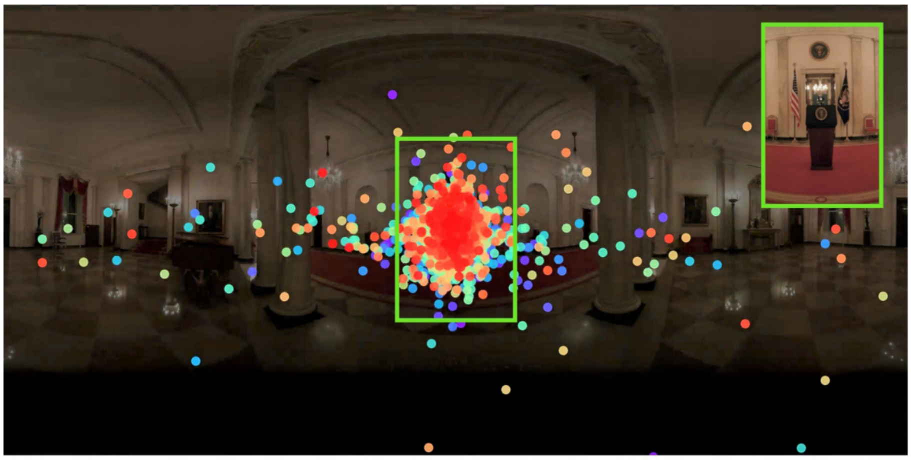


Figure 4.9: Representative image of the work of Marañes et al. [215], where they analyze users' gaze behavior during visualization of VR cinematic content. One of the key open problems in VR is the generation of engaging virtual experiences that meet users' expectations. To that end, it is necessary to understand users' behavior in such virtual experiences.

that are hard or expensive to replicate in real life. On the other hand, and while VR training is effective, the lack of a particular modality (e.g., haptic feedback when learning to manipulate pumps [419]) could diminish the effectiveness of VR with regard to traditional *hands-on* experiences. Hence, it is important to include all the useful sensory information that is needed for each particular experience, and make it as realistic and reliable as possible. A list of some representative applications of multimodal VR in training and education can be found in Table 4.3.

4.8.3 *Entertainment*

Entertainment is undergoing an important revolution with the re-emergence of VR as a new medium: as VR devices become more affordable, their use at consumer level is rapidly increasing. Leisure by means of VR videogames, cinematography, or narrative experiences is becoming increasingly common, and creating realistic, engaging experiences is the main goal of content creators. Multimodality can be instrumental in improving both realism and engagement.

Videogames allow users to interact with a virtual environment, controlling characters or avatars that respond based on their actions. Traditional videogames have leveraged narrative characteristics to connect with the player, to immerse them in the virtual world, so that the experience feels more engaging. With the appearance of VR, immersive games are evolving: higher realism, and stronger feelings of presence and agency can potentially be achieved now with this technology.

Nesbitt and Hoskens [253] hypothesized that integrating information from different senses could assist players in their performance. They evaluated visual, auditory and haptic information combinations, and although no significant performance improvement was achieved, players reported improved immersion, confidence and satisfaction in the multisensory cases. Since haptic devices may enhance the experience, some works have been devoted to developing different toolkits to offer these interactions in VR (e.g., vibrotactile interactions [222]), whilst other works have exploited somatic interactions, including not only haptic but whole proprioceptive cues. Alves et al. [8] studied user experience in games which included hand gestures and body movements, identifying problems and potential uses of gestural interaction devices in an integrated manner Many narrative experiences may require the user to have the feeling of walking, and it may be one of the hardest scenarios to get a realistic response, since multiple sensory information is combined. In this scope, some works investigated the addition of multisensory walking-related cues in locomotion [173], showing that adding auditory cues (i.e., footstep sounds), visual cues (i.e., head motions during walking), and vibrotactile cues (under participants' feet) could all enhance participants' sensation of self-motion (vection) and presence. Sometimes, full locomotion is not permitted, however realism can still be achieved: Colley et al. [61] went a step further in exploiting

Application	Example work	Additional involved senses (other than vision)				Brief description
		Audition	Proprioception	Haptics	Other	
Videogames	Nesbitt et al. [253]	✓	✗	✓	✗	Multimodality to assist players' performance
	Martinez et al. [222]	✗	✗	✓	✗	Vibrotactile toolkit for immersive videogames
	Alves et al. [8]	✗	✓	✗	✗	Serious games for children education on history
Physical activity simulation	Kruijff et al. [173]	✓	✓	✓	✗	Walking simulation for leisure applications
	Colley et al. [61]	✗	✓	✗	✗	Proprioceptive cues to simulate skiing
Cognitive and emotional effects	Kruijff et al. [174]	✓	✗	✓	Olfactory	Study of emotional responses in virtual experiences
	Deng et al. [76]	✓	✗	✓	✗	Cognitive load and processes in serious games
Narrative experiences	Rothe et al.. [301, 300]	✓	✗	✗	✗	Attention guidance in narrative experiences
	Ranasinghe et al. [289]	✓	✗	✓	Olfactory	Enhancing engagement in narrative experiences

Table 4.4: Example works of different applications in entertainment where multimodality plays an important role.

body proprioception, presenting a work that proposed using an HMD in skiing and snowboarding training while the user was on a real slope, so that proprioceptive cues were completely realistic.

Although many of the current VR videogames exploit audiovisual and somatic cues (which are the easiest to provide with current technology), some have tried to work with additional cues. As in previously mentioned learning processes, some works have explored the use of olfactory cues to investigate how enabling olfaction can contribute to users' learning performance, engagement, and quality of experience [63], although this modality still remains in an early exploratory phase.

In a similar manner, gustatory cues have been studied in several works. Arnold et al. [14] presented a game involving eating real food to survive, which combined with the capture and reproduction of chewing sounds increased the realism of the experience. Following this line, Mueller et al. [248] highlighted the potential technologies and designs to support eating as a form of play.

Multisensory feedback can enhance many of the high and complex cognitive processes involved in VR [76]. Particularly, multimodality can trigger different emotional responses in immersive games: Kruijff et al. [174] investigated those emotional effects and proposed guidelines that can be applied to reproduce diverse emotional responses in multimodal games. Within the wide area of entertainment, cinematographic and narrative experiences in VR have been emerging during the last years.

As explored in Section 4.4, guiding users' attention is specially challenging in virtual environments, where users cannot see the whole scenario at once. Although some traditional continuity editing rules may still apply [334], and visual cuts may impact users' behavior [215], the presence of directional sounds can also influence how users explore immersive environments [224], thus special attention must be paid to sound design when considering narrative experiences in VR [119].

To explore how different cues may actually define how users drive their attention in cinematic VR, Rothe et al. [300] investigated implicitly guiding the attention of the viewer by means of lights, movements, and sounds, integrating auditory and visual modalities, while Ranasinghe et al. [289] proposed adding olfactory and haptic (thermal and wind) stimuli to virtual narrative experiences, in order to achieve enhanced sensory engagement. A compilation of representative applications of multimodal VR in entertainment can be found in Table 4.4.

4.9 conclusions

Virtual reality can dramatically change the way we create and consume content in many aspects of our everyday life, including entertainment, training, design and manufacturing, communication, or advertising. In the last years, it has been rapidly growing and evolving as a field, with the thrust of impressive technical innovations in both acquisition and visualization hardware and software. However, if this new medium is going to succeed, it will be based on its ability to create *compelling user experiences*. The interaction between different sensory modalities (such as the five senses, or proprioception) has always been of interest to content creators, but in a VR setting, in which the user is immersed in an alternative reality, the importance of multimodal sensory input plays a more relevant role, since the feedback either from any modality, or from the combination of multiple of them, affects the final experience. In fact, it becomes both a possible liability, if not handled properly, and a potential strength, that if adequately leveraged can boost realism, help direct user attention, or improve user performance. Throughout this survey, we have summarized not only the main lines of research in these areas, but also outlined relevant insights for future directions in each of them.

While making use of multimodal setups can provide benefits to the experience, it also increases costs and complexity. From the point of view of the hardware, however, audiovisual integration is almost always present in current systems (see Table 4.1), and this is also the case for proprioception (except for smartphone-based and related headsets). Most controllers also include some kind of haptic feedback, although in this case it is quite simple and rudimentary, with ample room for improvement and sophistication in consumer-level systems. While research-level technology in haptics is quite advanced, transforming it into consumer-level solutions has been and still is a challenge, due to systems complexity, durability, or cost. We are currently witnessing the first attempts at providing more sophisticated haptic interactions with simpler, consumer-level hardware by using ultrasound; and certainly more advances are to be expected in this area, given the importance of haptics to the multimodal experience. Taste and smell are almost untapped in terms of hardware. Unlike the case of touch, where haptics is abundantly explored at research level, these senses are in their infancy from a research standpoint as well. Thus, special effort should be made towards developing hardware that is able to simulate compelling stimuli for these underexplored senses. From the point of view of the software, inclusion of multimodal input increases the bandwidth and computational resources needed, both current stumbling blocks of VR experiences, particularly collaborative ones. Thus, compression techniques and computational optimizations (both hardware and software-based) are two of the most active areas of research in VR that would also help an increased use of multimodal input. At the same time, works have shown that multimodal input can help maintain realism and immersion with lower quality visual input, so it can also be an advantage in these areas. Additionally, even if it implies an increase in cost and complexity, and depending on the final application scenario, these increased costs may still be more than advantageous if the alternative is setting up a similar, real scenario, in, e.g., emergency or medical training.

The inherent increased complexity resulting from the interaction between sources also poses a challenge for researchers in this area. We have reviewed a number of studies analyzing the interaction of *two* sensory modalities. Most of them were based on constrained experiments under laboratory conditions. However, the final goal of VR is to be present at consumer level, where more complex phenomena and interactions are likely to happen. Thus, lifting constraints on the experimental conditions, and exploring to what extent the insights found generalize and hold in free-viewing scenarios with more confounding factors, remains a critical avenue of future work, which undoubtedly needs to be built on the findings from controlled, constrained experiments. Works exploring three or more modalities are more rare. The integration of input from multiple senses has been an open area of research for over a century, partly because of the curse of dimensionality into which one runs when tackling this problem: the size of the parameter space grows exponentially and soon becomes intractable. Even when the data was available, deriving models to explain it has been a challenge, and analytical models often failed short to explain phenomena outside the particular scenario and parameter space explored, partly because of their lack of generality, partly because the type of data gathered can be very sensitive to the particular experimental setup. Current data-driven approaches certainly provide a new tool to address the problem, and some works have already started to rely on them, as is the case with audiovisual attention modeling. For this to be a solid path forward, however, we need public, carefully-crafted datasets that can be used by the community and in benchmarks, and we need reproducible experimental setups. Incidentally, VR is in itself a great experimental scenario for reproducibility, as opposed to physical, real-world setups.

Effect	How	Sensory modalities (other than visual)				Related works
		Auditory	Haptic	Proprioc.	Other	
Better perception and understanding of the environment (Section 4.3)	Integrating the visual appearance of an object or scene in the VE with a realistic and coherent sound of the same object	✓	✗	✗	✗	[212, 148]
	Correctly spatializing audio in a 360 environment	✓	✗	✗	✗	[246, 146, 405, 120, 110]
	Enhancing realism by touching a physical object and a virtual object at the same time	✗	✓	✓	Olfactory, gustatory	[138, 140]
	Proprioceptive cues (e.g. a hand instead of a pointer) are suitable for interaction scenarios	✗	✗	✓	✗	[280]
Increase in spatial awareness (Section 4.3, 4.7)	Training users to navigate a VE using auditory reflections and reverberation	✓	✗	✗	✗	[11]
	Creating sensory illusions with worn haptic devices that map different points in the VE to the body	✗	✓	✓	✗	[185]
Increase in the feeling of presence, realism or immersion (Section 4.3)	Being able to control a virtual body	✗	✗	✓	✗	[320]
	Body-centered perception (hand, face and trunk) achieved via spatio-temporal multisensory information integration within peripersonal space	✓	✓	✓	Olfactory	[33, 276, 308]
	Achieving partial body ownership considering first person visuo-spatial viewpoint and anatomical similarity	✗	✓	✓	✗	[283, 424]
	Integrating as much multisensory information as possible (spatialized sound, vibrations, wind, real objects, physical movement, scent) coherently with the VE	✓	✓	✓	Olfactory	[116, 157, 131]
	Use of soundscapes or zeitgebers in the virtual scenes	✓	✗	✗	✗	[328, 192]
Self-motion (also related to presence) (Section 4.7)	Adding walking-related multimodal cues, integrating vibrations in standing or even sitting users	✓	✓	✓	✗	[173, 170, 228, 240]
Modification of the saliency of the environment (Section 4.4)	Users tend to pay attention to regions with more sensory information (mostly explored in traditional media).	✓	✗	✗	✗	[69, 160, 70, 252, 80, 242]
	Auditory and visual stimuli are spatio-temporally correlated to increase saliency	✓	✗	✗	✗	[50]
Guidance or direction of attention (Section 4.4)	Using diegetic sound and moving cues to trigger exploratory behavior	✓	✗	✗	✗	[301, 300, 310, 44, 273]
	Spatialized auditory stimuli can also direct attention to specific parts of the environment, including those outside the field of view	✓	✗	✗	✗	[39, 20, 21, 175]
Improvement of user performance (Section 4.5)	Simultaneous stimuli of irrelevant sensory modalities increase detection of target stimulus and can also decrease search times or increase memory retention	✓	✓	✗	Olfactory	[199, 262, 391, 207]
	Auditory stimuli improve spatial awareness and thus reducing search times and improving spatial reasoning	✓	✗	✗	✗	[405, 120, 297, 109]
	Cognitive process starts faster in the presence of multimodality, allowing users to pay attention to more cues and details	✓	✓	✗	✗	[131]
	Combination of audio-haptic tempos to convey spatial information and relevant stimuli increase performance in search tasks	✓	✓	✗	✗	[9]
	Auditory stimuli facilitate touch of virtual objects outside the field of view	✓	✓	✗	✗	[167]
	Improvement of depth perception via simulated touch with a force-feedback system	✗	✓	✗	✗	[38, 372]
	Increasing the space that the user can explore with their body modifying the virtual avatar	✗	✗	✓	✗	[93, 94]
Simulation of physical properties (Section 4.3, 4.5)	Weight or touch simulation by haptic feedback	✗	✓	✓	✗	[46, 311]

Table 4.5: Illustrative examples of different effects that can be achieved through multimodality in VR.

Being aware of how the different sensory inputs interact thus helps researchers and practitioners in the field in two ways: in a first level, it aids them to create believable, successful experiences with the limited hardware and software resources available to them.

At the next level, they can leverage the way the different sensory inputs will interact to overcome some of the limitations imposed by the hardware and software available, and even to improve the *design* of such hardware and software. As multimodal interactions become known and well understood, they can then be leveraged for algorithm design, content generation, or even hardware development, essentially contributing to create better virtual experiences for users, and helping unleash the true potential of this medium.

5

Effects of Spatially Incongruent Auditory Stimuli on Visual Perception

We have seen the potential benefits of multimodality in immersive environments in Chapter 4. However, in Chapter 4 Section 4.6 we have also seen how multimodality can be used to alter how the users perceive themselves or their surroundings through multimodal illusions. Multimodal illusions can be leveraged to increase the quality of user experience and also allow us to better understand underlying cognitive processes. In this Chapter we present an inhibitory audiovisual illusion. When a series of visual targets are presented in a temporally congruent, spatially incongruent manner with a set of auditory cues, visual performance is significantly degraded at detection (is the target perceived, binary response) and recognition (which shape is the target, categorical response) levels. Additional eye tracking data analysis shows the suppressive effect occurs even in the absence of saccades towards the sound source, which suggests the underlying cause must be neural and not oculomotor. This work has been published in Scientific Reports (2020) and presented as a poster in ACM SIGGRAPH 2022.

S. Malpica, A. Serrano, D. Gutierrez, & B. Masia
Auditory stimuli degrade visual performance in virtual reality
Scientific Reports, 2020, vol. 10

5.1 INTRODUCTION

The two most used sensory modalities that help humans perceive extrapersonal space are sight and hearing [396]. While sight is the dominant sensory modality when perceiving the outside world, we rely on hearing to retrieve information for regions of space that we cannot see (i.e., rear space or occluded objects) [361]. The human brain processes the visual information to yield a coherent image. As part of this processing, it has evolved to discard or suppress some of this visual information in order to maintain a stable and congruent vision. This suppression happens consistently: during blinks it usually goes unnoticed thanks to a neural inhibitory mechanism in the brain [403]. For saccades (a quick, simultaneous movement of both eyes between two phases of fixation), our vision remains clear since the blurry images produced by high-speed eye movements are suppressed by the brain [227]. In addition to blinks and saccades, other visual suppression effects exist, triggered by different neural mechanisms [298].

Sensory cortices of different modalities (visual, auditory, etc.) are anatomically separated. However, several studies show that a multimodal interplay exists even between primary sensory cortices [379]. In particular, crossmodal inhibitory interactions have been found in humans for auditory and visual modalities [135], and for tactile and visual modalities [136, 150]. Several brain imaging studies have also shown crossmodal inhibitory or modulatory cortical responses [183, 158, 239, 151] in what were previously considered unimodal processing areas. The areas where neural suppression occurs are also identified, including parts of the sensory cortices. A deep and comprehensive understanding of crossmodal effects can be leveraged for applications beyond vision science, such as visual computing, immersive environments, or the design of novel display hardware.

In this work, we focus on how sound can degrade visual performance in VR. Despite the recent success of this emerging technology, the viewing behavior and mechanisms triggered by this new medium are not yet well understood [353]. Specifically, we investigate whether the presence of an auditory stimulus can degrade the detection and recognition of visual stimuli that appear in a temporally congruent manner with the auditory stimulus, as compared to the performance in the presence of the visual stimuli only (i.e., without an associated, temporally congruent auditory stimulus). Facilitating effects with spatially congruent modalities have been assessed in several studies [187, 233, 396]. We thus choose to present sounds in a spatially incongruent manner with the visual stimuli. Moreover, previous work has shown that crossmodal interactions taking place in rear space are often different from those in frontal space, since in rear space we have to rely on sounds to obtain information that cannot be retrieved visually [361]. Hidaka and Ide [135] showed that white noise bursts can degrade visual performance significantly in laboratory conditions; they used a fixed-head experiment setup in which the visual stimuli were tilted Gabor patches displayed on a conventional monitor. We differ from this previous work in several aspects, which aim to increase our knowledge of the phenomenon, generalize the findings, and bring them closer to their potential application scenarios.

First, we extend the analysis from low-level Gabor patches to realistic environments, including a task of higher cognitive load: We broaden the task from binary recognition in previous work, to detection *and* recognition of five possible visual targets. Crossmodal effects are still present in such higher cognitive load conditions. For example, it has been demonstrated that auditory spatialization can facilitate speech recognition [304]. When stimuli of different modalities are presented in a temporally congruent manner, it has also been shown that attention can be selectively diverted from a target to a secondary speaker [117]. Crossmodal effects in VR can even help creating illusions of different sensory modalities [348]. Second, we explore a wider range of different sounds with varying complexity. While previous work used only white noise, we also analyze pure frequencies, pink and brown noise, and two different sounds with semantic content for a total of six different sound types. Beyond their characteristics (i.e., frequency content), these sounds have been chosen due to how they affect perceptual processes; more details on this can be found in the Methods section. In addition, we explore the influence of the type and spatial location of sound, as well as its interaction with the shape and spatial location of the visual target. Further, instead of using a regular monitor, we conduct our experiments in an immersive VR setting. The reason for this is three-fold: First, VR offers a greater control over the conditions of the experiment, increasing reproducibility and repeatability; second, it allows for a more natural exploratory behavior of the subject, including walking around the scene, in contrast with previous approaches that required a fixed head position; and third, auditory-triggered visual performance degradation can find key applications in VR, where control of the user's attention is a fundamental challenge [299]. Moreover, it has been assumed in the literature that the visual performance degradation is caused by neural inhibitory interactions between the auditory and visual sensory pathways. However, it remains unknown if saccades towards the sound source (and hence saccadic suppression) are related to this effect. To explore this, we record gaze data by means of an eye tracker built into the head mounted display (HMD) and analyze gaze behavior during the experiment. We will make the data and stimuli available for reproducibility and further analyses.

Our main findings include:

- We find that the visual performance degradation effect is robust even for viewing conditions that impose a higher cognitive load, including natural exploratory behavior. This is important since these factors could potentially affect or mask the inhibitory effects reported in the literature.

- We find a consistent and significant degradation of both detection and recognition of the visual targets regardless of both sound location and the location or shape of the visual target.

- Our gaze data reveals that gaze behavior does not change even in conditions where visual performance decreases significantly, suggesting that the effect is not caused by oculomotor phenomena.

5.2 METHODS

Our experiments were designed with a two-level task. Participants explored an indoor scene in VR. They had to detect (binary response, a visual target was seen) and recognize (categorical response, after detection the shape of the visual target was identify) a set of white and plain simple geometrical shapes. A visual only baseline experiment (Experiment 0) indicates that the visual targets can be perceived in the absence of sound. When sounds are introduced (Experiment 1) there is a significant drop in the performance of the detection and recognition tasks.

Participants and apparatus. Fifty-six participants took part in the experiments described in this work. Seven of them in the baseline experiment (Exp. 0), and 49 in the main experiment (Exp. 1). The mean age was 24 years (±3.21). Twenty of them were women. All of them had normal or corrected-to-normal vision and did not report hearing problems. Participants were not aware of the experiment's goal. The visual and auditory stimuli were presented through an HTC Vive Pro VR headset with built-in headphones and a nominal field of view of 110 degrees (1440×1600 pixels resolution per eye and a framerate of 90fps). A single computer was used, with an Intel i7-7700 processor at 3.6GHz and 16GB of RAM. The graphics card was an Nvidia 1060GTX (6GB of dedicated DDR5 memory). All the scenes were created using Unity 3D (2018 version) with the Vive VR plug-in on Windows 10. The VR headset included a Pupil-Labs eye tracker. This add-on eye tracker was used to record the participant's gaze behavior through the experiment at 120Hz, with accuracy of one degree of visual angle.

Participants in Exp. 1 were presented with 18 audiovisual (*biCond*) stimuli, 18 visual-only (*visCond*) stimuli and 18 auditory-only stimuli. Participants in the baseline experiment (Exp. 0) were presented with 50 visual-only stimuli. Visual-only stimuli were the same for all participants in their respective experiments, while auditory-only stimuli were randomly chosen in Exp. 1. The auditory part of *biCond* stimuli was the same for all participants: Six different sounds in three possible locations each. The presentation of the different stimuli was randomized across participants to avoid order effects both in Exp. 0 and Exp. 1. We follow a conservative approach and consider for the analysis those participants with good detection and recognition percentages in *visCond* stimuli, setting a minimum detection and recognition threshold of 33% and 20%, respectively. As a result, only five participants were rejected from Exp. 1; their data was not considered for the analysis presented in the Results section.

Visual stimuli (targets). The visual targets consisted of five simple geometric white shapes with a gray outline, as shown in Figure 5.1. In order of increasing complexity: circle, square, rhombus, pentagon and star. They were chosen not to have any semantic meaning compared to the visual background scene. The target size is one degree of visual angle. Visual targets remain for 24ms in the participant's FOV. Both the target size and its duration had been fixed following Hidaka and Ide's work [135]. In our experiment, the target could appear randomly at one of three different locations, always at the same latitude (FOV equator line): FOV center, four degrees of visual angle to the left or four degrees of visual angle to the right of it. These stimuli were used both in Exp. 1 and in the baseline experiment, where their visibility was assessed. Visual-only (*visCond*) stimuli were maintained in Exp. 1 as sentinels.

Auditory stimuli. Auditory stimuli included six different sounds inspired by previous literature. *Pure frequency:* We are not used to pure frequency sounds in nature [361]. Being less common, this sound could deviate the participant's attention from the visual stimuli. *White noise:* This is the sound used by Hidaka and Ide [135]. It has proven to degrade performance in visual recognition tasks in traditional displays. *Brown noise:* Random changes between tones can stand out from uniform noises [361]. *Pink noise:* Pink noise is known to trigger an acoustic reflex response that protects the eardrum from loud noises [64]. Given the relationship between visual and auditory neural processing, we hypothesized that pink noise could also have an inhibitory effect on visual stimuli. *Survival sound:* Critical sounds for our survival also stand out, especially if they come from outside our FOV [361]. In particular we used a train horn in Exp. 1. *Human voice:* It has been shown that human voices draw our attention powerfully [361]. The duration of each sound was 400ms, to allow for the more complex sounds to play completely. Sounds were spatially located at random in one of three possible locations, always at 0.2m (Unity distance) from the head: directly behind the participant's head, shifted to the right (50 degrees rotation from the center of the head's position) or to the left (also 50 degrees), always outside their FOV. Auditory-only stimuli served as distractors, to avoid an association of the visual target appearance with the auditory stimuli onset.

Audiovisual stimuli. Audiovisual stimuli were created by presenting simultaneously an auditory stimulus and a target. As shown in Figure 5.2 B, the auditory stimuli start playing 100ms before the visual stimuli onset. Every participant was presented with the six possible sounds in the three described locations, making a total of 18 different audiovisual stimuli. The visual part of each stimuli was chosen pseudo-randomly (as close to a uniform distribution as possible) across participants. Figure 5.2 A shows all the possible locations of the bimodal stimuli.

Procedure. The baseline experiment (Experiment 0) procedure was designed with visual stimuli only. Its main purpose is establishing the performance of visual targets in the absence of auditory stimuli. Meanwhile, Experiment 1 presents both visual and auditory stimuli to study how auditory cues can affect visual perception. The participants were located inside a virtual scene that resembled a living room, shown in Figure 5.1. They could freely move in a physical space of 4x1.5m with a 1:1 mapping between the real and virtual spaces. Before starting the experiment, the participants were shown the same room without furniture so that they got used to the VR headset and the VE. Participants were informed of their task until they declared they had understood. Simple geometric shapes would appear and disappear in front of them randomly throughout the experiment; each time they detected one such shape, they had to notify the experimenter. The experimenter would then show them a question within the VE: *What did you see?* When the participant answered, the experimenter would log the answer and the experiment resumed. No new stimuli appeared until the participant had answered the question. This was an open-ended task, as the participants did not know *a priori* what specific shapes could appear during the experiment. If the participant **detected** the onset of a visual stimulus but did not **recognize** its specific shape they still had to notify it. The participants were also told that they would hear several sounds through the experiment, but that they had to stay focused on the appearance of the visual target. There was an additional background sound played throughout the whole experiment: the sound of a park through an open window and a news podcast that played through one of the speakers near the TV. The intention of this background sound was to increase the scene complexity and realism, as well as to avoid for the auditory stimuli being the only sounds in the scene.

Throughout the experiment, the three different types of stimuli (visual, auditory and audiovisual) appeared in random order with a random in-between interval that varied from five to ten seconds. The experiment took 15 to 20 minutes, including the initial explanation and the questionnaires that the participants filled before and after the experiment. The participants were informed to stop the experimenter if they felt any kind of sickness or discomfort during the experiment (none did). Before they started to use the VR headset, participants filled a questionnaire with sociodemographic questions including age, gender, and previous experience with VR. None of the sociodemographic factors had an influence on the obtained results. After the experiment had concluded, there was a short debriefing in which they filled a set of questions about the experiment (*Did you see or hear something remarkable?*, *Did you feel any discomfort?*, *Do you want to say something else about the experiment?*). None of the participants experienced sickness or discomfort after the experiment. Six of them reported either the train horn or the human voice were surprising at least the first time they appeared in the experiment. Nine found the task *interesting* or *engaging*.

Statistical analysis. A GLM assumes that the measured data samples are independent. In our case, we cannot assume that the samples are independent, since each participant was measured several times under different conditions. Using a GLMM we can account for mixed effects, and therefore account at the same time for both the fixed effects of our variables and the random effect corresponding to user variability. The dependent variable was binary (for detection) or categorical (for recognition). The independent variables in both cases were the visual target shape, the visual target location, the sound type, and the sound location; they were set as fixed effects. Different participants (in particular, the recorded subject ID) were considered as random effects. We used *Matlab* `fitglme` function with a `logit` link function.

5.3 RESULTS

5.3.1 *Experiment 0 (baseline): Visual detection and recognition in the absence of auditory stimuli*

To ensure that the visual targets were detectable and recognizable we first ran an experiment in the absence of concurrent auditory stimuli. The 360° virtual environment (VE) displayed in the VR headset showed a realistic living room as shown in Figure 5.1. The visual targets were five different simple shapes (circle, square, rhombus, pentagon, and a five-pointed star), placed at one of three possible locations inside the field of view (FOV) of the subject: center, four degrees to the left, or four degrees to the right, always on the FOV equator (see Figure 5.2 A, green area). All stimuli were white with a grey outline to help differentiate them from the VE. The subtended visual angle of each visual target was one degree.

A background auditory context was added, consisting on diegetic, localized audio (sounds from a park coming through the window, and a news podcast playing through one of the speakers near the TV). Each visual target appeared for 24ms [135] and the interval between targets was randomly chosen between five and ten seconds to prevent potential learning effects. The participants had to verbally report each time they saw a visual target, and specify its shape. They were explicitly told to notify the appearance of a target even if they could not recognize its shape. Each participant saw a total of 50 targets. The mean percentage of target **detection** (*binary response; the participant was able to identify the appearance of a target*) was 88.10% (±4.20%, 2*SEM). Wilcoxon tests were used to check for differences between experiments or between conditions (pairwise comparisons), while GLMM models were used to analyze the influence of the studied factors in the detection and recognition tasks. More details can be found in the Methods Section. All the GLMM results can be found in the supplementary material of our work [213]. We establish the significance level at $p = 0.05$. Neither the shape nor the location of the target had a significant influence on detection. The mean percentage of target **recognition** (*one of five possible responses; the participant could distinguish the shape of the stimulus*) for detected stimuli is 71.96% (±12.36%). Recognition is calculated relatively to the detection percentage. A percentage of 100% recognition means that all detected visual targets have been correctly recognized. Different from detection, shape had a significant influence ($\beta = -0.311$, $t(293) = -3.324$, $p = 0.001$) in recognition, with post-hoc Wilcoxon pairwise tests revealing that star shapes where better recognized. This may be related to the increased geometrical complexity of the star, which is the only non-convex shape in the stimuli.

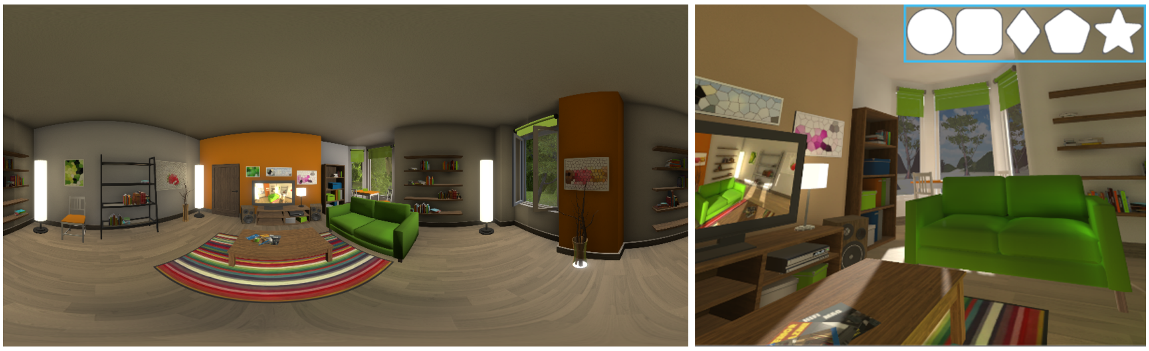


Figure 5.1: *Left:* 360° panorama of the virtual environment used in the experiments, rendered from the central point of view. *Right:* Representative close-up view of the VE. The inset shows the five different visual targets (a 3.2x scale is used here for visualization purposes). Users could move freely in a physical space of $4 \times 1.5m$. Scene by Barking Dog for Unity 3D.

5.3.2 *Experiment one: Visual detection and recognition in the presence of temporally coherent auditory stimuli*

In this experiment, each trial contained a single stimulus, which could be auditory-only, visual-only or bimodal (audiovisual). The background noise used in the baseline experiment was always present. A total of 54 stimuli were presented to each participant. Eighteen of them were visual-only, and followed the characteristics of the baseline experiment (we term them *visCond*). These stimuli also served as sentinels to make sure that all the participants had a good performance on detection and recognition tasks in the absence of confounding auditory stimuli. Another 18 stimuli were auditory-only, acting as distractors to make sure that participants would not expect a visual target to always appear in the presence of a sound. The last 18 stimuli were bimodal (*biCond*); these stimuli included both a visual target as in the baseline experiment, and a sound. Figure 5.2 illustrates the spatial and temporal layout of the experiment. No participant reported target detections in the auditory-only condition; in the following, we thus analyze the *visCond* and *biCond* conditions.

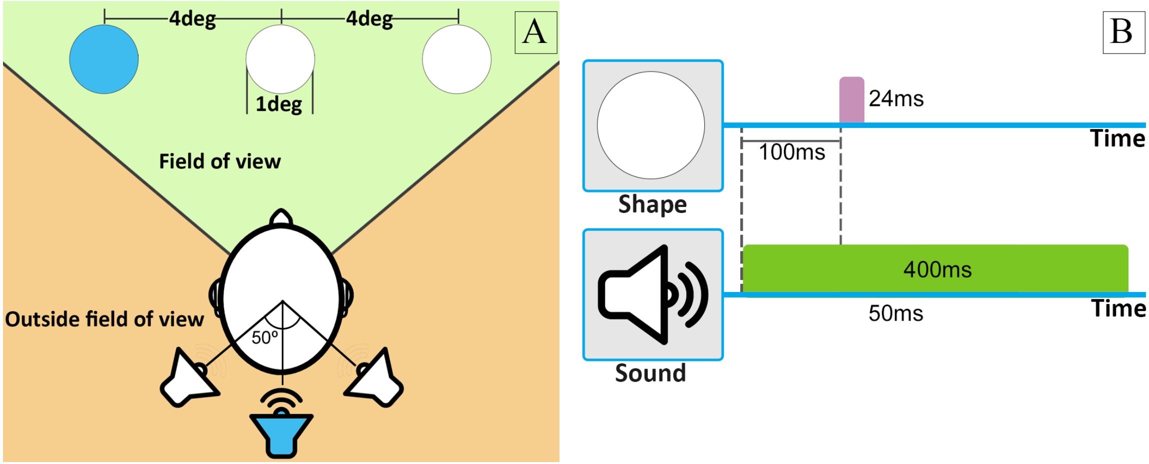


Figure 5.2: *A*: Spatial layout of the experiments. The three possible locations of the visual targets inside the participant's FOV: center, and four degrees of visual angle to the left and right. The targets subtended a visual angle of one degree. Auditory stimuli were spatially located outside the FOV, also in one of three possible locations, all of them 0.2m (Unity distance) from the user position: behind the participant, 50 degrees left or 50 degrees right. Both the visual targets and the auditory stimuli kept their positions fixed relative to the participant's head. One possible combination for a *biCond* stimuli is highlighted in blue. *B*: Temporal layout. Visual targets are shown 100ms after the sound starts, for a duration of 24ms. The auditory stimulus lasts 400ms. Gaze behavior is quantitatively analyzed in those 400ms to study the relationship between the presence of sound and the visual performance degradation effect.

Influence of sound in detection and recognition. For the visual-only stimuli (*visCond*), the mean percentage of detection is 82.07% (±4.81%), similar to the results from the baseline experiment. Adding sound (*biCond*) results in a large drop, yielding a mean percentage of detection of just 20.02% (±4.86%). Similarly, recognition for *visCond* is 59.93% (±6.76%), decreasing to only to 7.93% (±4.12%) for *biCond*. This is shown in Figure 5.3. A Wilcoxon signed rank test ($z = 5.783$, $p < 0.001$ for detection; $z = 5.777$, $p < 0.001$ for recognition) shows that both conditions are significantly different both in stimuli detection and recognition. In particular, we find a **decrease of both detection and recognition for *biCond* stimuli in relation to *visCond* stimuli**. A Wilcoxon rank sum test shows a significant difference ($z = 7.919$, $p < 0.001$) between *biCond* and the baseline experiment for both detection and recognition. Recognition drops from 71.96% to 59.93% for *visCond* compared to the baseline results. We hypothesize that this may be due to the greater cognitive load imposed on the participants, being exposed to three different stimuli conditions.

Effect of the different factors on detection and recognition. Here we analyze the influence of the different factors of the experiment (target location, target shape, sound location, and type of sound) on the detection and recognition tasks. As in the baseline experiment, the location of the visual target does not have a significant influence on detection nor recognition. The sound location does not have any significant influence either. Target shape has a significant influence ($\beta = -0.249$, $t(787) = -4.266$, $p < 0.001$) only for *visCond* during recognition tasks, but not for *biCond*. The type of sound in the experiment has a significant influence ($\beta = 0.658$, $t(613) = 1.481$, $p = 0.048$) on stimuli recognition (see Figure 5.3). We found anecdotal evidence that pink and white noise had dominant effects in the degradation of visual performance, although these are not significant. A deeper study about which types of sound or which particular features (e.g., frequency content [241]) may have a deeper impact on visual performance degradation would be an interesting line for future work.

Analysis of gaze data. Auditory stimuli have the potential to trigger visual saccades [104]. Here we investigate saccades as a possible cause for the visual performance degradation effect. In particular, even though participants were explicitly told to ignore the sounds and focus on the visual targets, it is still possible that the auditory stimuli in the *biCond* condition were inducing a saccadic suppression effect, preventing the visual target from being seen. To analyze this, we leverage the data collected through the eye tracker and analyze gaze behavior around the visual target onset, focusing on the differences between *visCond* and *biCond* stimuli. However, accurate saccade detection is challenging, especially in our case where participants are allowed to move while wearing the VR headset. We thus study the differences in fixation rates between *visCond* and *biCond* stimuli as a more robust way of analyzing gaze behavior. We calculate fixation rates using fixation detection by two-means clustering [134], which is robust in the presence of noise. We take into account a two-second window centered around the visual target onset, and a region of interest of ten visual degrees [27] around the position of the visual target (as shown in Figure 5.4). We find that each participant fixates in that region 50.24% of the time on average in the *visCond* condition, and 49.13% in the *biCond* condition, with no significant difference between conditions ($z = 0.671$, $p = 0.502$, Wilcoxon signed rank test). If we reduce this window to the 400ms around the visual target onset (the same 400ms where sound is present in the *biCond* condition, as shown in Figure 5.2), there is no significant difference either (72.40% vs 72.38% of the time on average, $z = 0.933$, $p = 0.3507$, Wilcoxon signed rank test). This suggests that the auditory part of *biCond* stimuli does not cause a significant change in gaze behavior. In particular, if saccadic suppression (a saccade triggered towards the sound source) was the underlying cause of the visual performance degradation, we would have expected to find a change in gaze behavior between *visCond* and *biCond*, with maybe a decrease of fixation time in the latter condition. In contrast, participants fixate similarly regardless of the presence of sound, while their visual performance varies significantly between *visCond* and *biCond*. This is confirmed by a qualitative analysis of gaze behavior, an example of which can be seen in Figure 5.4. Visual performance degradation happens even when gaze is fixated close to the target location at its onset. Therefore, we believe that the degradation effect is not caused by oculomotor phenomena.

5.4 DISCUSSION

Interactions between the human visual and auditory systems are complex and not completely understood yet. Frens et al. [105] showed that an auditory stimulus can improve performance of visual search tasks. At the same time, stimuli of one modality can alter [74, 337] or even suppress [135] the perception of stimuli of another modality. Inspired by these works, we have investigated the auditory-triggered visual performance degradation effect under immersive and realistic viewing conditions, including natural exploratory behavior. We have verified that this crossmodal, sound-induced visual inhibitory effects exist in VR. In particular, we found that the effect is robust to different sound types, sound locations, as well as varying visual target shapes and locations along the FOV equator. The used VE also imposes a higher cognitive load on participants when compared to previous work. Even then, the degradation effect is robust to these potentially masking effects. Given that the visual degradation is robust to modifications of the four factors studied in this work, we hypothesize that the mechanism responsible for the

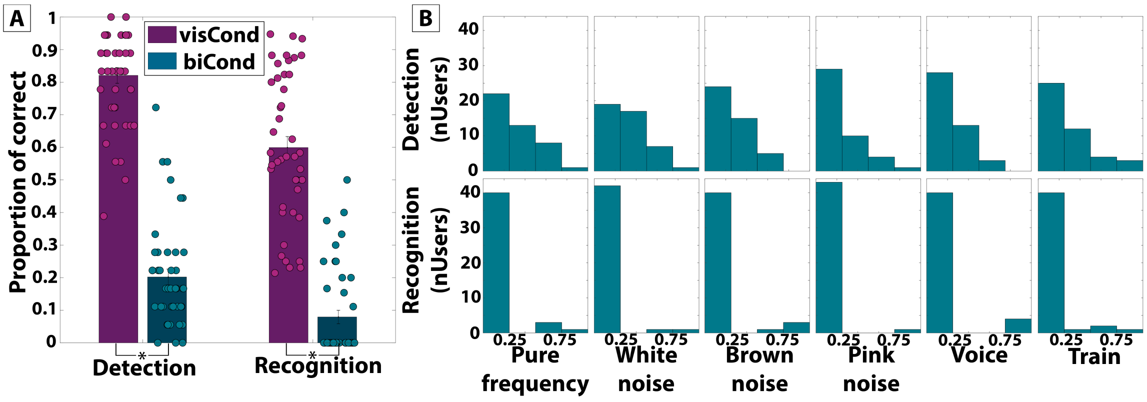


Figure 5.3: *A:* Mean detection and recognition for *visCond* and *biCond* conditions. Error bars show 2*SEM. Both detection and recognition are significantly lower for biCond stimuli (marked with an asterisk). Individual performance is shown as scattered points over the bars. *B:* Mean detection (top row) and recognition (bottom row) histograms, by type of sound. Y-axes show the number of participants with a given performance rate (X-axes, from 0 to 1, divided in 4 bins). Note that a 0 detection rate for a given sound type implies a 0 recognition rate for the same sound type. TODO: increase font sizes. Remove B part?

degradation effect does not depend on the particular characteristics of the sound or the visual target to be inhibited, but rather encompasses a larger aspect of sensory perception.

We find a consistent and significant degradation of detection and recognition of visual targets in the presence of temporally congruent auditory stimuli. Recognition further decreases for pink and white noise. In their experiments about inhibitory audiovisual interactions, Hidaka and Ide[135] used white noise bursts concurrent with the onset of their visual targets, under carefully controlled laboratory conditions including a chin rest and a conventional display. The authors reported a performance drop from 70.7% with visual-only stimuli to approximately 60% in the presence of sound. Our results show the same trend, and further suggest that, in the presence of higher cognitive loads, the degradation effect is more prominent, across a wide range of audiovisual stimuli. There is an important difference in the magnitude of the effect observed between our work and Hidaka and Ide's. Possible reasons include the increased realism of our experiment, the increased complexity of the task (which includes binary detection as well as an additional five-level recognition task), and the fact that users were able to move during the experiment, all these may lead to higher cognitive loads. Besides cognitive load, the fact that sound is spatialized inside the virtual environment, and presented in rear space (in a spatially incongruent manner with respect to the visual targets) might further influence the effect magnitude. Additionally, Hidaka and Ide report a bigger effect when sounds were presented in an ipsilateral, spatially congruent manner. We did not find this effect with binaural spatialized sounds. Hence, their findings might be related with monoaural sounds rather than with the spatial congruency of visual and auditory stimuli.

We chose to use target stimuli that were not semantically related to the background scene, both in its visual and auditory aspects. We took a conservative approach, and designed the target visual stimuli as simple, white geometrical shapes that clearly stand out from the rest of the scene, to minimize the risk of fortuitous oversights. More contextually integrated visual stimuli may have lower detection percentages when compared to the visual targets used in this experiment.

Our analysis of gaze behavior shows that visual degradation occurs even in the presence of fixations and with gaze near or at the location of the visual target. Traditionally, sound has proven to increase performance of visual related tasks. For example, Corneil et al. [62] show that saccades triggered by audiovisual stimuli have faster reaction times than those triggered by visual-only stimuli. However, other studies have also reported both facilitatory and inhibitory responses of audiovisual inputs, mostly depending on the spatiotemporal congruency of both modalities [187, 141]. The more congruent the different modalities of the input stimuli presented are, the easier a facilitative integration will occur. On the other hand, if the stimuli are spatially or temporally incongruent, an inhibitory effect is more likely to occur. In our experiment, the

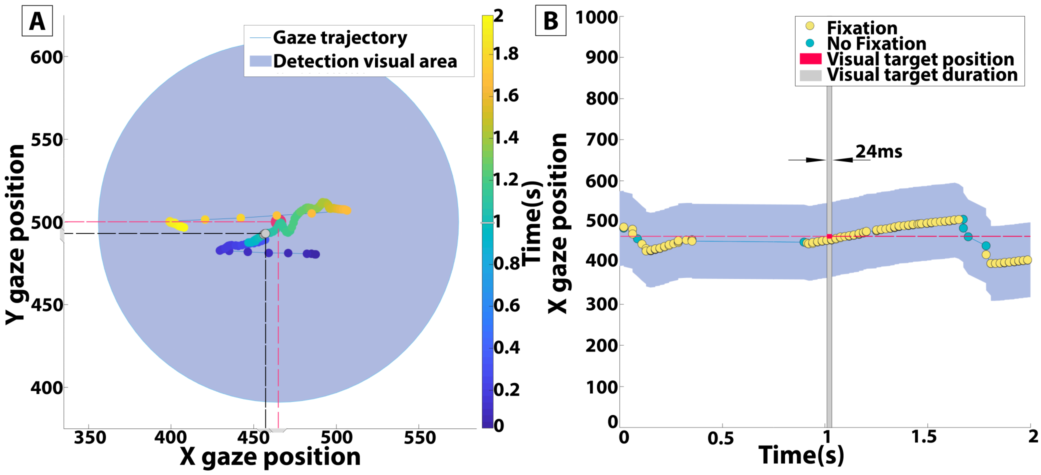


Figure 5.4: *A:* Illustrative gaze behavior during presentation of a visual target that was not detected, corresponding to a *biCond* stimulus. The colored points represent gaze position during a two-second window centered around the visual target onset. Target onset thus occurs at time $t = 1s$, and its spatial position is marked with a red point and associated red dashed lines. The blue-shaded area represents a region of ten degrees of visual angle [27] centered around the target. Gaze positions at the interval during which the target is present is marked with gray points and an associated black dashed line. Note that, despite gaze being very close to the target location during its onset, there is no detection. *B:* 1D visualization of gaze position (only x coordinate) over time during the same trial presented in *A*. Gaze samples corresponding to fixations are shown as yellow points, other gaze samples as blue points, the temporal interval during which the target is present is marked in gray, and the target position in x is marked by a dashed red horizontal line.

visual and the auditory modalities of the stimuli (in *biCond*) were always presented in a temporally congruent and spatially incongruent manner. As to what is the underlying cause of the visual performance degradation effect, there are several possibilities, including oculomotor, neural and attentional effects. Our analysis of gaze behavior suggests that this phenomenon does not seem to be related to oculomotor effects. One possible explanation is that the auditory stimuli (a salient exogenous cue presented slightly before the visual target) is causing an involuntary shift of attention [359, 360]. This attentional shift, either spatial [203] or modal [362], might in turn result in the degradation of visual performance or crossmodal deactivation of the visual input [247]. Note that in Hidaka and Ide's work [135] crossmodal attentional effects could not fully explain their findings, since the degradation effect was still present when the auditory stimuli were shown after the visual target. The authors concluded that the effect occurred based on neural interactions among auditory and visual modalities. One of the key differences in our experiment is that the auditory part of *biCond* stimuli is always shown 100ms before the visual target onset, which may cause auditory stimuli to compete with the processing of visual stimuli [158]. Further studies are necessary in order to determine what is the exact cause behind the observed effect for both experimental conditions.

Besides increasing knowledge about the human visual system, leveraging visual performance degradation can also entail a direct benefit for several applications [375, 19, 12, 32]. In particular, VR technology still faces challenging limitations that could be addressed with a deeper understanding of multimodal human perception. For instance, visual suppression has been used in conjunction with the *change blindness* phenomenon [351] to introduce changes in the virtual world that go unnoticed by the users, allowing them to avoid obstacles in the physical world [369, 35]. In general, a better understanding of the interplay of the different sensory modalities will lead to improved user experiences [131]. Apart from novel applications, we hope that our work can also motivate additional experiments to further study the scope of the visual performance degradation effect.

We have shown how it affects both detection and recognition of a flashing visual target. Does it also affect the perceived motion of a dynamic visual target? Can we integrate inhibitory effects from different sensory modalities? It would also be interesting to analyze other sound properties: can we make the sound barely (if at all) noticeable while still degrading visual performance? Modeling and extending the parameter space of sounds that degrade visual perception might also give us some additional insights on the underlying perceptual mechanisms at work.

6 Effects of Auditory Stimuli on Material Perception

In this chapter we study how a correctly synchronized multimodal cue can increase perceived realism and change how materials are perceived in immersive environments. First, we make sure that crossmodal effects (interactions between two or more different sensory modalities) are correctly perceived in VR. We do so with a series of experiments replicating in VR a well-known crossmodal audiovisual effect. Then, in a context of material perception in virtual environments, we show how the use of crossmodal audiovisual cues can increase perceived realism and disambiguate material perception even when visual quality is degraded. This work has been published in Multimedia Tools and Applications.

S. Malpica*, A. Serrano*, M. Allue, M. Bedia, & B. Masia
Crossmodal perception in virtual reality
Multimedia Tools and Applications 2020, 79(5)
$*$ Joint first authors

6.1 INTRODUCTION

During the last years, we are witnessing a reappearance of virtual reality (VR). New applications are developed every day, going far beyond entertainment and gaming, and including advertising [368], virtual tourism [123], prototyping [335], medicine [181], scientific visualization [171], or education [336], to name a few. There are still important stumbling blocks that hinder the development of more applications and reduce the visual quality of the results; examples include limited spatial resolution, chromatic aberrations, tracking issues, limited processing capability leading to lag, subsequent motion sickness, or content generation [397]. A relevant area which has received quite some interest but remains full of unanswered questions and open problems is how our perception is modified or altered when immersed in a virtual environment. Knowledge of human perception in virtual environments can help overcome the aforementioned current limitations. In the past, perception has been leveraged in many computer graphics-related areas such as rendering [287], material modeling and acquisition [350], or display [225]; a good review of applied perception in graphics can be found in the course by McNamara and colleagues [237].

In this work, within the much-studied area of perception in virtual environments, we chose to look into the less explored area of crossmodal perception in HMDs, that is, the interaction of different senses when perceiving a virtual environment through a headset. HMDs are different from traditional displays in that they provide a more realistic and immersive experience, as well as introducing additional degrees of freedom (the user now controls the camera), spatialized sound, increased field of view, and more visual cues (e.g., motion parallax). Specifically, we looked at the *influence of sound on visual perception in a virtual reality scenario*.

Crossmodal perception, and in particular the interaction between visual and auditory stimuli, has been studied before in real scenes and on conventional displays. The crossmodal effect between these two sensory inputs has been assessed and documented in different works [327, 346, 341], which state, among other conclusions, that the presence of sound can alter the visual perception.

This work is an extension of our previous work [7], where we replicated a well-known crossmodal perception experiment [327]. We found that crossmodal interaction was indeed present in

VR, and that its effects persisted even in the presence of more complex stimuli. These experiments are described in Section 6.2. We further extend this initial work by, once we have asserted the presence of a visual-auditory crossmodal effect, analyzing the effects of sound in the visual perception of materials, in order to find practical applications for VR. This new experiment is described in Section 6.3 and constitutes the main contribution of the present work. Generating content for VR headsets requires rendering complex scenes in real time, at high resolution and, ideally, at least 60 fps, which comes at a large computational cost, specially if the aim is to obtain a realistic appearance. Different works have investigated how visual perception is affected in VR, partly with the aim of reducing this rendering cost [31, 270]; conversely, other works have analyzed the effect of sound in material perception, but not in an immersive environment [36, 221]. In this work we have taken the first steps towards analyzing the influence of a visual-auditory effect on material perception in VR (Section 6.3), providing insights that can be used in the future to reduce computational costs, or improve the quality when rendering complex appearances. In particular, the research questions we investigate in this work are the following:

- Manifestation of the crossmodal effect in VR environments with increasing complexity.
- Influence of crossmodal interactions in material perception in immersible VR environments.

6.2 crossmodal interaction

We have first performed two experiments in order to determine how much an immersive environment interferes with the crossmodal interaction between the visual and auditive systems. Our experiments are based in the work of Sekuler et al. [327], where they explore the perceptual consequences of sound altering visual motion perception. In their experiments, they showed two identical disks that moved steadily towards each other, coincided, and then continued in the same direction. This scenario is consistent with two different interpretations: either the two spheres did not collide and continued in their original directions (they *streamed*), or they collided and *bounced*, changing their traveling direction. The goal of the experiment is to analyze whether a sound at the moment of the impact can affect the interpretation of the scenario.

We build upon Sekuler et al.'s work, and extend their experiment to virtual reality, aiming to explore the consequences on crossmodal interactions of introducing the user inside a more realistic and complex environment presented with a *head mounted display* (HMD).

6.2.1 *Experiment 1*

Goal. We first reproduce the experiment described in Sekuler et al.'s work both in a regular screen and in a HMD (*Oculus Rift DK2*). The goal of this experiment was to test whether the effect of sound altering visual motion perception as reported in the experiments carried out by Sekuler et al. is also observed when reproduced in a virtual environment with an HMD.

Stimuli. The visual stimuli were rendered with *Unity*. They consisted of two spheres with radius *0.5 degrees*, placed over a white plane. The material of the spheres was brown and very diffuse to avoid introducing additional visual cues. The two spheres were initially separated by a distance of *4.2 degrees*, and moved towards each other at a constant speed of *6 degrees per second*. After they coincided, they continued moving without changing their original direction. We show in Figure 6.1 the initial layout of the scene. In this scenario we presented three different visual conditions: the spheres moved continuously, paused one frame at the point of their coincidence, or paused two frames at the point of their coincidence. The original experiment [327] reported frames in a regular analog screen whose typical framerate is 25 frames per second. Since the framerate of our screen and the HMD (*Oculus Rift*) were very different, we adjusted the pause to last $1/25$ seconds. Therefore, throughout this work the terminology is as follows: one frame is equivalent to $1/25$ seconds, and two frames are equivalent to $2/25$ seconds.

These three visual conditions were presented together with one of the four following auditory conditions: no sound, accompanied by a brief click sound (frequency of *2000 Hz*, duration of *3*

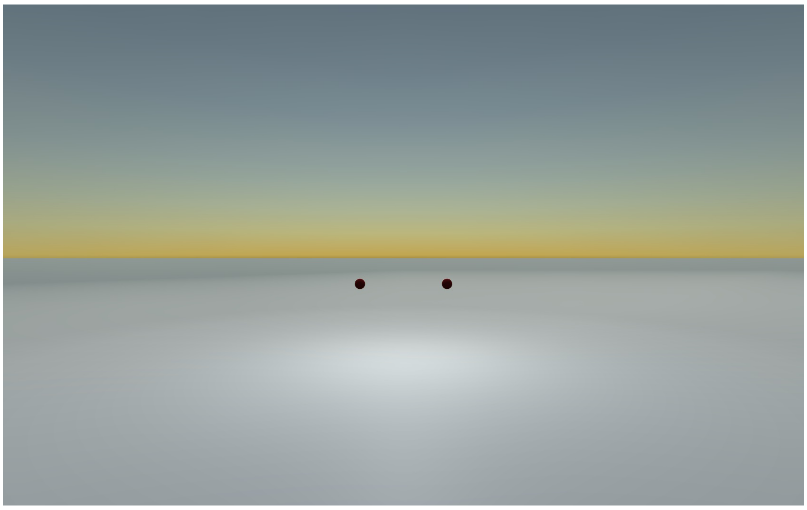


Figure 6.1: Initial layout of the scene for Experiment 1.

milliseconds) triggered *150 milliseconds* before or after the coincidence, or accompanied by a brief click sound at the point of coincidence.

Participants. Thirteen participants took part in the experiment (three female, ten male), with ages ranging from 18 to 28 years. All the participants volunteered to perform our experiments, and they were not aware of the purpose of each experiment. They were requested to fill a questionnaire about visual health, and we conducted a stereoscopic vision test to discard those participants with defective depth perception. They all had normal or corrected-to-normal vision.

Procedure. During the experiment we presented a total of twelve different conditions to each participant, three visual (continuous movement, pause one or two frames at the coincidence) and four auditory (no sound, sound at, before, or after the coincidence). Each of these conditions was presented *ten* times, making a total of 120 trials that appeared in a random order. We performed two blocks of the same experiment ordered randomly: one displayed on a regular screen (*Acer AL2216W TFT 22"*), and the other one displayed on an HMD (*Oculus Rift DK2*).

Before the HMD block, the lenses of the *Oculus Rift DK2* were adjusted to the participant eyes. We additionally introduced a training session before this block, where we showed two spheres at different depths and the participant had to choose which one was closer. We presented ten trials of the training with spheres at random depths. With this training the user gets used to the device, setup, and answering procedure.

We guided the participants through the test by showing several slides with descriptions of each phase of the experiment. After each trial, a slide was displayed with the question *"Did the spheres bounce or stream?"*, and a visual aid indicating the participant to answer with a mouse click (right or left).

Analysis and results. We use repeated measures ANOVA to test the influence of each of the conditions independently in the observed responses. For every participant, we take into account the answer (*bounce* or *stream*) in each of the ten trials. We need the repeated measures scheme because we measure the same independent variables (e.g., frames paused) under different conditions performed by the same subjects. We fix a significance value (p-value) of 0.05 in all the tests, and in those cases in which results from Mauchly's test of sphericity indicate that variances and covariances are not uniform, we report the results with the corresponding correction applied to the degrees of freedom (Greenhouse Geisser correction [66]). Prior to the analysis, we perform outlier rejection as detailed in the Appendix. We have three factors or variables of influence: (i) the overall influence of the display (2D scene presented on a *screen*, or 3D environment presented on an *HMD*); (ii) the influence of the *sound* when the spheres collide; and (iii) the influence of the length of the *pause* at the point of coincidence between the spheres. Results are presented in Table 6.1.

Table 6.1: Results (*F-test* and *significance*) of the analysis of the data with repeated measures ANOVA for Experiment 1. We test the influence of three factors in the perceived percentages of bounce responses.

	F	Sig.
Sound vs percent. bounce	83.664	0.000
Pause vs percent. bounce	63.528	0.000
Display vs percent. bounce	13.176	0.000

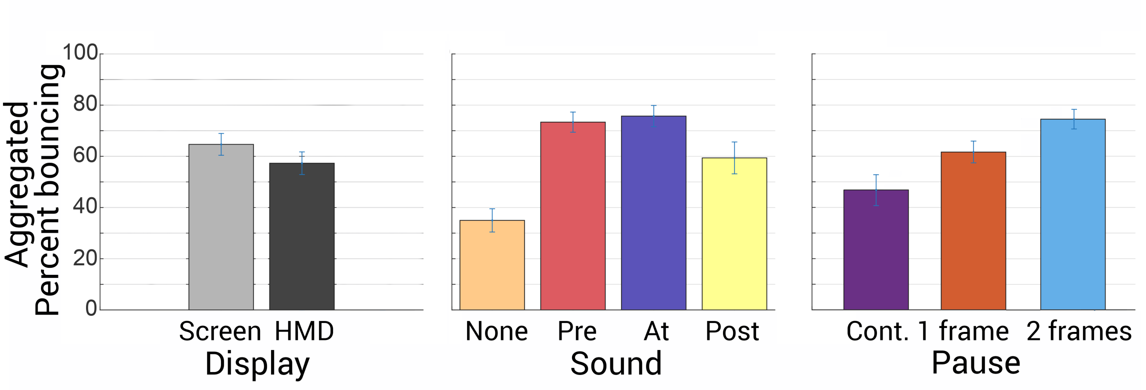


Figure 6.2: Aggregated percentages of bounce responses and corresponding error bars (standard error of the mean) for the Experiment 1. From left to right: Percentages for two display conditions (screen or HMD), percentages for four auditory conditions (no sound, sound at, before, or after the moment of coincidence of the spheres), and percentages for three visual conditions (continuous movement, pause one, or two frames at the point of coincidence of the spheres).

We can conclude that all three factors have a significant effect in the percentage of bounce responses, since all the p-values are below 0.05. We show in Figure 6.2 the mean percentages of bounce responses for the tested factors (error bars represent the standard error of the mean). We observe that the percentage of bounce responses decreases when using the HMD display. However, the main findings of Sekuler et al.'s work hold: a sound at the moment of coincidence, and a pause of two frames at the point of coincidence promote the perception of bouncing. We believe that the decrease in perceived bouncing in the tests with the HMD comes from the increase in the amount of visual cues due to the stereoscopic view. Sound promotes perception of bouncing when compared with the absence of sound; however, it has significantly less effect when reproduced after the point of coincidence. Still, there is a high tolerance for asynchrony between the sound and the visual input: even when the sound is delayed, the percentage of bounce responses increases. Also, as reported previously by Sekuler and others [327, 30, 326], the overall percentage of bounce responses increases with the duration of the pause.

6.2.2 *Experiment 2*

Goal. The goal of this experiment was to test whether a more complex scene could influence the crossmodal effect of sound altering visual motion perception. In order to do this, we increase the realism of the scene in three different ways (we term them three *blocks*) while keeping the proportions between distances and speed of the spheres of the original experiment.

Stimuli. The visual stimuli were rendered once again with *Unity*. We designed a new scene where the spheres are placed on a white table, inside a furnished room, and with a more realistic illumination. With respect to the first experiment we also increased the size of the spheres to *1 degree* of radius, and the distance between them to *8.4 degrees*, to make them more visible. A screenshot of the initial layout of the scene for the first block of the experiment is shown in Figure 6.3, left. For the second block of the experiment, starting from the scene in the first block,

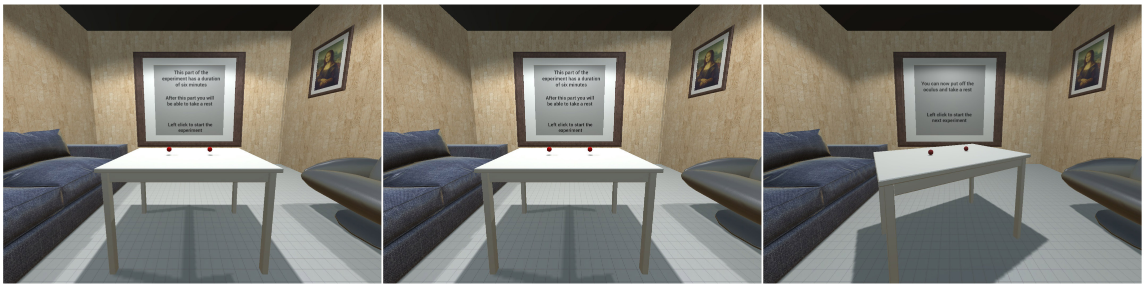

This part of the experiment has a duration of six minutes
After this part you will be able to take a rest
Left click to start the experiment
This part of the experiment has a duration of six minutes
After this part you will be able to take a rest
Left click to start the experiment
You can now put off the oculus and take a rest
Left click to start the next experiment

Figure 6.3: Initial layout of the scene for the three different blocks in Experiment 2. Left: increased radius of the spheres (block 1), middle: increased radius of the spheres and additional visual cues (block 2), and right: increased radius of the spheres and rotated plane of the collision (block 3)

Table 6.2: Results (*F-test* and *significance*) of the analysis of the data with repeated measures ANOVA for Experiment 2. We test the influence of three factors in the perceived percentages of bounces.

	F	**Sig.**
Sound vs percent. bounce	124.137	0.000
Pause vs percent. bounce	845.386	0.000
Scene vs percent. bounce	0.022	0.979

we additionally introduced two more visual cues to the spheres. First, we increased the glossiness of the material of the spheres, and second, we slightly lifted the spheres over the table in order to have more visible shadows (see Figure 6.3 middle). Finally, for the third block of the experiment, starting from the scene in the first block, we also rotated the plane of the collision between the spheres. We show a screenshot of the initial layout for this block in Figure 6.3 right.

Participants. Twenty seven participants took part in the experiment (two female, twenty-five male) with ages ranging from 19 to 32 years. As in the previous experiment, participants volunteered and took a questionnaire about visual health, and a stereoscopic depth test to assure that they all had correct depth vision. They all had normal or corrected-to-normal vision.

Procedure. During the experiment we presented a total of six different conditions, two visual (continuous movement, pause two frames at the coincidence), and three auditory (no sound, *click sound* at, or after the coincidence). Based on the results of the first experiment we removed the visual condition with a pause of one frame because the percentage of bouncing perceived was similar to the one perceived with the pause of two frames, and the auditory condition corresponding to the sound before the coincidence, also because of its similarity with the sound after the coincidence. Each of these conditions was presented *ten* times, making a total of 60 trials that appeared in a random order. All the blocks of the experiment were presented in the *HMD*, and each participant performed three randomly ordered blocks that corresponded to the three scenes described in the *Stimuli* section, totaling 180 trials per subject. Before starting the test, the participants performed the same training described in Experiment 1. Finally, in this experiment the slides with instructions about the test were shown on a frame on the back of the room striving to preserve as much as possible the realism of the environment.

Analysis and results. Again, we wanted to test three factors: the influence of each of the three scenes (three blocks), the influence of the *sound* when the spheres collide, and the influence of the *pause* at the point of coincidence between the spheres. Similarly to Experiment 1, we perform a repeated measures ANOVA; results are presented in Table 6.2. In Figure 6.4 we show the mean percentages of bounce responses for the tested factors, and the associated error bars representing the standard error of the mean. The analysis with the ANOVA reveals that, as before, there is a significant effect of the *sound*, and the *pause* in the perceived percentage of bounces. However, the p-value for the test with different scenes is very high, therefore we cannot draw any significant conclusion about the relationship between the three different scenes and the observed percentage of bouncing. When comparing Experiments 1 and 2 we can see that even when increasing the

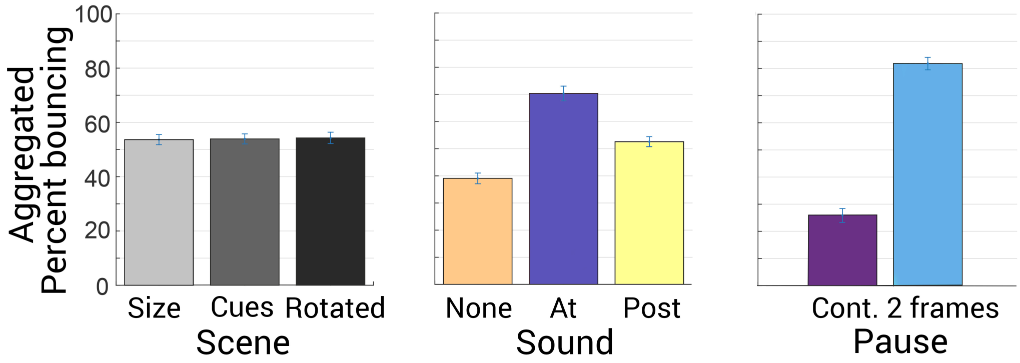


Figure 6.4: Aggregated percentages and error bars (standard error of the mean) for the Experiment 2. From left to right: Percentages for the three different scenes or blocks (increase in the size of the spheres, additional visual cues in the spheres, or rotated plane of the movement); percentages for three auditory conditions (no sound, sound at, or after the moment of coincidence of the spheres); and percentages for two visual conditions (continuous movement, or pause two frames at the point of coincidence of the spheres).

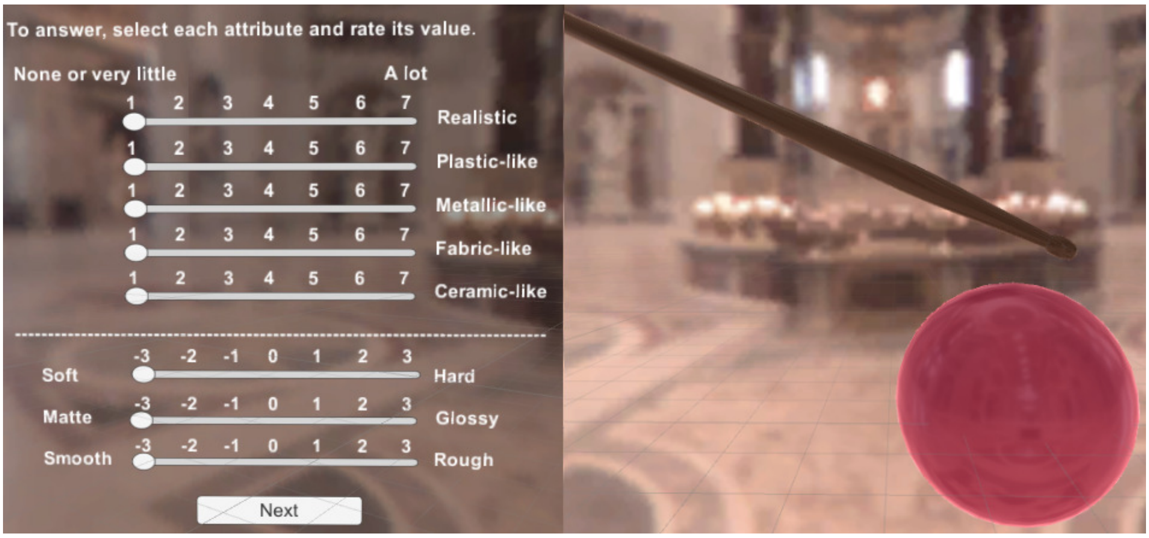
level of realism of the scene, the crossmodal effect of the sound altering the perceived motion still holds, although there is a general shift downwards of the percentage of bounce responses which can be observed by comparing the corresponding percentages of Figures 6.2 and 6.4. This shift downwards is possibly due to the presence of additional cues; however the high p-value of the scene factor, further indicates that there is no significant difference on the effect on crossmodal interaction between the three scenes (blocks) tested (i.e., no cue has proven to be significantly stronger or weaker in the detection of bouncing).

6.3 crossmodal material perception

Once we've proven that crossmodal interactions hold in VR we aim to analyze whether these interactions influence material perception. We have performed an experiment in order to determine how much the perception of material appearance is affected in virtual environments when a crossmodal interaction (visual and auditory stimuli) is presented in comparison with unimodal stimuli (only visual stimuli).

Goal. Our goal is twofold: we want to increase once more the stimuli complexity (not just a single sound with equal spheres, but different sounds paired with different visual stimuli), as well as determine if the presence of sound could help improve the immersion experience in VR environments, or even reduce its rendering costs.

Stimuli. We use *Unity* to render a set of spheres of different materials, including a phenolic sphere, metallic sphere, plastic sphere and fabric-like sphere. All the spheres are rendered with low and high visual resolution. The visual stimuli were rendered with the default material model (*GGX*). In the visual-only stimuli, they consisted on a sphere placed in front of the camera. In the audiovisual stimuli, the same sphere was presented, but this time with a wooden drumstick hitting it periodically from behind. Figure 6.5 shows an example of an audiovisual stimulus. The auditory stimuli were recorded mono sounds from the MIT hit sounds dataset [267], that were synchronized to play when the drumstick hits the sphere (in the MIT hit sounds database, it is also a wooden drumstick that is used to produce the sounds). We virtually placed sound sources in the 3D scene, effectively spatializing the mono sound regarding the participant and the sphere's relative position. Note that this is different from using stereo sound tracks, since participants actually perceive a 3D audio effect (i.e., they perceive effects such as head-shadowing). The same sound was always presented for the same material, regardless of its rendering quality. We used four different materials for the sphere. The materials were modeled in Unity and chosen to cover a range of material categories, which are chosen based on the types of materials present in the MERL database. In particular, we have: *metal*, *fabric*, *plastic*, and a *phenolic* material, (a specular


To answer, select each attribute and rate its value.

None or very little
A lot
1 2 3 4 5 6 7 Realistic
1 2 3 4 5 6 7 Plastic-like
1 2 3 4 5 6 7 Metallic-like
1 2 3 4 5 6 7 Fabric-like
1 2 3 4 5 6 7 Ceramic-like
Soft -3 -2 -1 0 1 2 3 Hard
Matte -3 -2 -1 0 1 2 3 Glossy
Smooth -3 -2 -1 0 1 2 3 Rough
Next

Figure 6.5: Left: The panel with the attributes that the participants had to rate. With the controller's joystick they could set the rating value and move between the attributes and the "next" button. Right: Presentation of a stimulus in the scene, showing both a sample sphere and the wooden drumstick.

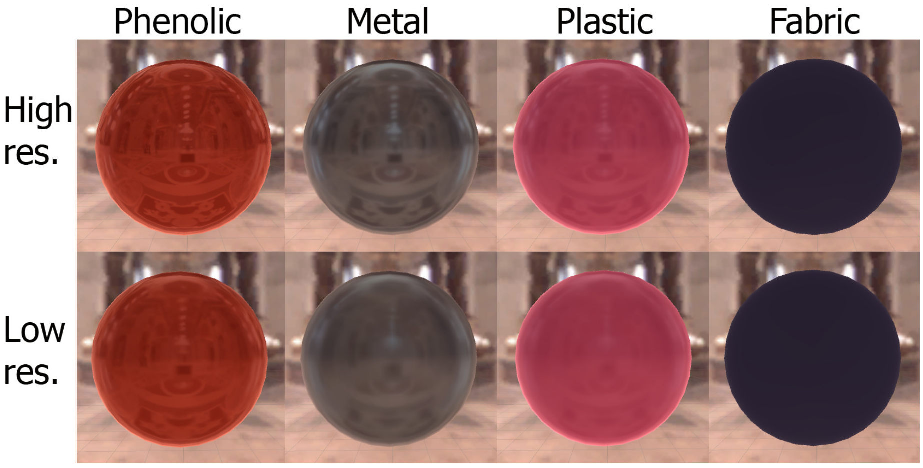


Figure 6.6: Each column shows one of the four possible materials used in the experiment. From left to right: Phenolic, metal, plastic, and fabric. Each row shows the material on high resolution (top) and low resolution (bottom).

material typically used as coating and to which we associated a ceramic-like sound). Each of the materials was presented twice: one with Unity's light-probe default rendering illumination quality (high resolution, 128 samples) and another with a reduced quality (low resolution, 32 samples). Figure 6.6 shows these eight combinations. The illumination in all cases was the environment map *St. Peters*, from the Light Probe Image Gallery [72], since real-world illumination, and that environment map in particular, facilitates material discrimination tasks [98].

Participants. The participants wore isolating headphones (Vic Firth SIH1) during the experiment and they provided answers to the rating questions with an Xbox controller. Thirteen new participants took part in the experiment (two female, eleven male), with ages ranging from 19 to 29 years. They all had normal or corrected-to-normal vision. Similarly to the two previous experiments, all participants took part in a questionnaire and a stereoscopic depth test.

Procedure. We use an HMD to determine if the presence of a collision sound can alter the perceived appearance of a material in a virtual environment. We presented different materials and asked the participants to rate a set of perceptual attributes. This attributes included low-level perceptual traits (*soft/hard*, *glossy/matte*, and *rough/smooth*), and high-level descriptors of appearance

Table 6.3: Conditions in our experiment.

	Low res.	High res.
Visual only	C_0	C_1
Audiovisual	C_2	C_3

(*realistic*, *metallic-like*, *plastic-like*, *fabric-like*, and *ceramic-like*). We chose these attributes because they are discriminatory [332], and they have also been used previously for assessing the interactions of sound and visual stimuli [221]. During the experiment we presented a total of 24 different stimuli to the participants (4 (materials) × 2 (quality levels) × 2 (modalities) + 3 (control materials) × 2 (modalities) + 2 (training stimuli)). Each of the stimuli was shown once. First, a brief explanation of the procedure and the attributes to be used was made. Then, the participants underwent a training with two different stimuli to make sure they understood what they were being asked to do and to learn how the controller worked. This training helped the user to get used to the device, setup, and answering procedure. The experiment was divided in *two* different blocks, with a total of four conditions (see Table 6.3): visual-only stimuli ($\{C_0, C_1\}$ for the low and high quality rendering, respectively) and audiovisual stimuli ($\{C_2, C_3\}$ corresponding to the low and high quality rendering, respectively).

The order of these two blocks was randomized: half the participants started with visual-only stimuli and the other half with audiovisual stimuli. Each of the blocks had 11 different stimuli (the four materials were presented in low and high quality, and there were 2 control materials). The presentation order of the stimuli within a block was also randomized, although ensuring that two qualities of the same material did not appear successively. To the left of the stimuli, a panel with the questions of the experiment was presented (Figure 6.5, left). Each stimulus, together with the questions, was displayed for 60 seconds. At the end of the 60 seconds, only the questions panel remained. A counter showing the remaining time before the stimulus disappeared was also displayed to make the user aware of the remaining time. Each question pertained to an attribute and a 7-point scale was used to provide the rating answer.

If the participant had rated all the attributes before the 60 seconds had passed, she could move forward to the next stimulus. Between each pair of stimuli, a gray screen with a red cube appeared so that the participants could take a rest if needed before continuing the experiment. The next stimulus appeared when the participants aligned a visual target with the red cube; in this way we also ensured that they were all looking at the same point of the scene when each stimulus is first presented.

Analysis and results. For the analysis we first performed outlier rejection by using our control materials: subjects were discarded when they did not provide a reasonable answer to the attribute *glossiness* in our control materials (see Figure 6.7). We discarded *two* subjects with this procedure, leaving a total of *eleven* users to analyze. We tested our data for normality using the *Shaphiro-Wilk* test, which is well suited for small samples. The ratings for all our attributes did not present a normal distribution ($p < 0.05$), we therefore turned to non-parametric methods to carry out the analysis of our four conditions. For each material and for each attribute we perform pairwise comparisons between the four conditions ($\{C_0, C_1, C_2, C_3\}$) by using the *Wilcoxon Signed-Rank* test. This test is a nonparametric equivalent to the dependent *t-test*, and can be used to investigate changes in ratings when subjects are presented with several conditions. Following Kerr and Pellacini [163] we consider significant p-values below 0.1, which indicates a 90% confidence that the means of the two different conditions differ. Our main insights are summarized in Table 6.4, and described in detail in the following.

Influence of resolution. The resolution of the light-probe plays an important role in the perceived *glossiness* of the material, as can be seen in Figure 6.8. This resolution affects the specular reflections (see Figure 6.6), therefore it is particularly noticeable in very specular materials, i.e., there is a significant difference between the high and low resolution stimuli for the *metallic* material while for the *fabric* material this difference is barely noticeable. We found a significant interaction in the *metallic* material between the *resolution* and the perceived *glossiness* ($p = 0.041$ for $\{C_0,\ C_1\}$).

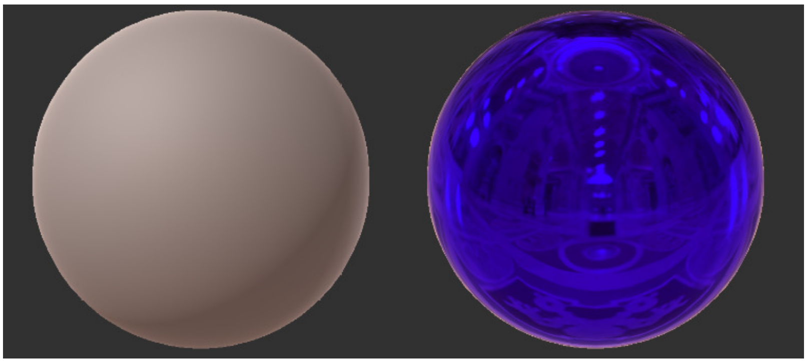


Figure 6.7: Control materials used to discard outliers. We discarded a subject if her rating for the attribute *glossiness* was above 2 for a very diffuse material (*left*), or below 6 for a very specular material (*right*), on a 7-point scale.

Table 6.4: Summary of the results (significance) of the analysis of the data with *Wilcoxon Signed-Rank* tests for Experiment 3. We compare the mean value of the *attribute* assigned to the *material* for the specified *conditions*.

	Mat.	**Att.**	**Cond.**	**Sig.**
Influence of resolution	Metallic	Glossy	$C_0 < C_1$	0.041
Influence of sound	Metallic	Plastic	$C_0 > C_2$	0.041
	Metallic	Metallic	$C_0 < C_2$	0.048
	Phenolic	Plastic	$C_0 > C_2$	0.027
			$C_1 > C_3$	0.017
	Phenolic	Ceramic	$C_0 < C_2$	0.027
			$C_1 < C_3$	0.017

The same trend can be observed for the conditions $\{C_2, C_3\}$. For the other three materials, interestingly, we observe no significant difference in the perception of glossiness regardless of resolution. These findings could be useful to save rendering costs by adjusting the resolution of light-probes according to the material, since the resolution of the light-probe has little effect in the perception of diffuse materials.

Influence of sound. We have found several interactions describing a significant effect of the presence of sound in the ratings for the high-level attributes. For the *metallic* material the ratings for the *plastic* attribute are significantly lower when the stimuli is presented together with sound ($p = 0.041$ for $\{C_0, C_2\}$). Conversely, the ratings for the *metallic* attribute are significantly higher ($p = 0.048$ for $\{C_0, C_2\}$). This effect is significant when we compare the low resolution conditions $\{C_0, C_2\}$, but not when we compare the high resolution conditions $\{C_1, C_3\}$. We believe this can be due to the high resolution visual stimuli better conveying the visual traits of the material; this undermines the effect of the auditory stimuli, since the user recognizes the material well enough just with the visual stimuli. This suggests that the effect of sound in material identification tasks may be more relevant when the visual stimuli has a low quality. For the *phenolic* material the mean of the *plastic* attribute significantly decreases when the user is presented with the multimodal stimuli. In this case, the effect is noticeable both for the low resolution ($p = 0.027$ for $\{C_0, C_2\}$) and high resolution ($p = 0.017$ for $\{C_1, C_3\}$) conditions. For this same material, the mean of the *ceramic* attribute increases ($p = 0.078$ for $\{C_0, C_2\}$ and $p = 0.077$ for $\{C_1, C_3\}$), which indicates that the sound effectively helps the users identifying the material. We did not find significant interactions for the *fabric* and the *plastic* materials, however, a similar trend can be seen in Figure 6.9: for every material there is an increase in the mean rating of its corresponding attribute (bars outlined in orange in Figure 6.9) when the user is presented with the audiovisual stimuli. These findings agree

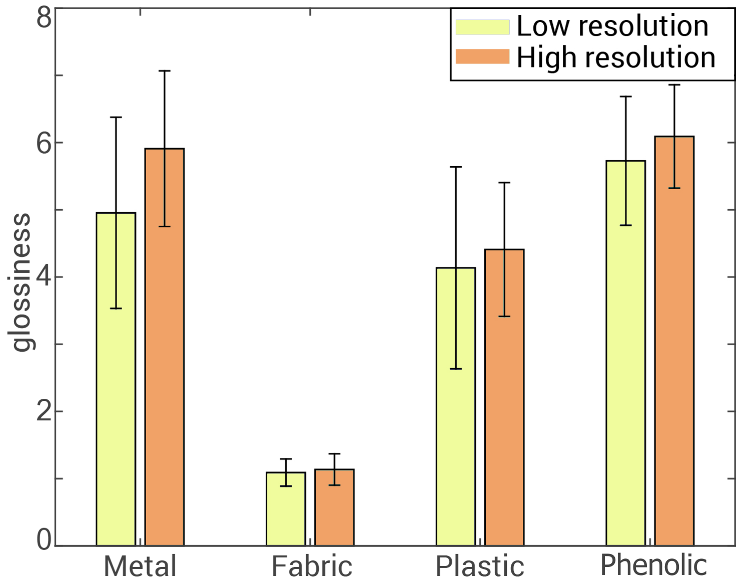


Figure 6.8: Mean ratings for the *glossy* attribute when the user is presented with the low resolution (yellow) and the high resolution visual stimuli (orange) for our four materials analyzed. Error bars show ±1 SEM. There is a trend indicating that the perceived glossiness increases in the high resolution stimuli.

with those of Giordano and McAdams [113], which supported that impact sounds were good descriptors for material identification tasks, and they suggest that the sound also benefits material discrimination tasks in VR, particularly when such materials are not easily recognizable only by its visual traits. Our findings indicate that a high resolution is required for material identification when its representation consists on visual stimuli only, however if additional auditory stimuli are introduced, the resolution could be lowered while keeping the perceived appearance, thus saving rendering costs.

6.4 conclusions

In this work, we have performed an exploration of crossmodal perception in virtual reality scenarios. We have studied the influence of auditory signals in the perception of visual motion. To do so, we first replicated an existing experiment which demonstrated the existence of a crossmodal interaction between both senses with simple stimuli on a 2D conventional display. We were able to successfully replicate it, obtaining the same trends in the results, and then extended it to virtual reality with a HMD. We found that the same trends hold on a HMD (i.e., the factors explored had the same influence on the crossmodal effect), but that there is a reduction in the crossmodal effect. This reduction essentially means that there is a shift in the results towards a better accuracy of subjects in performing the tasks assigned in the HMD setup. This can be due to the presence of additional cues, in particular depth cues including binocular disparity and possibly motion parallax. A similar conclusion can be drawn in our second experiment: We repeated the first experiment (only on the HMD), with new subjects, and with more complex stimuli (we had three different variations of the initial stimulus) to see whether the effect would still hold with more realistic scenery. We observed a further reduction of the crossmodal effect (subjects were better at detecting the correct behavior of the stimuli), which we hypothesize is due to the presence of additional cues, in this case pictorial cues (shading, perspective, texture).

We then move on to the particular case of material appearance perception, with the aim of laying the foundation for future practical applications. When analyzing crossmodal effects in a VR setup, we have observed that findings previously reported for conventional displays hold: the presence of sound improves material recognition. We have also included two different rendering qualities

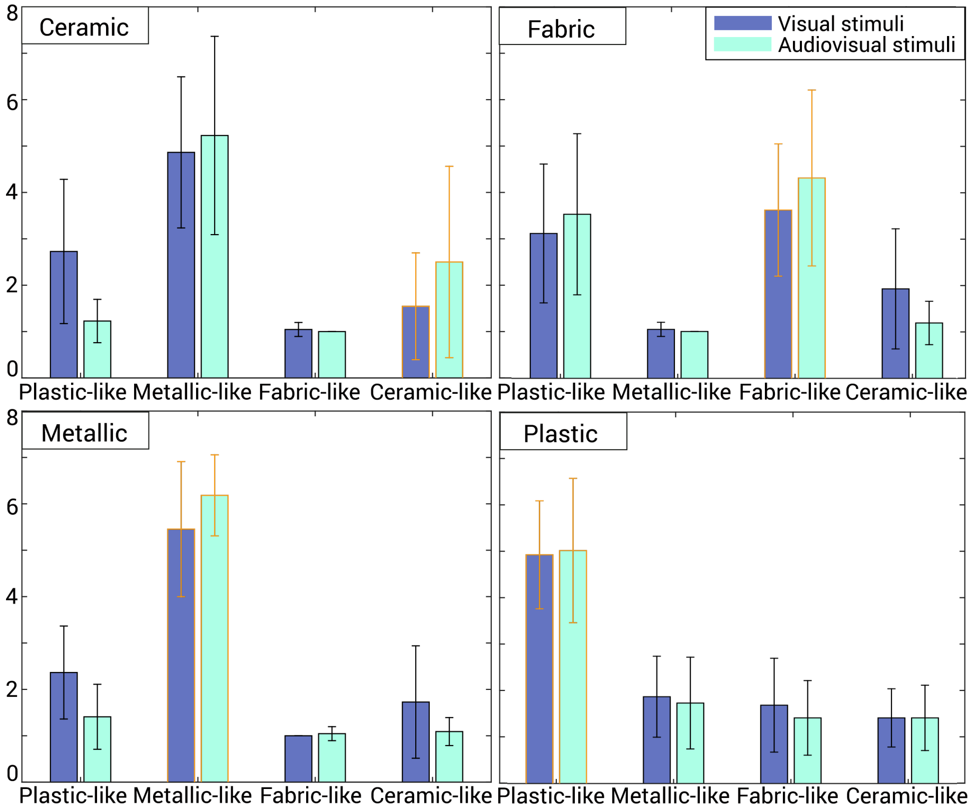


Figure 6.9: Mean ratings for the high-level attributes when the user is presented with the visual only stimuli (blue) and the audiovisual stimuli (green) for our four materials. Error bars show ±1 SEM. For every material, there is an increase in the mean rating of its corresponding attribute (marked by an orange outline) when the visual stimuli is accompanied by sound.

for the material, and observed two main findings: First, that the influence of the rendering quality on the perception of low-level attributes such as glossiness varies between material categories. Second, that the effect of sound in the recognition of materials is more relevant for the low quality-rendering case than for the high quality one.

In summary, regarding the research questions posed in Section 6.1, we can conclude that:

- The crossmodal effect holds in VR environments, even when increasing the complexity of scenes.
- Crossmodal interactions influence the perception of material traits in VR environments. More research is necessary to be able to quantify this effect and further understand it.

As in all studies of similar nature, some of our findings may not generalize to conditions outside our study. We have focused on simple sounds and scenes with a controlled increase of complexity. This allows us to isolate the effects of each condition, and perform a systematic analysis. We believe these are just a few steps in the exploration of crossmodal perception in virtual reality. In the future, we would like to expand these experiments by including other potentially influencing factors or effects, and by further increasing the complexity of the stimuli. An interesting avenue for future research would be to use different sound types and qualities in addition to the rendering qualities. In the area of material perception, we hope this work serves as the foundation for future explorations. Here we have employed representative materials of four main categories, future works should further delve into the problem, analyzing a larger variety of materials, especially among specular ones where there is more to be gained from exploitation of this crossmodal interaction. This could result in the development of quantitative prediction models to enable further practical applications of crossmodal perception in VR environments.

Part IV

CONCLUSION

7 Conclusion and Future Work

In this thesis we have used VR as a tool to better understand cognitive processes, establishing behavioral guidelines (improving the knowledge of how humans behave in immersive environments and helping practitioners to create better experiences) or modeling user behavior. This knowledge is key to improve user experience. Through this thesis we have focused on visual perception and its interaction with other modalities considering embodiment as a key component to understand human behavior. We have learned how to use behavioral data, both subjective and objective in order to build useful information on a variety of topics, from appearance similarity metrics to audiovisual suppressive effects in immersive environments. Thanks to the use of VR we can better control the sensory information that users receive and safely research in a complex, natural and reproducible environment. In the future we may see VR systems integrated with wearable biosensors, which would allow us to collect additional quantitative data in complex virtual environments.

Visual perception of realistic stimuli. In this part we have presented two different lines of work. The first one is devoted to appearance similarity metrics. We have learned how to integrate human perception information with a deep learning model. We have been able to gather enough data for the model to learn a similar enough behavior thanks to our rendered dataset and the large-scale, crowdsourced user studies. The applications we propose suggest the potential of this combination in a variety of fields. This work has exciting potential avenues for future work. For example, we have only gathered human judgements based on static images, with a single geometry and illumination. These factors have been further explored by other subsequent works [331, 178] yielding more robust models with a better generalization capability. For future work, the model could also be trained with real data instead of synthetic data only to allow for richer, more realistic learning and the latent space of the network could be studied for a better understanding of how perceptual appearance similarity is derived from visual input.

In the second line of work of this part, we have focused on the interplay between visual and time perception in immersive environments. We have studied how the manipulation of low-level visual factors affect perceived time, finding that larger visual changes compress perceived time in intervals of up to three minutes. A possible explanation for this effect lies in the attentional gate model, which suggests that attention has to be divided between visual and temporal perception. When there are more (or larger) visual changes, our attention is focused on the visual domain. The limited remaining attention that can be devoted to the temporal domain in such situations results in a perceived shortening of the experienced time interval. It is possible that how attention is divided between modalities also has an effect on the findings described in the second part of this thesis, being an important factor in contexts of high cognitive load.

As for future avenues of work, we would like to further explore this relationship by means of additional user studies to validate and consolidate our findings. Besides, there are works that show how a perceived shortening of time can be useful in several applications like medical treatment [318]. We believe that our findings could be used as guidelines to modify the visual aspect of VR applications in order to trigger different temporal alterations: for example, a more vivid contrast palette may help increase the shortening of perceived time while using VR; faster cuts in 360 movies could increase the perceived pacing of the story, etc. Before directly using our findings in these applications, a thorough study of the effects of confounding factors (including other high-level cognitive processes like cognitive load, emotional valence, arousal, boredom and tiredness, etc.) in the observed behavior should be carried out. Finally, since time perception is affected by subjective experience (with age affecting how we perceive time) we would like to repeat our studies with a more diverse set of participants.

Multimodal perception in immersive environments. In this part we have focused on how different modalities can affect user experience in virtual reality. We first provide a thorough, up-to-date overview of the uses of multimodality in immersive environments. We focus on the main improvements that multimodality has shown so far: an increase of realism, contributions on guiding attention, increasing user performance, improving navigation in virtual environments, etc. In applications ranging from medicine to education, training and entertainment the inclusion of different modalities is key for a complete and improved user experience. Our survey reveals several gaps in multimodality research which make up for promising future research. The research on the interplay between three or more modalities is scarce. This kind of research is a challenge for several reasons: the space to explore grows exponentially with each modality added, the interplay between different modalities is complex and difficult to measure directly, the environment has to be carefully designed to avoid the sensory overload of the users which would hinder the experience and the existing hardware for multimodal feedback should be improved. However, we believe that a sound and scalable user study methodology similar to that shown through this work, together with a parametric derivation of models and metrics correlated with human cognitive processes can push forward the understanding of multimodal perception in immersive environments.

Then we show how visual and auditory cues presented in a temporally congruent, spatially incongruent manner can significantly degrade visual performance when compared to a baseline visual-only condition. We devise a user study to explore and record this effect in which we move on from tight laboratory conditions to a realistic, complex virtual environment in which users can freely move. This elicits higher cognitive loads, as well as a more natural behavior. In the future, we would like to implement a redirected walking algorithm which applies crossmodal suppressive effects orthogonally to unimodal suppressive effects [369] which could potentially lead to better virtual to real compression maps. Further studies are needed to identify the underlying cause of the suppressive effect.

Finally, we delve into the importance of temporal synchronization between sensory modalities. In particular, we study how auditory and visual cues need to be correctly synchronized in order to be perceived as a single, multimodal event. In the context of material appearance perception, participants can better identify materials in a virtual environment if accompanied by sound. Moreover, the realism of the stimuli is judged as better in the presence of sound even if the visual quality is degraded. Backtracking to appearance similarity distance metrics, it would be really interesting to study how perceived similarity of different materials is affected by an immersive environment.

Personal conclusions. I started my thesis after graduating from a computer engineering degree and a biomedical engineering Master, following my own personal interests: I was fascinated by how the human brain works. During this thesis we have tried to bring the knowledge of cognitive sciences to computer science, with the firm belief that this union can only enhance the scope of what we can achieve with new technologies. Throughout this thesis we have worked on a variety of seemingly different topics. However, always inspired by the same motivation and with the invaluable guidance of my supervisors and colleagues we have been able to establish a common methodological approach to extract behavioral data (regardless of the specific topic to work on) and turn it into useful information and models. Not only have I improved and polished my technical skills, I have also grown over the years a series of soft skills that now I believe are essential for any researcher. I have had the amazing luck of working in interdisciplinary, international teams learning how to integrate diverse ideas and knowledge into coherent works. I will always cherish the time I spent in my research internships, where I could meet researchers of every possible background, each of them working on different projects. It helped me get out of my comfort zone and grow as a person. Supervising other students has also made me improve: having a better understanding of the high-level picture, being organized, identifying potential bottlenecks and planning in advance. I started to better integrate these skills into my workflow when I saw how in need my students were of them. Through the last few years I have also learned to fail. Not all of our projects have been successful and that is ok. Not all of our publications have been accepted in the first submission but they did improve with each rejection. I also know now that sometimes we have to prioritize: the moment or the circumstances may not be the best, or the idea may simply not be good enough. All in all, I have spent the last years of my life learning about topics I am excited about, meeting amazing people, growing as a person and doing all the things I love in my work. I did not know what to expect when I started this thesis, but could not be happier to have chosen this path. I only hope that what the future holds for me will be just as good.

8 Conclusiones y trabajo futuro

En esta tesis hemos utilizado la VR como herramienta para comprender mejor los procesos cognitivos, creando guías con información de alto nivel (mejorando el conocimiento de cómo se comportan los seres humanos en entornos inmersivos y ayudando a los profesionales a crear mejores experiencias) o modelando el comportamiento cuando es posible. Este conocimiento es clave para mejorar la experiencia del usuario. En esta tesis nos hemos centrado en la percepción visual y su interacción con otras modalidades considerando la propriocepción como un componente clave para entender el comportamiento humano. Hemos aprendido a utilizar datos de comportamiento tanto subjetivos como objetivos para construir información útil, desde métricas de similitud de apariencia hasta efectos de supresión audiovisual en entornos inmersivos. Gracias al uso de la VR podemos controlar mejor la información sensorial que reciben los usuarios e investigar con seguridad en un entorno complejo, natural y reproducible. En el futuro podríamos ver sistemas de VR integrados con biosensores, lo que nos permitiría recoger datos cuantitativos adicionales en entornos virtuales complejos.

Percepción visual de estímulos realistas. En esta parte hemos presentado dos líneas de trabajo diferentes. La primera está dedicada a las métricas de similitud de apariencia. Hemos aprendido a integrar la información de la percepción humana con un modelo de *deep learning*. Hemos sido capaces de recopilar suficientes datos para que el modelo aprenda un comportamiento similar al humano gracias a nuestro conjunto de datos renderizados y a los estudios de usuarios a gran escala. Las aplicaciones que proponemos sugieren el potencial de esta combinación en diversos campos. Este trabajo tiene interesantes posibilidades para el futuro. Por ejemplo, sólo hemos recogido juicios humanos basados en imágenes estáticas, con una única geometría e iluminación. Estos factores han sido explorados más a fondo por otros trabajos posteriores [331, 178] dando lugar a modelos más robustos con una mejor capacidad de generalización. En el futuro el modelo también podría ser entrenado con datos reales en lugar de sólo con datos sintéticos para permitir un aprendizaje más rico y realista, y el espacio latente de la red podría ser estudiado para una mejor comprensión de cómo la similitud de apariencia perceptual se deriva de la entrada visual.

En la segunda línea de trabajo de esta parte, nos hemos centrado en la interacción entre la percepción visual y del tiempo en entornos inmersivos. Hemos estudiado cómo la manipulación de factores visuales de bajo nivel afecta a la percepción del tiempo, descubriendo que los cambios visuales más grandes comprimen el tiempo percibido en intervalos de hasta tres minutos. Una posible explicación de este efecto reside en el modelo de la puerta atencional, que sugiere que la atención tiene que dividirse entre la percepción visual y la temporal. Cuando hay más cambios visuales (o cambios mayores), nuestra atención se centra en el ámbito visual. La limitada atención restante que puede dedicarse al dominio temporal en tales situaciones da lugar a un acortamiento percibido del intervalo de tiempo experimentado. En cuanto a futuras vías de trabajo, nos gustaría seguir explorando esta relación mediante estudios adicionales con usuarios para validar y consolidar nuestros hallazgos. Además, hay trabajos que muestran cómo la compresión percibida del tiempo puede ser útil en varias aplicaciones como el tratamiento médico [318]. Creemos que nuestros hallazgos podrían servir como guía para modificar el aspecto visual de las aplicaciones de VR con el fin de desencadenar diferentes alteraciones temporales: por ejemplo, una paleta de contrastes más viva podría ayudar acortar más el tiempo percibido mientras se utiliza la VR; unos cortes más rápidos en las películas de 360 podrían aumentar el ritmo percibido de la historia, etc. Antes de utilizar directamente nuestros hallazgos en estas aplicaciones, debería realizarse un estudio exhaustivo de los efectos de los factores de confusión (incluidos otros procesos cognitivos de alto nivel como la carga cognitiva, la valencia emocional, el nivel de emoción, el aburrimiento y el cansancio, etc.) en el comportamiento observado. Por último, dado que la percepción del tiempo se ve afectada por la experiencia subjetiva (y la edad afecta a la forma en que percibimos el tiempo), nos gustaría repetir nuestros estudios con un conjunto más diverso de participantes.

Percepción multimodal en entornos inmersivos. En esta parte nos hemos centrado en cómo las diferentes modalidades pueden afectar a la experiencia del usuario en la realidad virtual. En primer lugar, ofrecemos una visión general y actualizada de los usos de la multimodalidad en los entornos inmersivos. Nos centramos en las principales mejoras que la multimodalidad ha mostrado hasta el momento: aumento del realismo, contribución a la orientación de la atención, aumento del rendimiento del usuario, mejora de la navegación en entornos virtuales, etc. En aplicaciones que van desde la medicina hasta la educación, pasando por la formación y el entretenimiento, la inclusión de diferentes modalidades es clave para una experiencia de usuario completa y mejorada. Nuestro estudio del estado del arte revela varias lagunas en la investigación sobre la multimodalidad que constituyen una prometedora investigación futura. La investigación sobre la interacción entre tres o más modalidades es escasa. Este tipo de investigación supone un reto por varias razones: el espacio a explorar crece exponencialmente con cada modalidad añadida, la interacción entre las diferentes modalidades es compleja y difícil de medir directamente, el entorno tiene que diseñarse cuidadosamente para evitar la sobrecarga sensorial de los usuarios (lo que empeoraría la experiencia), y el hardware existente para la información sensorial multimodal debería mejorarse. Sin embargo, creemos que una metodología de estudio de usuarios sólida y escalable, similar a la mostrada en esta tesis, junto con una derivación paramétrica de modelos y métricas correlacionadas con los procesos cognitivos humanos, puede impulsar la comprensión de la percepción multimodal en entornos inmersivos.

A continuación, mostramos cómo las señales visuales y auditivas presentadas de forma temporalmente congruente y espacialmente incongruente pueden degradar significativamente el rendimiento visual en comparación con una condición de base sólo visual. Diseñamos un estudio con usuarios para explorar y registrar este efecto, en el que pasamos de condiciones estrictas de laboratorio a un entorno virtual realista y complejo en el que los usuarios pueden moverse libremente. Esto provoca una mayor carga cognitiva, así como un comportamiento más natural. En el futuro, nos gustaría implementar un algoritmo de marcha redirigida (*redirected walking*) que aplique los efectos de supresión que hemos encontrado ortogonalmente a los efectos de supresión unimodal [369], lo que podría potencialmente conducir a mejores mapas de compresión virtual a real. Además, se necesitan más estudios para identificar la causa subyacente del efecto supresivo.

Por último, se profundiza en la importancia de la sincronización temporal entre las modalidades sensoriales. En concreto, se estudia cómo las señales auditivas y visuales deben estar correctamente sincronizadas para ser percibidas como un único evento multimodal. En el contexto de la percepción del aspecto de los materiales, los participantes pueden identificar mejor los materiales en un entorno virtual si van acompañados de un sonido correctamente sincronizado. Además, el realismo de los estímulos se juzga mejor en presencia del sonido aunque la calidad visual se vea degradada. Volviendo a las métricas de distancia de similitud de apariencia, sería realmente interesante estudiar cómo la similitud percibida de los diferentes materiales se ve afectada por un entorno inmersivo.

Conclusiones personales. Empecé mi tesis después de graduarme en un grado de ingeniería informática y un máster de ingeniería biomédica, siguiendo mis propios intereses personales: me fascinaba el funcionamiento del cerebro humano. A lo largo de esta tesis hemos tratado de acercar los conocimientos de las ciencias cognitivas a la informática, con la firme convicción de que esta unión no puede sino potenciar el alcance de lo que podemos conseguir con las nuevas tecnologías. A lo largo de esta tesis hemos trabajado en una variedad de temas aparentemente diferentes. Sin embargo, siempre inspirados por la misma motivación y con la inestimable guía de mis supervisores y compañeros, hemos sido capaces de establecer un enfoque metodológico común para extraer datos de comportamiento (independientemente del tema específico a trabajar) y convertirlos en información y modelos útiles. No sólo he mejorado y pulido mis habilidades técnicas, sino que también he cultivado a lo largo de los años una serie de habilidades transversales que ahora creo que son esenciales para cualquier investigador. He tenido la increíble suerte de trabajar en equipos interdisciplinarios e internacionales aprendiendo a integrar ideas y conocimientos diversos en trabajos comunes. Siempre apreciaré el tiempo que pasé en mis prácticas de investigación, donde pude conocer a investigadores de distintas procedencias trabajando en proyectos variados. Encontrarme en ese entorno me ayudó a salir de mi zona de confort y a crecer como persona. Supervisar a otros estudiantes también me ha hecho mejorar: comprender mejor la vista a alto nivel de un proyecto, ser organizado, identificar con antelación posibles cuellos de botella y planificar de forma acorde. Empecé a integrar mejor estas habilidades en mi trabajo cuando vi que mis alumnos las necesitaban. En los últimos años también he aprendido a fracasar. No todos nuestros proyectos han tenido éxito, lo que no tiene nada de malo. No todas nuestras publicaciones han sido aceptadas a la primera pero han mejorado con cada rechazo. Ahora también sé que a veces hay que priorizar: puede que el momento o las circunstancias no sean los mejores, o que la idea simplemente no sea lo suficientemente buena. En definitiva, he pasado los últimos años de mi vida aprendiendo sobre temas que me entusiasman, conociendo a gente increíble, creciendo como persona y haciendo cosas que me gustan en mi trabajo. No sabía qué esperar cuando empecé esta tesis, pero no podría estar más contenta de haber elegido este camino. Sólo espero que lo que me depare el futuro sea igual de bueno.

Bibliography

[1] Adelson, E. H. On seeing stuff: the perception of materials by humans and machines. In *Human Vision and Electronic Imaging* (2001), vol. 4299, pp. 1–13.

[2] Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., and Belongie, S. Generalized non-metric multidimensional scaling. In *Artificial Intelligence and Statistics* (2007), pp. 11–18.

[3] Akhtar, Z., and Falk, T. H. Audio-visual multimedia quality assessment: A comprehensive survey. *IEEE Access 5* (2017), 21090–21117.

[4] Alais, D., and Burr, D. The ventriloquist effect results from near-optimal bimodal integration. *CurrentBiology* (2004).

[5] Ali, N., Ullah, S., Rabbi, I., and Alam, A. The effect of multimodal virtual chemistry laboratory on students' learning improvement. In *International Conference on Augmented and Virtual Reality* (2014), Springer, pp. 65–76.

[6] Allan, L. G. The perception of time. *Perception & psychophysics 26*, 5 (1979), 340–354.

[7] Allue, M., Serrano, A., Bedia, M. G., and Masia, B. Crossmodal Perception in Immersive Environments. In *Spanish Computer Graphics Conference (CEIG)* (2016).

[8] Alves Fernandes, L. M., et al. Exploring educational immersive videogames: an empirical study with a 3d multimodal interaction prototype. *Behaviour & Information Technology 35*, 11 (2016), 907–918.

[9] Ammi, M., and Katz, B. F. Intermodal audio-haptic metaphor: improvement of target search in abstract environments. *International journal of human-computer interaction 30*, 11 (2014), 921–933.

[10] Anderson, B. L. Visual perception of materials and surfaces. *Current Biology 21*, 24 (2011), R978–R983.

[11] Andreasen, A., Geronazzo, M., Nilsson, N., Zovnercuka, J., Konovalov, K., and Serafin, S. Auditory feedback for navigation with echoes in virtual environments: training procedure and orientation strategies. *IEEE Trans. on Visualization and Computer Graphics 25*, 5 (2019), 1876–1886.

[12] Arabadzhiyska, E., Tursun, O. T., Myszkowski, K., Seidel, H.-P., and Didyk, P. Saccade landing position prediction for gaze-contingent rendering. *ACM Transactions on Graphics (TOG) 36*, 4 (2017), 50.

[13] Armbrüster, C., Wolter, M., Kuhlen, T., Spijkers, W., and Fimm, B. Depth perception in virtual reality: distance estimations in peri-and extrapersonal space. *Cyberpsychology & Behavior 11*, 1 (2008), 9–15.

[14] Arnold, P. You better eat to survive! exploring edible interactions in a virtual reality game. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (2017), pp. 206–209.

[15] Arons, B. A review of the cocktail party effect. *Journal of the American Voice I/O Society 12*, 7 (1992).

[16] Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems 16*, 6 (2010), 345–379.

[17] Baddeley, A. D., and Hitch, G. J. Developments in the concept of working memory. *Neuropsychology 8*, 4 (1994), 485.

[18] Bailey, H. D., Mullaney, A. B., Gibney, K. D., and Kwakye, L. D. Audiovisual integration varies with target and environment richness in immersive virtual reality. *Multisensory Research 31*, 7 (2018).

[19] Bailey, R., McNamara, A., Sudarsanam, N., and Grimm, C. Subtle gaze direction. *ACM Transactions on Graphics (TOG) 28*, 4 (2009), 100.

[20] Bala, P., Masu, R., Nisi, V., and Nunes, N. Cue control: Interactive sound spatialization for 360° videos. In *International Conference on Interactive Digital Storytelling* (2018), Springer, pp. 333–337.

[21] Bala, P., Masu, R., Nisi, V., and Nunes, N. " when the elephant trumps" a comparative study on spatial audio for orientation in 360° videos. In *Proc. of Conference on Human Factors in Computing Systems* (2019), pp. 1–13.

[22] Banerjee, A., and Dave, R. N. Validating clusters using the hopkins statistic. In *Proc. of IEEE International Conference on Fuzzy Systems* (2004), vol. 1, pp. 149–153.

[23] Bangert, A. S., Reuter-Lorenz, P. A., and Seidler, R. D. Dissecting the clock: Understanding the mechanisms of timing across tasks and temporal intervals. *Acta psychologica 136*, 1 (2011), 20–34.

[24] Baños, R. M., Botella, C., Alcañiz, M., Liaño, V., Guerrero, B., and Rey, B. Immersion and emotion: their impact on the sense of presence. *Cyberpsychology & behavior 7*, 6 (2004), 734–741.

[25] Bansal, A., Weech, S., and Barnett-Cowan, M. Movement-contingent time flow in virtual reality causes temporal recalibration. *Scientific reports 9*, 1 (2019), 1–13.

[26] Barghout, A., Cha, J., El Saddik, A., Kammerl, J., and Steinbach, E. Spatial resolution of vibrotactile perception on the human forearm when exploiting funneling illusion. In *IEEE Intern. Works. on Haptic Audiovis. Envs. and Games* (2009).

[27] Battista, J., Kalloniatis, M., and Metha, A. Visual function: The problem with eccentricity. *Clinical and Experimental Optometry 88*, 5 (2005), 313–321.

[28] Bell, S., Upchurch, P., Snavely, N., and Bala, K. Opensurfaces: A richly annotated catalog of surface appearance. *ACM Trans. on Graphics 32*, 4 (2013), 111.

[29] Bell, S., Upchurch, P., Snavely, N., and Bala, K. Material recognition in the wild with the materials in context database. In *Proc. Computer Vision and Pattern Recognition* (2015), pp. 3479–3487.

[30] Bertenthal, B. I., Banton, T., and Bradbury, A. Directional bias in the perception of translating patterns. *Perception 22*, 2 (1993), 193–207.

[31] Billger, M., and d'Elia, S. Color appearance in virtual reality: a comparison between a full-scale room and a virtual reality simulation. In *9th Congress of the International Color Association* (2002), International Society for Optics and Photonics, pp. 122–126.

[32] Blake, R. A neural theory of binocular rivalry. *Psychological Review 96*, 1 (1989), 145.

[33] Blanke, O., Slater, M., and Serino, A. Behavioral, neural, and computational principles of bodily self-consciousness. *Neuron 88*, 1 (2015), 145–166.

[34] Bohil, C. J., Alicea, B., and Biocca, F. A. Virtual reality in neuroscience research and therapy. *Nature reviews neuroscience 12*, 12 (2011), 752–762.

[35] Bolte, B., and Lappe, M. Subliminal reorientation and repositioning in immersive virtual environments using saccadic suppression. *IEEE Trans. on Visualization and Computer Graphics 21*, 4 (2015), 545–552.

[36] Bonneel, N., Suied, C., Viaud-Delmon, I., and Drettakis, G. Bimodal perception of audio-visual material properties for virtual environments. *ACM Trans. Appl. Percept. 7*, 1 (2010).

[37] Boud, A., Baber, C., and Steiner, S. Virtual reality: A tool for assembly? *Presence: Teleoperators & Virtual Environments 9*, 5 (2000), 486–496.

[38] Bouguila, L., Ishii, M., and Sato, M. Effect of coupling haptics and stereopsis on depth perception in virtual environment. In *Proc. of the 1st Workshop on Haptic Human Computer Interaction* (2000), pp. 54–62.

[39] Brown, A., Sheikh, A., Evans, M., and Watson, Z. Directing attention in 360-degree video. In *IBC 2016 Conference* (2016).

[40] Brown, S. W. Attentional resources in timing: Interference effects in concurrent temporal and nontemporal working memory tasks. *Perception & psychophysics 59*, 7 (1997), 1118–1140.

[41] Bruder, G., and Steinicke, F. Time perception during walking in virtual environments. In *2014 IEEE Virtual Reality (VR)* (2014), IEEE, pp. 67–68.

[42] Buhusi, C. V., and Meck, W. H. What makes us tick? functional and neural mechanisms of interval timing. *Nature reviews neuroscience 6*, 10 (2005), 755–765.

[43] Burns, E., Razzaque, S., Panter, A. T., Whitton, M. C., McCallus, M. R., and Brooks, F. P. The hand is slower than the eye: a quantitative exploration of visual dominance over proprioception. In *IEEE Proc. Virtual Reality* (2005), pp. 3–10.

[44] Çamci, A. Exploring the effects of diegetic and non-diegetic audiovisual cues on decision-making in virtual reality. In *SMC 2019. Proceedings of the 16th Sound and Music Computing Conference* (2019), pp. 28–31.

[45] Campos, J. L., Butler, J. S., and Bülthoff, H. H. Multisensory integration in the estimation of walked distances. *Experimental brain research 218*, 4 (2012), 551–565.

[46] Carlon, A. D. *Virtual Reality's Utility for Examining the Multimodal Perception of Heaviness.* PhD thesis, California State University, Fresno, 2018.

[47] Chadwick, A., and Kentridge, R. The perception of gloss: A review. *Vision Research 109* (2015), 221 – 235.

[48] Chalmers, A., Debattista, K., and Ramic-Brkic, B. Towards high-fidelity multi-sensory virtual environments. *The Visual Computer 25*, 12 (2009), 1101.

[49] Chang, E., Kim, H. T., and Yoo, B. Virtual reality sickness: a review of causes and measurements. *International Journal of Human–Computer Interaction 36*, 17 (2020), 1658–1682.

[50] Chao, F., Ozcinar, C., Wang, C., Zerman, E., Zhang, L., Hamidouche, W., Deforges, O., and Smolic, A. Audio-visual perception of omnidirectional video for virtual reality applications. In *IEEE Inter. Conf. on Mult. & Expo Workshops* (2020).

[51] Chauvel, G., Wulf, G., and Maquestiaux, F. Visual illusions can facilitate sport skill learning. *Psychonomic bulletin & review 22*, 3 (2015), 717–721.

[52] Checa, D., and Bustillo, A. A review of immersive virtual reality serious games to enhance learning and training. *Multimedia Tools and Applications* (2019), 1–27.

[53] Chen, K., Xu, K., Yu, Y., Wang, T.-Y., and Hu, S.-M. Magic decorator: automatic material suggestion for indoor digital scenes. *ACM Trans. on Graphics 34*, 6 (2015), 232.

[54] Cheng, D., Gong, Y., Zhou, S., Wang, J., and Zheng, N. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proc. Computer Vision and Pattern Recognition* (2016), pp. 1335–1344.

[55] Cheng, H., and Liu, S. Haptic force guided sound synthesis in multisensory virtual reality (vr) simulation for rigid-fluid interaction. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (2019), IEEE.

[56] Christopoulos, D., and Gaitatzes, A. Multimodal interfaces for educational virtual environments. In *2009 13th Panhellenic Conference on Informatics* (2009), IEEE, pp. 197–201.

[57] Cipresso, P., Albani, G., et al. Virtual multiple errands test (vmet): a virtual reality-based tool to detect early executive functions deficit in parkinson's disease. *Frontiers in Behavioral Neuroscience 8* (2014), 405.

[58] Cipresso, P., La Paglia, F., La Cascia, C., Riva, G., Albani, G., and La Barbera, D. Break in volition: A virtual reality study in patients with obsessive-compulsive disorder. *Experimental brain research 229*, 3 (2013), 443–449.

[59] Coelho, M., Ferreira, J. J., Dias, B., Sampaio, C., Martins, I. P., and Castro-Caldas, A. Assessment of time perception: The effect of aging. *Journal of the International Neuropsychological Society 10*, 3 (2004), 332–341.

[60] Cole, J., and Montero, B. Affective proprioception. *Janus Head 9*, 2 (2007), 299–317.

[61] Colley, A., Väyrynen, J., and Häkkilä, J. Skiing in a blended virtuality: an in-the-wild experiment. In *Proceedings of the 19th International Academic Mindtrek Conference* (2015), pp. 89–91.

[62] Corneil, B., Van Wanrooij, M., Munoz, D., and Van Opstal, A. Auditory-visual interactions subserving goal-directed saccades in a complex scene. *Journal of Neurophysiology 88* (2002), 438:454.

[63] Covaci, A., Ghinea, G., Lin, C., Huang, S., and Shih, J. Multisensory games-based learning-lessons learnt from olfactory enhancement of a digital board game. *Multimedia Tools and Applications 77*, 16 (2018).

[64] Creten, W., Vanpeperstraete, P., Van Camp, K., and Doclo, J. An experimental study on diphasic acoustic reflex patterns in normal ears. *Scandinavian Audiology 5*, 1 (1976), 3–8.

[65] Crison, F., Lecuyer, A., d'Huart, D., Burkhardt, J., Michel, G., and Dautin, J. Virtual technical trainer: Learning how to use milling machines with multi-sensory feedback in virtual reality. In *IEEE Proc. Virtual Reality* (2005), pp. 139–145.

[66] Cunningham, D., and Wallraven, C. *Experimental Design: From User Studies to Psychophysics*, 1st ed. A. K. Peters, Ltd., Natick, MA, USA, 2011.

[67] Cunningham, D. W., Wallraven, C., Fleming, R. W., and Strasser, W. Perceptual reparameterization of material properties. In *Proc. Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging* (2007), pp. 89–96.

[68] Dana, K. J., van Ginneken, B., Nayar, S. K., and Koenderink, J. J. Reflectance and texture of real-world surfaces. *ACM Trans. on Graphics 18*, 1 (Jan. 1999), 1–34.

[69] De Coensel, B., and Botteldooren, D. A model of saliency-based auditory attention to environmental sound. In *20th International Congress on Acoustics (ICA-2010)* (2010), pp. 1–8.

[70] De Coensel, B., Botteldooren, D., Berglund, B., and Nilsson, M. E. A computational model for auditory saliency of environmental sound. *The Journal of the Acoustical Society of America 125*, 4 (2009), 2528–2528.

[71] De Sa, A. G., and Zachmann, G. Virtual reality as a tool for verification of assembly and maintenance processes. *Computers & Graphics 23*, 3 (1999), 389–403.

[72] Debevec, P. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques* (1998), ACM, pp. 189–198.

[73] Delanoy, J., Lagunas, M., Galve, I., Gutierrez, D., Serrano, A., Fleming, R., and Masia, B. The role of objective and subjective measures in material similarity learning. In *ACM SIGGRAPH Posters* (2020).

[74] Delong, P., Aller, M., Giani, A. S., Rohe, T., Conrad, V., Watanabe, M., and Noppeney, U. Invisible flashes alter perceived sound location. *Scientific Reports 8*, 1 (2018), 12376.

[75] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proc. Computer Vision and Pattern Recognition* (2009), pp. 248–255.

[76] Deng, S., Kirkby, J. A., Chang, J., and Zhang, J. J. Multimodality with eye tracking and haptics: a new horizon for serious games? *International Journal of Serious Games 1*, 4 (2014), 17–34.

[77] Dichgans, J., and Brandt, T. Visual-vestibular interaction: Effects on self-motion perception and postural control. In *Perception*. Springer, 1978, pp. 755–804.

[78] Doerschner, K., Fleming, R. W., Yilmaz, O., Schrater, P. R., Hartung, B., and Kersten, D. Visual motion and the perception of surface material. *Current Biology 21*, 23 (2011), 2010–2016.

[79] Dorsey, J., Rushmeier, H., and Sillion, F. *Digital modeling of material appearance*. 2010.

[80] Duangudom, V., and Anderson, D. V. Using auditory saliency to understand complex auditory scenes. In *2007 15th European Signal Processing Conference* (2007), IEEE, pp. 1206–1210.

[81] Dumas, B., Lalanne, D., and Oviatt, S. Multimodal interfaces: A survey of principles, models and frameworks. In *Human machine interaction*. Springer, 2009, pp. 3–26.

[82] Dupuy, J., and Jakob, W. An adaptive parameterization for efficient material acquisition and rendering. *ACM Trans. on Graphics 37*, 6 (2018), 1–14.

[83] Eagleman, D. M. Using time perception to measure fitness for duty. *Military Psychology 21*, sup1 (2009), S123–S129.

[84] Eg, R., and Behne, D. M. Perceived synchrony for realistic and dynamic audiovisual events. *Frontiers in psychology 6* (2015), 736.

[85] Elbamby, M. S., Perfecto, C., Bennis, M., and Doppler, K. Toward low-latency and ultra-reliable virtual reality. *IEEE Network 32*, 2 (2018), 78–84.

[86] Emerge. Bringing touch and emotion to virtual experiences, 2021. Last accessed on 2021-11-02.

[87] Evangelopoulos, G., Zlatintsi, A., Potamianos, A., Maragos, P., Rapantzikos, K., Skoumas, G., and Avrithis, Y. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Trans. on Multimedia 15*, 7 (2013), 1553–1568.

[88] Fan, Y., Guthrie, A., and Levinson, D. Waiting time perceptions at transit stops and stations: Effects of basic amenities, gender, and security. *Transportation Research Part A: Policy and Practice 88* (2016), 251–264.

[89] Fei-Fei, L., Iyer, A., Koch, C., and Perona, P. What do we perceive in a glance of a real-world scene? *Journal of vision 7*, 1 (2007), 10–10.

[90] Feigl, T., Kõre, E., Mutschler, C., and Philippsen, M. Acoustical manipulation for redirected walking. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology* (2017), pp. 1–2.

[91] Feng, M., Dey, A., and Lindeman, R. W. The effect of multi-sensory cues on performance and experience during walking in immersive virtual environments. In *2016 IEEE Virtual Reality (VR)* (2016), IEEE, pp. 173–174.

[92] Fernandes, L., et al. Bringing user experience empirical data to gesture-control and somatic interaction in virtual reality videogames: an exploratory study with a multimodal interaction prototype. In *SciTecIn15-Conferência Ciências E Tecnologias Da Interação 2015* (2015).

[93] Feuchtner, T., and Müller, J. Extending the body for interaction with reality. In *Proceedings of the Conference on Human Factors in Computing Systems* (2017), pp. 5145–5157.

[94] Feuchtner, T., and Müller, J. Ownershift: Facilitating overhead interaction in virtual reality with an ownership-preserving hand space shift. In *Proc. of the 31st ACM Symposium on User Interface Software and Technology* (2018).

[95] Filip, J., and Vávra, R. Template-based sampling of anisotropic brdfs. In *Computer Graphics Forum* (2014), vol. 33, pp. 91–99.

[96] Fleming, R. W. Visual perception of materials and their properties. *Vision Research 94* (2014), 62–75.

[97] Fleming, R. W. Material perception. *Annual Review of Vision Science 3* (2017), 365–388.

[98] Fleming, R. W., Dror, R. O., and Adelson, E. H. Real-world illumination and the perception of surface reflectance properties. *Journal of Vision 3*, 5 (2003), 3–3.

[99] Fleming, R. W., Nishida, S., and Gegenfurtner, K. R. Perception of material properties. *Vision Research 115* (2015), 157 – 162.

[100] Fordell, H., Bodin, K., Eklund, A., and Malm, J. Rehatt–scanning training for neglect enhanced by multi-sensory stimulation in virtual reality. *Topics in stroke rehabilitation 23*, 3 (2016), 191–199.

[101] Fores, A., Ferwerda, J., and Gu, J. Toward a perceptually based metric for brdf modeling. In *Color and Imaging Conference* (2012), vol. 2012, Society for Imaging Science and Technology, pp. 142–148.

[102] Fraisse, P. The psychology of time.

[103] Freina, L., and Ott, M. A literature review on immersive virtual reality in education: state of the art and perspectives. In *The international scientific conference elearning and software for education* (2015), vol. 1, pp. 10–1007.

[104] Frens, M., and Van Opstal, A. Auditory-evoked saccades in two dimensions: dynamical characteristics, influence of eye position and sound spectrum. *Information Processing Underlying Gaze Control 12* (1994), 329.

[105] Frens, M. A., Van Opstal, A. J., and Van der Willigen, R. F. Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception & Psychophysics 57*, 6 (1995), 802–816.

[106] Gago, D., and Almeida, R. M. M. d. Effects of pleasant visual stimulation on attention, working memory, and anxiety in college students. *Psychology & Neuroscience 6* (2013), 351–355.

[107] Gallace, A., Ngo, M. K., Sulaitis, J., and Spence, C. Multisensory presence in virtual reality: possibilities & limitations. In *Multiple sensorial media advances and applications*. IGI Global, 2012, pp. 1–38.

[108] Gao, J., and Nevatia, R. Revisiting temporal modeling for video-based person reid. *arXiv preprint arXiv:1805.02104* (2018).

[109] Gao, R., Chen, C., Al-Halah, Z., Schissler, C., and Grauman, K. Visualechoes: Spatial image representation learning through echolocation. *arXiv preprint arXiv:2005.01616* (2020).

[110] Gao, R., and Grauman, K. 2.5d visual sound. In *Proc. of IEEE Conf. on Comp. Vision and Pattern Recogn.* (2019).

[111] Garces, E., Agarwala, A., Gutierrez, D., and Hertzmann, A. A Similarity Measure for Illustration Style. *ACM Transactions on Graphics (Proc. SIGGRAPH) 33*, 4 (2014).

[112] Georgoulis, S., Vanweddingen, V., Proesmans, M., and Van Gool, L. Material classification under natural illumination using reflectance maps. In *IEEE Winter Conference on Applications of Computer Vision* (2017), IEEE, pp. 244–253.

[113] Giordano, B. L., and McAdams, S. Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *The Journal of the Acoustical Society of America 119*, 2 (2006), 1171–1181.

[114] Gkioulekas, I., Walter, B., Adelson, E. H., Bala, K., and Zickler, T. On the appearance of translucent edges. In *Proc. Computer Vision and Pattern Recognition* (2015), pp. 5528–5536.

[115] Gkioulekas, I., Xiao, B., Zhao, S., Adelson, E. H., Zickler, T., and Bala, K. Understanding the role of phase function in translucent appearance. *ACM Transactions on graphics (TOG) 32*, 5 (2013), 1–19.

[116] Gonçalves, G., Melo, M., Vasconcelos-Raposo, J., and Bessa, M. Impact of different sensory stimuli on presence in credible virtual environments. *IEEE Trans. on Vis. and Computer Graphics 26*, 11 (2019).

[117] Gonzalez-Franco, M., Maselli, A., Florencio, D., Smolyanskiy, N., and Zhang, Z. Concurrent talking in immersive virtual reality: on the dominance of visual speech cues. *Scientific reports 7*, 1 (2017), 1–11.

[118] Gopher, D. Skill training in multimodal virtual environments. *Work 41*, Supplement 1 (2012), 2284–2287.

[119] Gospodarek, M., Genovese, A., Dembeck, D., Brenner, C., Roginska, A., and Perlin, K. Sound design and reproduction techniques for co-located narrative vr experiences. In *Audio Engineering Society Convention 147* (2019).

[120] Groehn, M., Lokki, T., Savioja, L., and Takala, T. Some aspects of role of audio in immersive visualization. In *Visual Data Exploration and Analysis VIII* (2001), vol. 4302, International Society for Optics and Photonics.

[121] Gugenheimer, J., Wolf, D., Haas, G., Krebs, S., and Rukzio, E. Swivrchair: A motorized swivel chair to nudge users' orientation for 360 degree storytelling in virtual reality. In *Proc. of the Conf. on Human Factors in Comp. Systems* (2016).

[122] Gürkök, H., and Nijholt, A. Brain–computer interfaces for multimodal interaction: A survey and principles. *International Journal of Human-Computer Interaction 28*, 5 (2012), 292–307.

[123] Guttentag, D. A. Virtual reality: Applications and implications for tourism. *Tourism Management 31*, 5 (2010), 637–651.

[124] Ha, S., Kim, L., Park, S., Jun, C.-s., and Rho, H. Virtual prototyping enhanced by a haptic interface. *CIRP annals 58*, 1 (2009), 135–138.

[125] Harders, M., Bianchi, G., and Knoerlein, B. Multimodal augmented reality in medicine. In *International Conference on Universal Access in Human-Computer Interaction* (2007), Springer, pp. 652–658.

[126] Hart, S. G. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (2006), vol. 50, Sage publications Sage CA: Los Angeles, CA, pp. 904–908.

[127] Havran, V., Filip, J., and Myszkowski, K. Perceptually Motivated BRDF Comparison using Single Image. *Computer Graphics Forum* (2016).

[128] Hayashi, O., Fujita, K., Takashima, K., Lindernan, R., and Kitarnura, Y. Redirected jumping: Imperceptibly manipulating jump motions in virtual reality. In *IEEE Conf. on Virtual Reality and 3D User Interfaces* (2019).

[129] Hayman, E., Caputo, B., Fritz, M., and Eklundh, J.-O. On the significance of real-world conditions for material classification. In *Proc. European Conference on Computer Vision* (2004), Springer, pp. 253–266.

[130] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CoRR abs/1512.03385* (2015).

[131] Hecht, D., Reiner, M., and Halevy, G. Multimodal virtual environments: response times, attention, and presence. *Presence: Teleoperators and virtual environments 15*, 5 (2006), 515–523.

[132] Hentschke, H. Harald's toolbox to compute measures of effect sizes in matlab. https://www.mathworks.com/matlabcentral/fileexchange/32398-hhentschke-measures-of-effect-size-toolbox. Accessed: 2021-12-09.

[133] Herrera, N. S., and McMahan, R. P. Development of a simple and low-cost olfactory display for immersive media experiences. In *Proceedings of the ACM International Workshop on Immersive Media Experiences* (2014), pp. 1–6.

[134] Hessels, R. S., Niehorster, D. C., Kemner, C., and Hooge, I. T. Noise-robust fixation detection in eye movement data: Identification by two-means clustering (i2mc). *Behavior research methods 49*, 5 (2017), 1802–1823.

[135] Hidaka, S., and Ide, M. Sound can suppress visual perception. *Scientific Reports 5*, 1 (2015), 1–9.

[136] Hidaka, S., Suzuishi, Y., Ide, M., and Wada, M. Effects of spatial consistency and individual difference on touch-induced visual suppression effect. *Scientific Reports 8*, 1 (2018), 17018.

[137] Ho, Y.-X., Landy, M. S., and Maloney, L. T. How direction of illumination affects visually perceived surface roughness. *Journal of Vision 6*, 5 (2006), 8–8.

[138] Hoffman, H. G. Physically touching virtual objects using tactile augmentation enhances the realism of virtual environments. In *Proceedings. IEEE 1998 Virtual Reality Annual International Symposium* (1998), IEEE, pp. 59–63.

[139] Hoffman, H. G., Garcia-Palacios, A., Carlin, A., Furness Iii, T. A., and Botella-Arbona, C. Interfaces that heal: coupling real and virtual objects to treat spider phobia. *Intern. Journal of HCI 16*, 2 (2003).

[140] Hoffman, H. G., Hollander, A., Schroder, K., Rousseau, S., and Furness, T. Physically touching and tasting virtual objects enhances the realism of virtual experiences. *Virtual Reality 3*, 4 (1998), 226–234.

[141] Holmes, N. P., and Spence, C. Multisensory integration: space, time and superadditivity. *Current Biology 15*, 18 (2005), R762–R764.

[142] Howard, I., Jenkin, H., and Hu, G. Visually-induced reorientation illusions as a function of age. *Aviation, space, and environmental medicine 71*, 9 Suppl (2000), A87–91.

[143] Hu, P., Sun, Q., Didyk, P., Wei, L.-Y., and Kaufman, A. E. Reducing simulator sickness with perceptual camera control. *ACM Trans. on Graphics (TOG) 38*, 6 (2019), 1–12.

[144] Hu, Z., Li, S., Zhang, C., Yi, K., Wang, G., and Manocha, D. Dgaze: Cnn-based gaze prediction in dynamic scenes. *IEEE Trans. on Visualization and Computer Graphics 26*, 5 (2020), 1902–1911.

[145] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proc. Computer Vision and Pattern Recognition* (2017), vol. 1, p. 3.

[146] Huang, H., Solah, M., Li, D., and Yu, L.-F. Audible panorama: Automatic spatial audio generation for panorama imagery. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (2019).

[147] Hutchins, M., Adcock, M., Stevenson, D., Gunn, C., and Krumpholz, A. The design of perceptual representations for practical networked multimodal virtual training environments. In *the Proceedings of the 11th International Conference on Human-Computer Interaction: HCI International'05* (2005).

[148] Iachini, T., Maffei, L., Ruotolo, F., Senese, V., Ruggiero, G., Masullo, M., and Alekseeva, N. Multisensory assessment of acoustic comfort aboard metros: a virtual reality study. *Applied Cognitive Psychology 26*, 5 (2012).

[149] Ichimura, A., Nakajima, I., and Juzoji, H. Investigation and analysis of a reported incident resulting in an actual airline hijacking due to a fanatical and engrossed vr state. *CyberPsychology & Behavior 4*, 3 (2001).

[150] Ide, M., and Hidaka, S. Tactile stimulation can suppress visual perception. *Scientific reports 3*, 1 (2013), 1–8.

[151] Ide, M., Hidaka, S., Ikeda, H., and Wada, M. Neural mechanisms underlying touch-induced visual perceptual suppression: An fmri study. *Scientific reports 6*, 1 (2016), 1–9.

[152] Ito, H. Cortical shape adaptation transforms a circle into a hexagon: A novel afterimage illusion. *Psychological Science 23*, 2 (2012), 126–132.

[153] Jaimes, A., and Sebe, N. Multimodal human–computer interaction: A survey. *Computer vision and image understanding 108*, 1-2 (2007), 116–134.

[154] Jarabo, A., Wu, H., Dorsey, J., Rushmeier, H., and Gutierrez, D. Effects of approximate filtering on the appearance of bidirectional texture functions. *IEEE Trans. on Visualization and Computer Graphics 20*, 6 (2014).

[155] Je, S., Kim, M. J., Lee, W., Lee, B., Yang, X.-D., Lopes, P., and Bianchi, A. Aero-plane: A handheld force-feedback device that renders weight motion illusion on a virtual 2d plane. In *Proceedings of the ACM Symposium on User Interface Software and Technology* (2019), pp. 763–775.

[156] Jensen, L., and Konradsen, F. A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies 23*, 4 (2018), 1515–1529.

[157] Jung, S., Wood, A., Hoermann, S., Abhayawardhana, P., and Lindeman, R. The impact of multi-sensory stimuli on confidence levels for perceptual-cognitive tasks in vr. In *IEEE Conf. on Virtual Reality and 3D User Interfaces* (2020).

[158] Kawashima, R., O'Sullivan, B. T., and Roland, P. E. Positron-emission tomography studies of cross-modality inhibition in selective attentional tasks: closing the" mind's eye". *Proceedings of the National Academy of Sciences 92*, 13 (1995), 5969–5972.

[159] Kaya, N., and Epps, H. H. Relationship between color and emotion: A study of college students. *College student journal 38*, 3 (2004), 396–405.

[160] Kayser, C., Petkov, C. I., Lippert, M., and Logothetis, N. K. Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology 15*, 21 (2005), 1943–1947.

[161] Kendall, M., and Gibbons, J. D. *Rank Correlation Methods*, 5 ed. A Charles Griffin Title, September 1990.

[162] Kendall, M. G., and Babington-Smith, B. On the method of paired comparisons. *Biometrica 31* (1940), 324–345.

[163] Kerr, W. B., and Pellacini, F. Toward evaluating material design interface paradigms for novice users. In *ACM SIGGRAPH 2010 Papers* (2010), ACM, pp. 35:1–35:10.

[164] Keshavarz, B., Hettinger, L. J., Vena, D., and Campos, J. L. Combined effects of auditory and visual cues on the perception of vection. *Experimental brain research 232*, 3 (2014), 827–836.

[165] Kilteni, K., Groten, R., and Slater, M. The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments 21*, 4 (2012), 373–387.

[166] Kim, K., Gu, J., Tyree, S., Molchanov, P., Niessner, M., and Kautz, J. A lightweight approach for on-the-fly reflectance estimation. In *Proc. International Conference on Computer Vision* (2017), pp. 20–28.

[167] Kimura, Z., and Sato, M. Auditory stimulation on touching a virtual object outside a user's field of view. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (2020), IEEE.

[168] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[169] Koelewijn, T., Bronkhorst, A., and Theeuwes, J. Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta psychologica 134*, 3 (2010), 372–384.

[170] Kouakoua, K., Duclos, C., Aissaoui, R., Nadeau, S., and Labbe, D. R. Rhythmic proprioceptive stimulation improves embodiment in a walking avatar when added to visual stimulation. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (2020), IEEE, pp. 573–574.

[171] Koutek, C. D. M., and Koutek, M. Scientific visualization in virtual reality: interaction techniques and application development.

[172] Křivánek, J., Ferwerda, J. A., and Bala, K. Effects of global illumination approximations on material appearance. *ACM Trans. on Graphics (Proc. SIGGRAPH) 29*, 4 (2010), 112:1–112:10.

[173] Kruijff, E., Marquardt, A., Trepkowski, C., Lindeman, R., Hinkenjann, A., Maiero, J., and Riecke, B. On your feet! enhancing vection in leaning-based interfaces through multisensory stimuli. In *Proc. Symp. on Spatial User Interact.* (2016).

[174] Kruijff, E., Marquardt, A., Trepkowski, C., Schild, J., and Hinkenjann, A. Enhancing user engagement in immersive games through multisensory cues. In *International Conference on Games and Virtual Worlds for Serious Applications* (2015).

[175] Kytö, M., Kusumoto, K., and Oittinen, P. The ventriloquist effect in augmented reality. In *2015 IEEE International Symposium on Mixed and Augmented Reality* (2015), IEEE, pp. 49–53.

[176] Lagunas, M., Garces, E., and Gutierrez, D. Learning icons appearance similarity. *Multimedia Tools and Applications* (2018), 1–19.

[177] Lagunas, M., Malpica, S., Serrano, A., Garces, E., Gutierrez, D., and Masia, B. A similarity measure for material appearance. *ACM Trans. on Graphics (Proc. SIGGRAPH) 38*, 4 (2019).

[178] Lagunas, M., Serrano, A., Gutierrez, D., and Masia, B. The joint role of geometry and illumination on material recognition. *Journal of Vision 21*, 2 (2021), 2–2.

[179] Lalanne, C., and Lorenceau, J. Crossmodal integration for perception and action. *Journal of Physiology-Paris 98*, 1-3 (2004), 265–279.

[180] Langbehn, E., Steinicke, F., Lappe, M., Welch, G., and Bruder, G. In the blink of an eye: leveraging blink-induced suppression for imperceptible position and orientation redirection in virtual reality. *ACM Trans. on Graph.* (2018).

[181] Larsen, C. R., Soerensen, J. L., Grantcharov, T. P., Dalsgaard, T., Schouenborg, L., Ottosen, C., Schroeder, T. V., and Ottesen, B. S. Effect of virtual reality training on laparoscopic surgery: randomised controlled trial. *Bmj 338* (2009), b1802.

[182] Lathan, C., Tracey, M., Sebrechts, M., Clawson, D., and Higgins, G. Using virtual environments as training simulators: Measuring transfer. *Handbook of Virtual Environments: Design, Implementation, and Applications* (2002).

[183] Laurienti, P. J., Burdette, J. H., Wallace, M. T., Yen, Y.-F., Field, A. S., and Stein, B. E. Deactivation of sensory-specific cortex by cross-modal stimuli. *Journal of cognitive neuroscience 14*, 3 (2002), 420–429.

[184] LaViola Jr, J. J. A discussion of cybersickness in virtual environments. *ACM Sigchi Bulletin 32*, 1 (2000).

[185] Lawson, G., Roper, T., and Abdullah, C. Multimodal "sensory illusions" for improving spatial awareness in virtual environments. In *Proc. of the European Conference on Cognitive Ergonomics* (2016).

[186] Laycock, S. D., and Day, A. A survey of haptic rendering techniques. In *Computer Graphics Forum* (2007), vol. 26, Wiley Online Library, pp. 50–65.

[187] Leone, L. M., and McCourt, M. E. The roles of physical and physiological simultaneity in audiovisual multisensory facilitation. *i-Perception 4*, 4 (2013), 213–228.

[188] Leung, T., and Malik, J. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision 43*, 1 (2001), 29–44.

[189] Lewis, P. A., and Miall, R. C. Brain activation patterns during measurement of sub-and supra-second intervals. *Neuropsychologia 41*, 12 (2003), 1583–1592.

[190] Lewis, P. A., and Miall, R. C. Distinct systems for automatic and cognitively controlled time measurement: evidence from neuroimaging. *Current opinion in neurobiology 13*, 2 (2003), 250–255.

[191] Li, L., Yu, F., Shi, D., Shi, J., Tian, Z., Yang, J., Wang, X., and Jiang, Q. Application of virtual reality technology in clinical medicine. *American journal of translational research 9*, 9 (2017).

[192] Liao, H., Xie, N., Li, H., Li, Y., Su, J., Jiang, F., Huang, W., and Shen, H. T. Data-driven spatio-temporal analysis via multi-modal zeitgebers and cognitive load in vr. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (2020), IEEE, pp. 473–482.

[193] Lin, Y.-C., Chang, Y.-J., Hu, H.-N., Cheng, H.-T., Huang, C.-W., and Sun, M. Tell me where to look: Investigating ways for assisting focus in 360 video. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), pp. 2535–2545.

[194] Lin, Y.-T., Liao, Y.-C., Teng, S.-Y., Chung, Y.-J., Chan, L., and Chen, B.-Y. Outside-in: visualizing out-of-sight regions-of-interest in a 360 video using spatial picture-in-picture previews. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (2017), pp. 255–265.

[195] Liu, P., Forte, J., Sewell, D., and Carter, O. Cognitive load effects on early visual perceptual processing. *Attention, Perception, & Psychophysics 80*, 4 (2018), 929–950.

[196] Lombardi, S., and Nishino, K. Reflectance and natural illumination from a single image. In *Proc. European Conference on Computer Vision* (2012), pp. 582–595.

[197] Long, B., Seah, S. A., Carter, T., and Subramanian, S. Rendering volumetric haptic shapes in mid-air using ultrasound. *ACM Trans. on Graphics (TOG) 33*, 6 (2014), 1–10.

[198] Lopes, P., You, S., Ion, A., and Baudisch, P. Adding force feedback to mixed reality experiences and games using electrical muscle stimulation. In *Proc. of the Conference on Human Factors in Computing Systems* (2018), pp. 1–13.

[199] Lovelace, C., Stein, B., and Wallace, M. An irrelevant light enhances auditory detection in humans: a psychophysical analysis of multisensory integration in stimulus detection. *Cognitive brain research 17*, 2 (2003), 447–453.

[200] Lu, F., Chen, X., Sato, I., and Sato, Y. Symps: Brdf symmetry guided photometric stereo for shape and light source estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence 40*, 1 (2018), 221–234.

[201] Lu, J., Li, L., and Sun, G. P. A multimodal virtual anatomy e-learning tool for medical education. In *International Conference on Technologies for E-Learning and Digital Entertainment* (2010), Springer, pp. 278–287.

[202] Lun, Z., Kalogerakis, E., and Sheffer, A. Elements of style: learning perceptual shape style similarity. *ACM Trans. on Graphics 34*, 4 (2015), 84.

[203] Macaluso, E., Frith, C. D., and Driver, J. Modulation of human visual cortex by cross-modal spatial attention. *Science 289*, 5482 (2000), 1206–1208.

[204] MacDonald, J. A., Balakrishnan, J., Orosz, M. D., and Karplus, W. J. Intelligibility of speech in a virtual 3-d environment. *Human Factors 44*, 2 (2002), 272–286.

[205] Maculewicz, J., Nilsson, N. C., and Serafin, S. An investigation of the effect of immersive visual and auditory feedback on rhythmic walking interaction. In *Proceedings of the Audio Mostly 2016*. 2016, pp. 194–201.

[206] Magalhães, E., Jacob, J., Nilsson, N., Nordahl, R., and Bernardes, G. Physics-based concatenative sound synthesis of photogrammetric models for aural and haptic feedback in virtual environments. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (2020), pp. 376–379.

[207] Maggioni, E., Cobden, R., Dmitrenko, D., and Obrist, M. Smell-o-message: integration of olfactory notifications into a messaging application to improve users' performance. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (2018), pp. 45–54.

[208] Maloney, L. T., and Brainard, D. H. Color and material perception: Achievements and challenges. *Journal of Vision 10*, 9 (2010), 19–19.

[209] Malpica, S., Masia, B., Herman, L., Wetzstein, G., Eagleman, D., Gutierrez, D., Bylinskii, Z., and Sun, Q. Has half the time passed? investigating time perception at long time scales. *Journal of Vision 20*, 11 (2020), 489–489.

[210] Malpica, S., Masia, B., Herman, L., Wetzstein, G., Eagleman, D. M., Gutierrez, D., Bylinskii, Z., and Sun, Q. Larger visual changes compress time: The inverted effect of asemantic visual features on interval time perception. *PloS one 17*, 3 (2022), e0265591.

[211] Malpica, S., Serrano, A., Allue, M., Bedia, M., and Masia, B. Crossmodal perception in virtual reality. *Multimedia Tools and Applications* (2019), 1–21.

[212] Malpica, S., Serrano, A., Allue, M., Bedia, M., and Masia, B. Crossmodal perception in virtual reality. *Multimedia Tools and Applications 79*, 5 (2020), 3311–3331.

[213] Malpica, S., Serrano, A., Gutierrez, D., and Masia, B. Auditory stimuli degrade visual performance in virtual reality. *Scientific Reports 10* (2020).

[214] Mantiuk, R., Daly, S., and Kerofsky, L. Display adaptive tone mapping. In *ACM Trans. on Graphics* (2008), vol. 27, ACM, p. 68.

[215] Marañes, C., Gutierrez, D., and Serrano, A. Exploring the impact of 360° movie cuts in users' attention. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (2020), IEEE, pp. 73–82.

[216] Marbury, D. What does the future hold for ar and vr in healthcare?, 2021. Last accessed on 2021-11-02.

[217] Marchal, M., Gallagher, G., Lécuyer, A., and Pacchierotti, C. Can stiffness sensations be rendered in virtual reality using mid-air ultrasound haptic technologies? In *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications* (2020), Springer, pp. 297–306.

[218] Martin, D., Malpica, S., Gutierrez, D., Masia, B., and Serrano, A. Multimodality in vr: A survey. *ACM Computing Surveys (CSUR) 54*, 10s (2022), 1–36.

[219] Martin, D., Serrano, A., Bergman, A. W., Wetzstein, G., and Masia, B. Scangan360: A generative model of realistic scanpaths for 360° images. *IEEE Trans. on Vis. and Computer Graphics* (2022).

[220] Martin, D., Serrano, A., and Masia, B. Panoramic convolutions for 360° single-image saliency prediction. In *CVPR Workshop on Computer Vision for Augmented and Virtual Reality* (2020).

[221] Martín, R., Iseringhausen, J., Weinmann, M., and Hullin, M. B. Multimodal perception of material properties. In *Proceedings of the ACM SIGGRAPH symposium on applied perception* (2015), ACM, pp. 33–40.

[222] Martínez, J., García, A., Oliver, M., Molina, J., and González, P. Vitaki: a vibrotactile prototyping toolkit for virtual reality and video games. *International Journal of Human-Computer Interaction 30*, 11 (2014).

[223] Maselli, A., and Slater, M. The building blocks of the full body ownership illusion. *Frontiers in human neuroscience 7* (2013), 83.

[224] Masia, B., Camon, J., Gutierrez, D., and Serrano, A. Influence of directional sound cues on users exploration across 360 movie cuts. *IEEE Computer Graphics and Applications* (2021), 1–1.

[225] Masia, B., Wetzstein, G., Didyk, P., and Gutierrez, D. A Survey on Computational Displays: Pushing the Boundaries of Optics, Computation, and Perception. *Computers & Graphics 37*, 8 (2013), 1012 – 1038.

[226] Mast, F. W., and Oman, C. M. Top-down processing and visual reorientation illusions in a virtual reality environment. *Swiss Journal of Psychology/Schweizerische Zeitschrift für Psychologie 63*, 3 (2004), 143.

[227] Matin, E. Saccadic suppression: a review and an analysis. *Psychological Bulletin 81*, 12 (1974), 899.

[228] Matsuda, Y., Nakamura, J., Amemiya, T., Ikei, Y., and Kitazaki, M. Perception of walking self-body avatar enhances virtual-walking sensation. In *IEEE Conf. on Virtual Reality and 3D User Interfaces Abstracts and Workshops* (2020).

[229] Matsumoto, K., Langbehn, E., Narumi, T., and Steinicke, F. Detection thresholds for vertical gains in vr and drone-based telepresence systems. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (2020), pp. 101–107.

[230] Matsumoto, K., Narumi, T., Ban, Y., Yanase, Y., Tanikawa, T., and Hirose, M. Unlimited corridor: A visuo-haptic redirection system. In *International Conference on Virtual-Reality Continuum and its Applications in Industry* (2019), pp. 1–9.

[231] Matusik, W., Ajdin, B., Gu, J., Lawrence, J., Lensch, H. P., Pellacini, F., and Rusinkiewicz, S. Printing spatially-varying reflectance. *ACM Trans. on Graphics (Proc. SIGGRAPH Asia) 28*, 5 (2009).

[232] Matusik, W., Pfister, H., Brand, M., and McMillan, L. A data-driven reflectance model. *ACM Trans. on Graphics 22*, 3 (2003), 759–769.

[233] McDonald, J. J., Teder-SaÈlejaÈrvi, W. A., and Hillyard, S. A. Involuntary orienting to sound improves visual perception. *Nature 407*, 6806 (2000), 906–908.

[234] McFee, B., and Lanckriet, G. Learning Multi-modal Similarity. *Journal of Machine Learning Research 12* (2011), 491–523.

[235] McGill, M., Ng, A., and Brewster, S. I am the passenger: how visual motion cues can influence sickness for in-car vr. In *Proceedings of the Conference on Human Factors in Computing Systems* (2017), pp. 5655–5668.

[236] McInnes, L., and Healy, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[237] McNamara, A., Mania, K., and Gutierrez, D. Perception in graphics, visualization, virtual environments and animation. SIGGRAPH Asia Courses, 2011.

[238] Melo, M., Gonçalves, G., Monteiro, P., Coelho, H., Vasconcelos-Raposo, J., and Bessa, M. Do multisensory stimuli benefit the virtual reality experience? a systematic review. *IEEE Trans. on Vis. and Computer Graphics* (2020).

[239] Merabet, L. B., Swisher, J. D., McMains, S. A., Halko, M. A., Amedi, A., Pascual-Leone, A., and Somers, D. C. Combined activation and deactivation of visual cortex during tactile sensory processing. *Journal of neurophysiology 97*, 2 (2007), 1633–1641.

[240] Meyer, G. F., Shao, F., White, M. D., Hopkins, C., and Robotham, A. J. Modulation of visually evoked postural responses by contextual visual, haptic and auditory information: a 'virtual reality check'. *PloS one 8*, 6 (2013).

[241] Middlebrooks, J. C., and Green, D. M. Sound localization by human listeners. *Annual review of psychology 42*, 1 (1991), 135–159.

[242] Min, X., Zhai, G., Zhou, J., Zhang, X., Yang, X., and Guan, X. A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Trans. on Image Processing 29* (2020), 3805–3819.

[243] Miranda, M. I. Taste and odor recognition memory: the emotional flavor of life. *Reviews in the Neurosciences 23*, 5-6 (2012), 481–499.

[244] Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence 42*, 2 (2019), 502–508.

[245] Montague, W. P. A theory of time-perception. *The American Journal of Psychology 15*, 1 (1904), 1–13.

[246] Morgado, P., Nvasconcelos, N., Langlois, T., and Wang, O. Self-supervised generation of spatial audio for 360 video. In *Advances in Neural Information Processing Systems* (2018), pp. 362–372.

[247] Mozolic, J. L., Joyner, D., Hugenschmidt, C. E., Peiffer, A. M., Kraft, R. A., Maldjian, J. A., and Laurienti, P. J. Cross-modal deactivations during modality-specific selective attention. *BMC neurology 8*, 1 (2008), 35.

[248] Mueller, F., Kari, T., Khot, R., Li, Z., Wang, Y., Mehta, Y., and Arnold, P. Towards experiencing eating as a form of play. In *Proc. of the Symposium on Computer-Human Interaction in Play Companion Extended Abstracts* (2018).

[249] Mühlberger, A., Weik, A., Pauli, P., and Wiedemann, G. One-session virtual reality exposure treatment for fear of flying: 1-year follow-up and graduation flight accompaniment effects. *Psychotherapy Research 16*, 1 (2006).

[250] Muhlberger, A., Wiedemann, G., and Pauli, P. Efficacy of a one-session virtual reality exposure treatment for fear of flying. *Psychotherapy Research 13*, 3 (2003), 323–336.

[251] Mylo, M., Giesel, M., Zaidi, Q., Hullin, M., and Klein, R. Appearance bending: A perceptual editing paradigm for data-driven material models. *Vision, Modeling and Visualization. The Eurographics Association* (2017).

[252] Nakajima, J., Sugimoto, A., and Kawamoto, K. Incorporating audio signals into constructing a visual saliency map. In *Pacific-Rim Symposium on Image and Video Technology* (2013), Springer, pp. 468–480.

[253] Nesbitt, K. V., and Hoskens, I. Multi-sensory game interface improves player satisfaction but not performance. In *Proceedings of the ninth conference on Australasian user interface-Volume 76* (2008), pp. 13–18.

[254] Neyret, S., Navarro, X., Beacco, A., Oliva, R., Bourdin, P., Valenzuela, J., Barberia, I., and Slater, M. An embodied perspective as a victim of sexual harassment in virtual reality reduces action conformity in a later milgram obedience scenario. *Scientific Reports 10*, 1 (2020), 1–18.

[255] Ngan, A., Durand, F., and Matusik, W. Experimental Analysis of BRDF Models. In *Eurographics Symposium on Rendering* (2005), The Eurographics Association.

[256] Ngan, A., Durand, F., and Matusik, W. Image-driven navigation of analytical brdf models. In *Rendering Techniques* (2006), pp. 399–407.

[257] Nidiffer, A. R., Diederich, A., Ramachandran, R., and Wallace, M. T. Multisensory perception reflects individual differences in processing temporal correlations. *Scientific Reports 8*, 1 (2018), 1–15.

[258] Nielsen, L. T., Møller, M. B., Hartmeyer, S. D., Ljung, T. C., Nilsson, N. C., Nordahl, R., and Serafin, S. Missing the point: an exploration of how to guide users' attention during cinematic virtual reality. In *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology* (2016), pp. 229–232.

[259] Nilsson, N. C., Nordahl, R., Sikström, E., Turchet, L., and Serafin, S. Haptically induced illusory self-motion and the influence of context of motion. In *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications* (2012), Springer, pp. 349–360.

[260] Nilsson, N. C., Peck, T., Bruder, G., Hodgson, E., Serafin, S., Whitton, M., Steinicke, F., and Rosenberg, E. S. 15 years of research on redirected walking in immersive virtual environments. *IEEE Computer Graphics and Applications 38*, 2 (2018), 44–56.

[261] Nilsson, N. C., Suma, E., Nordahl, R., Bolas, M., and Serafin, S. Estimation of detection thresholds for audiovisual rotation gains. In *2016 IEEE Virtual Reality (VR)* (2016), IEEE, pp. 241–242.

[262] Noesselt, T., Bergmann, D., Hake, M., Heinze, H.-J., and Fendrich, R. Sound increases the saliency of visual events. *Brain Research 1220* (2008), 157–163.

[263] Nogalski, M., and Fohl, W. Acoustic redirected walking with auditory cues by means of wave field synthesis. In *IEEE Virtual Reality* (2016).

[264] Nogalski, M., and Fohl, W. Curvature gains in redirected walking: A closer look. In *2017 IEEE Virtual Reality (VR)* (2017), IEEE, pp. 267–268.

[265] Normand, J.-M., Giannopoulos, E., Spanlang, B., and Slater, M. Multisensory stimulation can induce an illusion of larger belly size in immersive virtual reality. *PloS one 6*, 1 (2011).

[266] Open-AI. Dall-e 2, a new text to image deep learning model. https://openai.com/dall-e-2/. Accessed: 2021-10-17.

[267] Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E. H., and Freeman, W. T. Visually indicated sounds. *CoRR abs/1512.08512* (2015).

[268] Paek, S. *The impact of multimodal virtual manipulatives on young children's mathematics learning*. PhD thesis, Teachers College, Columbia University, 2012.

[269] Pariyadath, V., and Eagleman, D. The effect of predictability on subjective duration. *PloS one 2*, 11 (2007), e1264.

[270] Patney, A., Salvi, M., Kim, J., Kaplanyan, A., Wyman, C., Benty, N., Luebke, D., and Lefohn, A. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics (TOG) 35*, 6 (2016), 179.

[271] Pavel, A., Hartmann, B., and Agrawala, M. Shot orientation controls for interactive cinematography with 360 video. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (2017).

[272] Peatfield, N., Mueller, N., Ruhnau, P., and Weisz, N. Rubin-vase illusion perception is predicted by prestimulus activity and connectivity. *Journal of vision 15*, 12 (2015), 429–429.

[273] Peck, T. C., Fuchs, H., and Whitton, M. C. Evaluation of reorientation techniques and distractors for walking in large virtual environments. *IEEE Trans. on Visualization and Computer Graphics 15*, 3 (2009).

[274] Pellacini, F., Ferwerda, J. A., and Greenberg, D. P. Toward a psychophysically-based light reflection model for image synthesis. In *Proc. on Computer Graphics and Interactive Techniques* (2000), pp. 55–64.

[275] Pereira, T., and Rusinkiewicz, S. Gamut mapping spatially varying reflectance with an improved BRDF similarity metric. In *Computer Graphics Forum* (2012), vol. 31, Wiley Online Library, pp. 1557–1566.

[276] Petkova, V., and Ehrsson, H. If i were you: perceptual illusion of body swapping. *PloS one 3*, 12 (2008).

[277] Pezent, E., O'Malley, M. K., Israr, A., Samad, M., Robinson, S., Agarwal, P., Benko, H., and Colonnese, N. Explorations of wrist haptic feedback for ar/vr interactions with tasbi. In *Extended Abstracts of the Conference on Human Factors in Computing Systems* (2020), pp. 1–4.

[278] Polti, I., Martin, B., and van Wassenhove, V. The effect of attention and working memory on the estimation of elapsed time. *Scientific reports 8*, 1 (2018), 1–11.

[279] Pont, S. C., and Te Pas, S. F. Material—illumination ambiguities and the perception of solid objects. *Perception 35*, 10 (2006), 1331–1350.

[280] Poupyrev, I., Ichikawa, T., Weghorst, S., and Billinghurst, M. Egocentric object manipulation in virtual environments: empirical evaluation of interaction techniques. In *Computer Graphics Forum* (1998), vol. 17.

[281] Powers Iii, A. R., Hillock-Dunn, A., and Wallace, M. T. Generalization of multisensory perceptual learning. *Scientific Reports 6* (2016), 23374.

[282] Poyade, M. Motor skill training using virtual reality and haptic interaction–a case study in industrial maintenance. *MÁLAGA* (2013).

[283] Pozeg, P., Galli, G., and Blanke, O. Those are your legs: the effect of visuo-spatial viewpoint on visuo-tactile integration and body ownership. *Frontiers in psychology 6* (2015), 1749.

[284] Prange, A., Barz, M., and Sonntag, D. Medical 3d images in multimodal virtual reality. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion* (2018), pp. 1–2.

[285] Prinz, J. J. Is the mind really modular? In *Contemporary Debates in Cognitive Science*. 2006, pp. 22–36.

[286] Ramachandran, V. S., and Rogers-Ramachandran, D. Synaesthesia in phantom limbs induced with mirrors. *Proceedings of the Royal Society of London. Series B: Biological Sciences 263*, 1369 (1996), 377–386.

[287] Ramanarayanan, G., Ferwerda, J., Walter, B., and Bala, K. Visual equivalence: Towards a new standard for image fidelity. *ACM Trans. Graph. 26*, 3 (july 2007).

[288] Ranasinghe, N., Cheok, A., Nakatsu, R., and Do, E. Y.-L. Simulating the sensation of taste for immersive experiences. In *Proceedings of the ACM International Workshop on Immersive Media Experiences* (2013).

[289] Ranasinghe, N., Jain, P., Thi Ngoc Tram, N., Koh, K., Tolley, D., Karwita, S., Lien-Ya, L., Liangkun, Y., Shamaiah, K., Eason Wai Tung, C., et al. Season traveller: Multisensory narration for enhancing the virtual reality experience. In *Proceedings of the Conference on Human Factors in Computing Systems* (2018), pp. 1–13.

[290] Reddi, S. J., Kale, S., and Kumar, S. On the convergence of adam and beyond.

[291] Rewkowski, N., Rungta, A., Whitton, M., and Lin, M. Evaluating the effectiveness of redirected walking with auditory distractors for navigation in virtual environments. In *IEEE Conf. on Virtual Reality and 3D User Interfaces* (2019).

[292] Richard, E., Tijou, A., and Richard, P. Multi-modal virtual environments for education: From illusion to immersion. In *International Conference on Technologies for E-Learning and Digital Entertainment* (2006), Springer.

[293] Richardt, C., Tompkin, J., and Wetzstein, G. Capture, reconstruction, and representation of the visual real world for virtual reality. In *Real VR–Immersive Digital Reality*. Springer, 2020, pp. 3–32.

[294] Riecke, B. E., Väljamäe, A., and Schulte-Pelkum, J. Moving sounds enhance the visually-induced self-motion illusion (circular vection) in virtual reality. *ACM Trans. on Applied Perception 6*, 2 (2009).

[295] Rock, I., and Victor, J. Vision and touch: An experimentally created conflict between the two senses. *Science 143*, 3606 (1964), 594–596.

[296] Rosen, J. M., Soltanian, H., Redett, R. J., and Laub, D. R. Evolution of virtual reality [medicine]. *IEEE Engineering in Medicine and Biology Magazine 15*, 2 (1996), 16–22.

[297] Rosenkvist, A., Eriksen, D. S., Koehlert, J., Valimaa, M., Vittrup, M. B., Andreasen, A., and Palamas, G. Hearing with eyes in virtual reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (2019), IEEE, pp. 1349–1350.

[298] Ross, J., Morrone, M. C., Goldberg, M. E., and Burr, D. C. Changes in visual perception at the time of saccades. *Trends in Neurosciences 24*, 2 (2001), 113–121.

[299] Rothe, S., Buschek, D., and Hussmann, H. Guidance in cinematic virtual reality-taxonomy, research status and challenges. *Multimodal Technologies and Interaction 3*, 1 (2019), 19.

[300] Rothe, S., and Hussmann, H. Guiding the viewer in cinematic virtual reality by diegetic cues. In *International Conference on Augmented Reality, Virtual Reality and Computer Graphics* (2018), Springer, pp. 101–117.

[301] Rothe, S., Hussmann, H., and Allary, M. Diegetic cues for guiding the viewer in cinematic virtual reality. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology* (2017), pp. 1–2.

[302] Rubinstein, M., Gutierrez, D., Sorkine, O., and Shamir, A. A comparative study of image retargeting. *ACM Trans. on Graphics (Proc. SIGGRAPH Asia) 29*, 6 (2010), 160:1–160:10.

[303] Rubio-Tamayo, J., Gertrudix Barrio, M., and García García, F. Immersive environments and virtual reality: Systematic review and advances in communication, interaction and simulation. *Multimodal Tech. and Interact.* (2017).

[304] Rungta, A., Rewkowski, N., Schissler, C., Robinson, P., Mehra, R., and Manocha, D. Effects of virtual acoustics on target-word identification performance in multi-talker environments. In *Proceedings of the 15th ACM Symposium on Applied Perception* (2018), pp. 1–8.

[305] Ruotolo, F., Maffei, L., Di Gabriele, M., Iachini, T., Masullo, M., Ruggiero, G., and Senese, V. P. Immersive virtual reality and environmental noise assessment: An innovative audio–visual approach. *Environmental Impact Assessment Review 41* (2013), 10–20.

[306] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision 115*, 3 (2015), 211–252.

[307] Sadowski, W., and Stanney, K. Presence in virtual environments. *Human factors and ergonomics. Handbook of virtual environments: Design, implementation, and applications (p. 791–806).* (2002).

[308] Sakhardande, P., Murugan, A., and Pillai, J. Exploring effect of different external stimuli on body association in vr. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (2020), pp. 689–690.

[309] Salazar, S., Pacchierotti, C., de Tinguy, X., Maciel, A., and Marchal, M. Altering the stiffness, friction, and shape perception of tangible objects in virtual reality using wearable haptics. *IEEE Trans. on Haptics 13*, 1 (2020), 167–174.

[310] Salselas, I., Penha, R., and Bernardes, G. Sound design inducing attention in the context of audiovisual immersive environments. *Personal and Ubiquitous Computing* (2020), 1–12.

[311] Samad, M., Gatti, E., Hermes, A., Benko, H., and Parise, C. Pseudo-haptic weight: Changing the perceived weight of virtual objects by manipulating control-display ratio. In *Proc. of the Conf. on Human Factors in Computing Systems* (2019).

[312] Sanchez-Vives, M. V., and Slater, M. From presence to consciousness through virtual reality. *Nature Reviews Neuroscience 6*, 4 (2005), 332–339.

[313] Sano, Y., Ichinose, A., Wake, N., Osumi, M., Sumitani, M., Kumagaya, S.-i., and Kuniyoshi, Y. Reliability of phantom pain relief in neurorehabilitation using a multimodal virtual reality system. In *Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2015).

[314] Sarlat, L., Warusfel, O., and Viaud-Delmon, I. Ventriloquism aftereffects occur in the rear hemisphere. *Neuroscience letters 404*, 3 (2006), 324–329.

[315] Satava, R., and Jones, S. Current and future applications of virtual reality for medicine. *Proceedings of the IEEE 86*, 3 (1998).

[316] Schatzschneider, C., Bruder, G., and Steinicke, F. Who turned the clock? effects of manipulated zeitgebers, cognitive load and immersion on time estimation. *IEEE transactions on visualization and computer graphics 22*, 4 (2016), 1387–1395.

[317] Schmitz, A., MacQuarrie, A., Julier, S., et al. Directing versus attracting attention: Exploring the effectiveness of central and peripheral cues in panoramic videos. In *IEEE Conf. on Virtual Reality and 3D User Interfaces* (2020).

[318] Schneider, S. M., Kisby, C. K., and Flint, E. P. Effect of virtual reality on time perception in patients receiving chemotherapy. *Supportive Care in Cancer 19*, 4 (2011), 555–564.

[319] Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proc. Computer Vision and Pattern Recognition* (2015), pp. 815–823.

[320] Schuemie, M. J., Van Der Straaten, P., Krijn, M., and Van Der Mast, C. A. Research on presence in virtual reality: A survey. *CyberPsychology & Behavior 4*, 2 (2001), 183–201.

[321] Schultz, M., and Joachims, T. Learning a Distance Metric from Relative Comparisons. In *Advances in Neural Information Processing Systems* (2003).

[322] Schwartz, G., and Nishino, K. Material recognition from local appearance in global context. *arXiv preprint arXiv:1611.09394* (2016).

[323] Schwartz, G., and Nishino, K. Recognizing material properties from images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* (2019).

[324] Seinfeld, S., Feuchtner, T., Pinzek, J., and Müller, J. Impact of information placement and user representations in vr on performance and embodiment. *arXiv preprint arXiv:2002.12007* (2020).

[325] Seitz, A. R., Kim, R., and Shams, L. Sound facilitates visual learning. *Current Biology 16*, 14 (2006).

[326] Sekuler, R., Sekuler, A., and Brackett, T. When visual objects collide: Repulsion and streaming. *Investigative Ophthalmology and Visual Science 36*, 50 (1995).

[327] Sekuler, R., Sekuler, A. B., and Lau, R. Sound alters visual motion perception. *Nature 385*, 6614 (1997), 308.

[328] Serafin, G., and Serafin, S. Sound design to enhance presence in photorealistic virtual reality. Georgia Inst. of Tech.

[329] Serafin, S., Geronazzo, M., Erkut, C., Nilsson, N. C., and Nordahl, R. Sonic interactions in virtual reality: state of the art, current challenges, and future directions. *IEEE Comp. Graph. and App. 38*, 2 (2018).

[330] Serafin, S., Nilsson, N. C., Sikstrom, E., De Goetzen, A., and Nordahl, R. Estimation of detection thresholds for acoustic based redirected walking techniques. In *2013 IEEE Virtual Reality (VR)* (2013), IEEE, pp. 161–162.

[331] Serrano, A., Chen, B., Wang, C., Piovarči, M., Seidel, H.-P., Didyk, P., and Myszkowski, K. The effect of shape and illumination on material perception: model and applications. *ACM Transactions on Graphics (TOG) 40*, 4 (2021), 1–16.

[332] Serrano, A., Gutierrez, D., Myszkowski, K., Seidel, H.-P., and Masia, B. An intuitive control space for material appearance. *ACM Trans. on Graphics 35*, 6 (Nov. 2016), 186:1–186:12.

[333] Serrano, A., Martin, D., Gutierrez, D., Myszkowski, K., and Masia, B. Imperceptible manipulation of lateral camera motion for improved virtual reality applications. *ACM Trans. on Graphics 39*, 6 (2020).

[334] Serrano, A., Sitzmann, V., Ruiz-Borau, J., Wetzstein, G., Gutierrez, D., and Masia, B. Movie editing and cognitive event segmentation in virtual reality video. *ACM Trans. on Graph. (SIGGRAPH) 36*, 4 (2017).

[335] Seth, A., Vance, J. M., and Oliver, J. H. Virtual reality for assembly methods prototyping: a review. *Virtual reality 15*, 1 (2011), 5–20.

[336] Seymour, N. E., Gallagher, A. G., Roman, S. A., O'brien, M. K., Bansal, V. K., Andersen, D. K., and Satava, R. M. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Annals of surgery 236*, 4 (2002), 458.

[337] Shams, L., Kamitani, Y., and Shimojo, S. Illusions: What you see is what you hear. *Nature 408*, 6814 (2000), 788.

[338] Shams, L., and Kim, R. Crossmodal influences on visual perception. *Physics of life reviews 7*, 3 (2010).

[339] Shams, L., Ma, W., and Beierholm, U. Sound-induced flash illusion as an optimal percept. *Neuroreport* (2005).

[340] Shams, L., and Seitz, A. R. Benefits of multisensory learning. *Trends in cognitive sciences 12*, 11 (2008).

[341] Shams L, K. R. Crossmodal influences on visual perception. *Physics of Life Reviews* (2010).

[342] Shapiro, A. G., and Todorovic, D. *The Oxford compendium of visual illusions*. Oxford University Press, 2016.

[343] Sharan, L., Rosenholtz, R., and Adelson, E. Material perception: What can you see in a brief glance? *Journal of Vision 9*, 8 (2009), 784–784.

[344] Sharif Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proc. Computer Vision and Pattern Recognition (CVPR) Workshops* (2014), pp. 806–813.

[345] Shiban, Y., Pauli, P., and Mühlberger, A. Effect of multiple context exposure on renewal in spider phobia. *Behaviour Research and Therapy 51*, 2 (2013), 68–74.

[346] Shimojo, S., Scheier, C., Nijhawan, R., Shams, L., Kamitani, Y., and Watanabe, K. Beyond perceptual modality: Auditory effects on visual perception. *Acoustical Science and Technology 22*, 2 (2001), 61–67.

[347] Siddig, A., Ragano, A., Jahromi, H., and Hines, A. Fusion confusion: Exploring ambisonic spatial localisation for audio-visual immersion using the mcgurk effect. In *ACM Workshop on Immersive Mixed and Virtual Env. Systems* (2019).

[348] Siddig, A., Ragano, A., Jahromi, H. Z., and Hines, A. Fusion confusion: exploring ambisonic spatial localisation for audio-visual immersion using the mcgurk effect. In *Proceedings of the 11th ACM Workshop on Immersive Mixed and Virtual Environment Systems* (2019), pp. 28–33.

[349] Siddig, A., Sun, P., Parker, M., and Hines, A. Perception deception: Audio-visual mismatch in virtual reality using the mcgurk effect.

[350] Sillion, F. X., Rushmeier, H., and Dorsey, J. *Digital Modeling of Material Appearance*. Morgan Kaufmann/Elsevier, 2008.

[351] Simons, D. J., and Levin, D. T. Change blindness. *Trends in Cognitive Sciences 1*, 7 (1997), 261–267.

[352] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[353] Sitzmann, V., Serrano, A., Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., and Wetzstein, G. Saliency in vr: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics 24*, 4 (2018), 1633–1642.

[354] Slater, M. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Trans. of the Royal Society B: Biological Sciences 364*, 1535 (2009), 3549–3557.

[355] Slater, M., Khanna, P., Mortensen, J., and Yu, I. Visual realism enhances realistic response in an immersive virtual environment. *IEEE Computer Graphics and Applications 29*, 3 (2009), 76–84.

[356] Slater, M., and Usoh, M. Presence in immersive virtual environments. In *Proceedings of IEEE Virtual Reality Annual International Symposium* (1993), IEEE, pp. 90–96.

[357] Soler, C., Subr, K., and Nowrouzezahrai, D. A versatile parameterization for measured material manifolds. In *Computer Graphics Forum* (2018), vol. 37, pp. 135–144.

[358] Spanlang, B. e. a. How to build an embodiment lab: achieving body representation illusions in virtual reality. *Frontiers in Robotics and AI 1* (2014), 9.

[359] Spence, C., and Driver, J. Audiovisual links in exogenous covert spatial orienting. *Perception & psychophysics 59*, 1 (1997), 1–22.

[360] Spence, C., and Ho, C. Tactile and multisensory spatial warning signals for drivers. *IEEE Transactions on Haptics 1*, 2 (2008), 121–129.

[361] Spence, C., Lee, J., and Van der Stoep, N. Responding to sounds from unseen locations: Crossmodal attentional orienting in response to sounds presented from the rear. *European Journal of Neuroscience 51*, 5 (2017).

[362] Spence, C., and Parise, C. Prior-entry: A review. *Consciousness and cognition 19*, 1 (2010), 364–379.

[363] Spence, C., Senkowski, D., and Röder, B. Crossmodal processing, 2009.

[364] Steuer, J. Defining virtual reality: Dimensions determining telepresence. *Journal of Comm. 42*, 4 (1992).

[365] Stojšić, I., Ivkov-Džigurski, A., and Maričić, O. Virtual reality as a learning tool: How and where to start with immersive teaching. In *Didactics of Smart Pedagogy*. Springer, 2019, pp. 353–369.

[366] Strandholt, P., Dogaru, O., Nilsson, N., et al. Knock on wood: Combining redirected touching and physical props for tool-based interaction in virtual reality. In *Proc. of the Conf. on Human Factors in Computing Systems* (2020).

[367] Strasnick, E., Holz, C., Ofek, E., Sinclair, M., and Benko, H. Haptic links: Bimanual haptics for virtual reality using variable stiffness actuation. In *Proceedings of the Conference on Human Factors in Computing Systems* (2018).

[368] Suh, K.-S., and Lee, Y. E. The effects of virtual reality on consumer learning: an empirical investigation. *Mis Quarterly* (2005), 673–697.

[369] Sun, Q., Patney, A., Wei, L., Shapira, O., Lu, J., Asente, P., Zhu, S., McGuire, M., Luebke, D., and Kaufman, A. Towards virtual reality infinite walking: Dynamic saccadic redirection. *ACM Trans. on Graph. (TOG) 37*, 4 (2018), 67.

[370] Sun, T., Jensen, H. W., and Ramamoorthi, R. Connecting measured brdfs to analytic brdfs by data-driven diffuse-specular separation. *ACM Trans. on Graphics 37*, 6 (2018), 1–15.

[371] Sun, T., Serrano, A., Gutierrez, D., and Masia, B. Attribute-preserving gamut mapping of measured brdfs. In *Computer Graphics Forum* (2017), vol. 36, pp. 47–54.

[372] Swapp, D., Pawar, V., and Loscos, C. Interaction with co-located haptic feedback in virtual reality. *Virtual Reality 10*, 1 (2006), 24–30.

[373] Szàkely, G., and Satava, R. Virtual reality in medicine. *BMJ 319*, 7220 (1999), 1305.

[374] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. arxiv.

[375] Takemori, S. Visual suppression test. *Annals of Otology, Rhinology & Laryngology 86*, 1 (1977), 80–85.

[376] Taljaard, J. A review of multi-sensory technologies in a science, technology, engineering, arts and mathematics (steam) classroom. *Journal of Learning Design 9*, 2 (2016), 46–55.

[377] Tamuz, O., Liu, C., Belongie, S., Shamir, O., and Kalai, A. T. Adaptively learning the crowd kernel. In *Proc. International Conference on Machine Learning* (2011), pp. 673–680.

[378] Tang, Z., Patel, A., Guo, X., and Prabhakaran, B. A multimodal virtual environment for interacting with 3d deformable models. In *Proceedings of the ACM International Conference on Multimedia* (2010).

[379] Teichert, M., and Bolz, J. How senses work together: Cross-modal interactions between primary sensory cortices. *Neural Plasticity 2018* (2018).

[380] Thompson, W., Fleming, R., Creem-Regehr, S., and Stefanucci, J. K. *Visual Perception from a Computer Graphics Perspective*, 1st ed. A. K. Peters, Ltd., 2011.

[381] Tinsley, J. N., Molodtsov, M. I., Prevedel, R., Wartmann, D., Espigulé-Pons, J., Lauwers, M., and Vaziri, A. Direct detection of a single photon by humans. *Nature Communications 7*, 1 (2016), 1–9.

[382] Treue, S. Visual attention: the where, what, how and why of saliency. *Current Opinion in Neurobiology* (2003).

[383] Tripathi, A. K., Mukhopadhyay, S., and Dhara, A. K. Performance metrics for image contrast. In *2011 International Conference on Image Information Processing* (2011), IEEE, pp. 1–4.

[384] Trujillo-Ortiz, A. Matlab statistics support scripts. https://es.mathworks.com/matlabcentral/profile/authors/869509?detail=fileexchange. Accessed: 2021-12-09.

[385] Tsakiris, M. The multisensory basis of the self: from body to identity to others. *The Quarterly Journal of Experimental Psychology 70*, 4 (2017), 597–609.

[386] Tse, P. U., Intriligator, J., Rivest, J., and Cavanagh, P. Attention and the subjective expansion of time. *Perception & psychophysics 66*, 7 (2004), 1171–1189.

[387] Tuthill, J. C., and Azim, E. Proprioception. *Current Biology 28*, 5 (2018), R194–R203.

[388] Valori, I., McKenna-Plumley, P., Bayramova, R., Zandonella C., C., Altoè, G., and Farroni, T. Proprioceptive accuracy in immersive virtual reality: A developmental perspective. *PloS one 15*, 1 (2020).

[389] Van Assen, J. J. R., Barla, P., and Fleming, R. W. Visual features in the perception of liquids. *Current Biology 28*, 3 (2018), 452–458.

[390] Van der Burg, E., Olivers, C. N., Bronkhorst, A., and Theeuwes, J. Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance 34*, 5 (2008).

[391] Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., and Theeuwes, J. Poke and pop: Tactile–visual synchrony increases visual saliency. *Neuroscience letters 450*, 1 (2009), 60–64.

[392] van der Ham, I. J., Klaassen, F., van Schie, K., and Cuperus, A. Elapsed time estimates in virtual reality and the physical world: The role of arousal and emotional valence. *Computers in Human Behavior 94* (2019), 77–81.

[393] Van Der Maaten, L., and Weinberger, K. Stochastic triplet embedding. In *IEEE International Workshop on Machine Learning for Signal Processing* (2012), pp. 1–6.

[394] Van der Meijden, O. A., and Schijven, M. P. The value of haptic feedback in conventional and robot-assisted minimal invasive surgery and virtual reality training: a current review. *Surgical endoscopy 23*, 6 (2009).

[395] van der Stoep, N., Serino, A., Farnè, A., Di Luca, M., and Spence, C. Depth: The forgotten dimension in multisensory research. *Multisensory Research 29*, 6-7 (2016), 493–524.

[396] Van der Stoep, N., Serino, A., Farnè, A., Di Luca, M., and Spence, C. Depth: The forgotten dimension in multisensory research. *Multisensory Research 29*, 6-7 (2016), 493–524.

[397] Van Krevelen, D., and Poelman, R. A survey of augmented reality technologies, applications and limitations. *International Journal of Virtual Reality 9*, 2 (2010), 1.

[398] Vangorp, P., Barla, P., and Fleming, R. W. The perception of hazy gloss. *Journal of Vision 17*, 5 (2017), 19–19.

[399] Vangorp, P., Laurijssen, J., and Dutré, P. The influence of shape on the perception of material reflectance. *ACM Trans. Graph. 26*, 3 (July 2007).

[400] Vatakis, A., Balci, F., Di Luca, M., and Correa, Á. *Timing and time perception: Procedures, measures, & applications*. Brill, 2018.

[401] Viaud-Delmon, I., Warusfel, O., Seguelas, A., Rio, E., and Jouvent, R. High sensitivity to multisensory conflicts in agoraphobia exhibited by virtual reality. *European Psychiatry 21*, 7 (2006).

[402] Vidaurre, R., Casas, D., Garces, E., and Lopez-Moreno, J. Brdf estimation of complex materials with nested learning. In *IEEE Winter Conference on Applications of Computer Vision* (2019).

[403] Volkmann, F. C., Riggs, L. A., and Moore, R. K. Eyeblinks and visual suppression. *Science 207*, 4433 (1980), 900–902.

[404] Vorländer, M., Schröder, D., Pelzer, S., and Wefers, F. Virtual reality for architectural acoustics. *Journal of Building Performance Simulation 8*, 1 (2015), 15–25.

[405] Walker, B. N., and Lindsay, J. Effect of beacon sounds on navigation performance in a virtual reality environment. Georgia Institute of Technology.

[406] Wallgrün, J., Bagher, M., Sajjadi, P., and Klippel, A. A comparison of visual attention guiding approaches for 360° image-based vr tours. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (2020), IEEE, pp. 83–91.

[407] Waltemate, T., Gall, D., Roth, D., Botsch, M., and Latoschik, M. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. *IEEE Tran. on Vis. and Com. Graph.* (2018).

[408] Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing 13*, 4 (2004), 600–612.

[409] Wang, Z., Lubetzky, A., Gospodarek, M., TaghaviDilamani, M., and Perlin, K. Virtual environments for rehabilitation of postural control dysfunction. *arXiv preprint arXiv:1902.10223* (2019).

[410] Warshawsky-Livne, L., and Shinar, D. Effects of uncertainty, transmission type, driver age and gender on brake reaction and movement time. *Journal of safety research 33*, 1 (2002), 117–128.

[411] Wearden, J. Passage of time judgements. *Consciousness and Cognition 38* (2015), 165–171.

[412] Weber, S., Weibel, D., and Mast, F. Time perception, movement and presence in virtual reality.

[413] Welinder, P., Branson, S., Perona, P., and Belongie, S. J. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems* (2010), pp. 2424–2432.

[414] Whitmire, E., Benko, H., Holz, C., Ofek, E., and Sinclair, M. Haptic revolver: Touch, shear, texture, and shape rendering on a reconfigurable virtual reality controller. In *Proc. of the Conf. on Human Factors in Computing Systems* (2018).

[415] Wilberz, A., Leschtschow, D., Trepkowski, C., Maiero, J., Kruijff, E., and Riecke, B. Facehaptics: Robot arm based versatile facial haptics for immersive environments. In *Proc. of the Conf. on Human Factors in Computing Systems* (2020).

[416] Wilcox, L. M., Allison, R. S., Elfassy, S., and Grelik, C. Personal space in virtual reality. *ACM Trans. on Perception 3*, 4 (2006), 412–428.

[417] Wills, J., Agarwal, S., Kriegman, D., and Belongie, S. Toward a perceptual space for gloss. *ACM Trans. on Graphics 28*, 4 (Sept. 2009), 103:1–103:15.

[418] Wilson, C. J., and Soranzo, A. The use of virtual reality in psychology: A case study in visual perception. *Computational and Mathematical Methods in Medicine 2015* (2015).

[419] Winther, F., Ravindran, L., Svendsen, K. P., and Feuchtner, T. Design and evaluation of a vr training simulation for pump maintenance. In *Extended Abstracts of the Conf. on Human Factors in Computing Systems* (2020), pp. 1–8.

[420] Xu, M., Li, C., Zhang, S., and Le Callet, P. State-of-the-art in 360 video/image processing: Perception, assessment and compression. *IEEE Journal of Selected Topics in Signal Processing 14*, 1 (2020), 5–26.

[421] Xuan, B., Zhang, D., He, S., and Chen, X. Larger stimuli are judged to last longer. *Journal of vision 7*, 10 (2007), 2–2.

[422] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* (2014), pp. 3320–3328.

[423] Yu, W., and Brewster, S. Multimodal virtual reality versus printed medium in visualization for blind people. In *Proceedings of the International ACM Conference on Assistive Technologies* (2002), pp. 57–64.

[424] Yuan, Y., and Steed, A. Is the rubber hand illusion induced by immersive virtual reality? In *2010 IEEE Virtual Reality Conference (VR)* (2010), IEEE, pp. 95–102.

[425] Zaal, F. T., and Bootsma, R. J. Virtual reality as a tool for the study of perception-action: The case of running to catch fly balls. *Presence 20*, 1 (2011), 93–103.

[426] Zakay, D. Psychological time as information: The case of boredom. *Frontiers in psychology* (2014), 917.

[427] Zakay, D., and Block, R. A. The role of attention in time estimation processes. In *Advances in psychology*, vol. 115. Elsevier, 1996, pp. 143–164.

[428] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. Computer Vision and Pattern Recognition* (2018), pp. 586–595.

[429] Zhu, Y., Zhai, G., and Min, X. The prediction of head and eye movement for 360 degree images. *Signal Processing: Image Communication 69* (2018), 15–25.

[430] Zsolnai-Fehér, K., Wonka, P., and Wimmer, M. Gaussian material synthesis. *ACM Trans. on Graphics 37*, 4 (July 2018), 76:1–76:14.

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*".

Final Version as of December 20, 2022 (version 1.0).